

# Contents

Preface	ix
1 Subjunctions as discourse markers? Stancetaking on the topic ‘insubordinate subordination’ Werner Abraham	1
2 Two-layer networks, non-linear separation, and human learning R. Harald Baayen & Peter Hendrix	13
3 John’s car repaired. Variation in the position of past participles in the verbal cluster in Dutch Sjef Barbiers, Hans Bennis & Lotte Dros-Hendriks	23
4 Perception of word stress Leonor van der Bij, Dicky Gilbers & Wolfgang Kehrein	33
5 Empirical evidence for discourse markers at the lexical level Jelke Bloem	45
6 Verb phrase ellipsis and sloppy identity: a corpus-based investigation Johan Bos	57
7 <i>Om</i> -omission Gosse Bouma	65
8 Neural semantics Harm Brouwer, Matthew W. Crocker & Noortje J. Venhuizen	75
9 Liberating Dialectology J. K. Chambers	85
10 A new library for construction of automata Jan Daciuk	93
11 Generating English paraphrases from logic Dan Flickinger	99

*Contents*

12	<b>Use and possible improvement of UNESCO's <i>Atlas of the World's Languages in Danger</i></b> Tjeerd de Graaf	109
13	<b>Assessing smoothing parameters in dialectometry</b> Jack Grieve	119
14	<b>Finding dialect areas by means of bootstrap clustering</b> Wilbert Heeringa	127
15	<b>An acoustic analysis of English vowels produced by speakers of seven different native-language backgrounds</b> Vincent J. van Heuven & Charlotte S. Gooskens	137
16	<b>Impersonal passives in German: some corpus evidence</b> Erhard Hinrichs	149
17	<b><i>In Hülle und Fülle</i> – quantification at a distance in German, Dutch and English</b> Jack Hoeksema	159
18	<b>The interpretation of Dutch direct speech reports by Frisian-Dutch bilinguals</b> Franziska Köder, J. W. van der Meer & Jennifer Spenader	171
19	<b>Mining for parsing failures</b> Daniël de Kok & Gertjan van Noord	181
20	<b>Looking for meaning in names</b> Stasinou Konstantopoulos	191
21	<b>Second thoughts about the Chomskyan revolution</b> Jan Koster	199
22	<b>Good maps</b> William A. Kretzschmar, Jr.	211
23	<b>The presentation of linguistic examples in the 1950s: an unheralded change</b> Charlotte Lindenbergh & Jan-Wouter Zwart	221
24	<b>Gravity, radiation, and dialectometry</b> Robert Malouf	233
25	<b>Exploring the role of extra-linguistic factors in defining dialectal variation patterns through cluster comparison</b> Simonetta Montemagni & Martijn Wieling	241

<b>26 Default inheritance and derivational morphology</b> Stefan Müller	253
<b>27 Licensing resultative phrases: the case of locatum subject alternation verbs in Japanese</b> Tsuneko Nakazawa	263
<b>28 Free word order and MCFLs</b> Mark-Jan Nederhof	273
<b>29 Keystroke dynamics for authorship attribution</b> Barbara Plank	283
<b>30 Quantitative diachronic dialectology</b> Jelena Prokić	293
<b>31 Write as you speak? A cross-linguistic investigation of orthographic transparency in 16 Germanic, Romance and Slavic languages</b> Anja Schüppert, Wilbert Heeringa, Jelena Golubovic & Charlotte Gooskens	303
<b>32 Vowel co-occurrence restriction in Ainu</b> Hidetoshi Shiraishi	315
<b>33 From dialectometry to semantics</b> Dirk Speelman & Kris Heylen	325
<b>34 On blended selfies and tainted smoothies</b> Oscar Strik, Muriel Norde & Karin Beijering	335
<b>35 “Featurometry”</b> Benedikt Szmrecsanyi	345
<b>36 Bootstrapping a dependency parser for Maltese – a real-world test case</b> Jörg Tiedemann & Lonneke van der Plas	355
<b>37 Identifying dialect regions from syntactic data</b> Erik Tjong Kim Sang	367
<b>38 Morphology changes faster than phonology</b> Esteve Valls	375
<b>39 Effective communication. A Platonic case study</b> Gerry C. Wakker	387
<b>40 Variation: from dialect to pragmatics, a progress report</b> Annie Zaenen, Brian Hicks, Cleo Condoravdi, Lauri Karttunen & Stanley Peters	399

*Contents*

**41 Talking about beliefs about beliefs without using recursion**

Denise Zijlstra, Marieke Wijnbergen, Margreet Vogelzang & Petra Hendriks **409**

## Preface

This *Festschrift* is meant as a celebration of John Nerbonne's career as a researcher. Two things immediately stand out. The first is the weight of this book, which is caused by the large number of contributions by an even greater number of people. More than 65 researchers have contributed to this impressive collection of scientific papers. This exemplifies not only John's productive career in which he has published extensively, and often with colleagues from different universities, but also his open, friendly and constructive approach to others. The second thing standing out is the great variation in terms of content of the contributions, ranging from semantics to dialectology. While in the last two decades John has focused primarily on his dialectometric work – thereby founding the approach which is known in the rest of the world as the “Groningen School of Dialectometry” – he is highly knowledgeable and has published in a large number of areas within Computer Science and Computational Linguistics.

When John Nerbonne came to Groningen in 1993 as the new professor in Alfa-informatica (see Figure 1), the Alfa-informatica department consisted of two temporary lecturers in computational linguistics (Gosse Bouma and Gertjan van Noord), one historian (George Welling), and one researcher (Harry Gaylord) with a background in biblical studies and an interest in digital texts. Alfa-informatica, often referred to as “Humanities Computing” in English, can be seen as a kind of Digital Humanities *avant la lettre*. Under Nerbonne's guidance, the department maintained an interest in the various aspects of Humanities Computing, in particular in computational linguistics.

Through the years, John has been much more than just the head of the Alfa-informatica department. Being an engineer with a background in industry, he was initially seen by many in the Faculty of Arts as somewhat of an outsider. This changed quickly, however, as John was always keen on establishing contacts with others both within and outside the Faculty of Arts. He was director of the Center for Language and Cognition Groningen (CLCG) for several years, thus positioning computational linguistics as one of the key themes of linguistics research in Groningen. He was also immensely important as teacher, especially of many statistics courses, for a range of programs at both the BA and MA level.

The current state of the Alfa-informatica study programme and the computational linguistics research group makes it easy to forget that, for a long time, Alfa-informatica attracted only a modest numbers of students, thus putting staff under constant pressure to find new teaching and funding opportunities. One of the results of this is that the department has merged recently with Communication Science to become the Communication and Information Science department. At the same time, student



Figure 1: Dutch newspaper article about John Nerbonne's appointment as professor in Groningen (*Nieuwsblad van het Noorden*, Friday May 8, 1992, p. 13).

numbers began to rise, which has led to the unprecedented situation that new staff members needed to be recruited for several years in a row.

Under Nerbonne's leadership, computational linguistics flourished enormously in Groningen. The computational linguistics group now consists of an astonishing number of twelve (!) staff members, including three professors. One important aspect of his leadership which helped in establishing Groningen as a hot spot for computational linguistics was his capability to identify opportunities. The following examples illustrate this.

It must have been around 1998 when Nerbonne suggested to van Noord that he should apply for a Dutch NWO Pionier grant. In order to prepare better for such a grant proposal, a successful grant proposal of a previous year was somehow obtained. Van Noord vividly remembers the enormously impressive list of publications of previous year's applicant, and his sincere conviction that he, van Noord, had absolutely no chance whatsoever to get that funding. Nerbonne convinced him to try anyway – and less than two years later the Pionier grant (over 2 million Dutch guilders which paid for four PhD positions and two post-doc positions) landed in Nerbonne's group.

#### JOHN AS SUPERVISOR

The first time I came into contact with John Nerbonne was when following his course on Machine Learning in 2005. As John was a very open and knowledgeable teacher, I was very glad that he accepted to be my Master thesis supervisor and two years later to be my PhD thesis supervisor. During the period of four years as John's PhD student, I came to know him as someone whose office door is always open.<sup>1</sup> We did not have many formal meetings, but given his open-door-policy this was also entirely unnecessary. Only when talking to PhD students at other departments, I learned that his policy was a (very positive) exception rather than the rule. I remember John as a supervisor who was able to bring out the best in you. He had several approaches for this. Frequently, he knew just how to ask the right questions to get you (back) on track. That he knew which questions to ask is in no small part due to his extensive knowledge about our field, but also many other fields. I was always impressed by his inevitable intelligent questions when attending a guest lecture of someone in a completely different field. Another approach John employed to bring out the best in you, was for example stimulating scientific creativity and productivity by casually remarking that there might be an interesting workshop in Singapore or some other exotic place, but that the deadline for submitting an eight-page paper was already in two weeks. Or he simply suggested attending a workshop of a visiting scientist who did not exactly work in our field, but he thought would be useful for my development. That this person eventually became my second promotor is in no small part due to John. What I remember most, however, is his advice to PhD students: "Work hard, play hard!" We all know that life in academia requires a lot of effort, but it's important not to forget that there should also be room for a personal life outside of the university. It's a mantra I still try to live by today. I would not have been where I am today without John. I am grateful that I had the opportunity to work with him, and hope we will continue collaborating for many years to come.

<sup>1</sup>Generally from about 7.30 AM to about 7.30 PM.

Several years later, the Executive Board (*College van Bestuur*) of the university initiated an "Endowed Chair" programme. The university made available extra funding for an Endowed Chair professorship – together with funding for PhD students and postdocs – for each faculty, for which an international top researcher ("a potential winner of the Dutch NWO Spinoza prize") could be convinced to come work in Groningen. Again, Nerbonne saw opportunities that most of his close colleagues did not see. As a result, a few years later Alfa-informatica alumnus Johan Bos was appointed Endowed Chair at the Faculty of Arts!

John Nerbonne was also very successful as a supervisor of PhD students. He managed to supervise more than 40 successful PhD projects – in itself already an enormous achievement. In the inset, Martijn Wieling reports on his memories of John Nerbonne in the role of PhD supervisor. Nerbonne was quite successful in super-

## *Preface*

vising brilliant PhD students – but perhaps in contrast to other supervisors, he was particularly good at motivating students working on projects that – for various reasons – were not finished within the four year period normally available for a PhD. In several of these cases, Nerbonne was very insistent in motivating these students – which often were distracted by a new job to pay for their living – and simply did not allow them to give up.

Internationally, his reputation as a successful and influential researcher in computational linguistics was confirmed by his election as President of the Association for Computational Linguistics in 2002. His reputation was also recognized at the national, Dutch, level. In 2005, Nerbonne was awarded membership of the Dutch Royal Academy (KNAW), and in 2014, Nerbonne received a royal decoration (*Ridder in de Orde van de Nederlandse Leeuw*).

The papers in this volume are all written by distinguished scholars, colleagues, former colleagues and former PhD students of John Nerbonne, and in many ways reflect the breadth of his interests. Although readers with a background in almost any field of linguistics will find papers worth reading, it is hard to imagine anyone reading all of the papers with equal interest – except for one person, we hope...

## Chapter 1

# Subjunctions as discourse markers? Stancetaking on the topic 'in subordinate subordination'

Werner Abraham

Ludwig Maximilian University of Munich

The present paper is about types of seeming subordination (i.e. subordination by form), which in fact bears all distributional assets of inordination. Yet, such inordinate subordination is characterized by sentential autonomy in terms of illocutive force (i.e., it is not only presupposed as subordinates usually are). It will be shown that a reliable diagnostic is provided by the selection of modal particles in the narrow sense, which German and Dutch excel in. It will be illustrated that even the property of matrix V2 as opposed to subordinate Vlast, otherwise a reliable diagnostic in German and Dutch, are not sufficient conditions for subordination vs. matrix status, sentential presupposition vs. assertive status, and independent (autonomous) illocutive force vs. dependent (inherited) illocutive force. Modal particles play a major diagnostic role in this categorial dichotomy. It will be argued that since formal subordination is an insufficient condition for the separation of formal and notional independence of sentential autonomy the theory of sentence types will have fundamentally to be modified with the main asset of a thorough notional rather than formal definition of sentential autonomy.<sup>1</sup>

## 1 What might tempt one to analyze *wenn/if* as a discourse marker

The goal of the present discussion is the clarification of the status of sentences with the looks of subordination, but which nevertheless appear autonomously like (1)-(4) below.<sup>2</sup> The recent debate on this phenomenon has been traced back to Evans (2007), but it has, in fact, a far longer tradition, at least in the literature on German data. The gist of the ensuing discussion is as follows:

---

<sup>1</sup> I have profited from critical remarks extended by Nick Evans on a draft version of this.

<sup>2</sup> Clausal autonomy is more than clausal root. In other words, they select illocutive force AND can be uttered without matrix support.

1. The subordinate *wenn/if* (and its equivalents in other languages, notably German), as an ‘in subordinate subordinator’/IS, may neither be a true subordinator in a number of cases, nor can it be a discourse marker for various reasons, among which that *wenn/if* retains its original conditional meaning and that other subordinations fail to undergo the same grammaticalizing result. The latter issue fails to substantiate with generalizing force the classification of *wenn/if* as a discourse marker.
2. Once *wenn/if* is taken to be possible and meaningful also in what is an embedded form (which, counter to English, is beyond doubt in German), we shall add it to the list of matrix clauses marking special speech acts counteracting, undoubtedly, all syntactic canons.
3. Since, furthermore, the phenomena under discussion show also prosodic traits of autonomous declaratives, we conclude that the pf-module hosting the prosodic rules overwrite syntactic and semantic-truth logical rules with the effect to autonomize phenomena of in subordinate subordination. Such steps, however, do not invalidate the canonic syntactic and semantic rules (such as word order, MP-insertibility, subjunctive cliticization, etc.).

The problem to be solved in the present discussion is briefly this: in the pertinent literature we find solutions to the extent that such original subordinators as *if* are reclassified as discourse operators licensing sentences in their own matrix like right. This position is here refuted. The phenomenon, more generally, is this: there appear to be exclamative sentences introduced by subordinators, which nevertheless stand by themselves yielding the speech act effect of independent clauses. See the following illustration (1)-(4), where the respective subjunction words appear underlined. [Italics for valid English sentences, flat Roman for glosses of the German sentences. ‘Syntactic form’ refers to ‘embedded’ vs. ‘non-embedded’; parentheses signal elided protasis or apodosis.]

	Syntactic form	speech act autonomy
(1) <u>Wenn</u> er heute nur hier wäre! <i>if he only were here today!</i>	C( $\cap$ ?)	autonomous
(2) <u>Was</u> DAS nun wieder ist! <i>what that again might be!</i>	(CP $\cap$ )CP	autonomous
(3) <u>Ob</u> das heute wohl noch geht? <i>if that today still works</i>	(CP $\cap$ )C	autonomous
(4) <u>Dass</u> DAS denn noch geht heute!? <i>that this still works today</i>	C( $\cap$ CP)	autonomous

Notice that (1), due to the insertion of the modal particle *nur* ‘only’, cannot be thought of as a conditional subordinate. There is no embedding matrix clause imaginable. (2) is possible as an apodosis sentence to a performance verb question (*Er fragte sich,...* ‘he asked himself...’). The same holds for a performance predicate protasis for (3) and, as an apodosis, for (4). However, any superordinate clause other than by the form of a performance predicate (*say, believe, ask*) appears impossible. Again, the insertions of the modal particles/MPs *wohl* ‘well’ and *denn* ‘then’ render (3) and

## 1 Subjunctions as discourse markers?

(4) as insubordinable to any matrix clause. Given that *wenn-dass/if-that* (henceforth IS) count as subjunctions subordinating complement sentences as well as the fact that such sentences fully comply with our communicative understanding as emotive utterances with wishful import may lead to the conclusion that all there remains to classify *if* in an autonomous utterance is a discourse marker.

All of this lets us conclude that the autonomy of sentences as speech acts in their own right cannot be made contingent alone on the presence of an embedding matrix clause. (1)-(4) are not embeddable while doubtlessly autonomous. We emphasize that one of the factors for clause typing is the insertion of MPs – grammatical (not lexical!) adverbials affecting the speech act quality in C(omp) (Abraham 2013; 2014; Struckmeier 2010). We will return to this type determinant.

Our refutation of the claim that *if, what (wenn, was, and a few more in German; Dutch als, wat and more)* are autonomous discourse operators will be based on the following empirical problems:

- all ‘insubordinate subordinations’ have the looks of subordinated sentences (V-final in German and Dutch);
- subordinated sentences generally disallow the insertion of modal particles for reasons of their presuppositional (rather than assertive) status;
- the clash between categorial status of the pertinent subjunctions and sentential autonomy is to be solved along two different lines:
- either the lexical subjunction is recategorized with a fundamentally different licensing force (as indicated in the literature referred to in the incipient paragraph) – this position will have to say what is behind the syntactic and semantic-pragmatic properties reserved to subordination;
- or the list of autonomous sentences is extended equally fundamentally. This step requires that the criteria for motivating sentence types/STs are extended such that not only the new STs, but also the old ones are covered under one common denominator. This line of argumentation need not change the syntactic-semantic status of subordinators. They remain subordinators both semantically and syntactically.

We will pursue the line of argumentation using MPs as diagnostics for sentential autonomy (second in the list above). Notice that much depends of the categories and sentential characteristics of subordinated sentences. The fact that subordination is clearly signaled by word order and MP insertion in German (and Dutch), but not in English and French<sup>3</sup> (since there are no such differences in the first place) indicates very strongly that our solution may not be valid cross-linguistically. The ensuing discussion follows basic insights on the distinction of declarative meaning and ‘locution’ meaning in regard to the assumption in update semantics that the meaning of

---

<sup>3</sup> Proclisis characteristic of matrix clauses is replicated in embeddings: *Moi je crois que... – Il a dit que moi je crois...* Proclisis is retained. French subordination is motivated semantically, not syntactically as in German. The same holds for English.

a sentence is its context change potential/CCP (cf. Gunlogson (2003: 50–51) before the background prepared by Heim (1992) and others). The modification made here is that the CCP of a sentence is defined in terms of an update to a substructure of the context, i.e. the commitment set (cs) of an individual participant. Consistent with the accompanying working hypotheses to be unfolded, intonational rise and fall will serve to identify the individual cs to be updated, given an utterance context, i.e., a context in which individual participants can be identified in the roles of Speaker/Sp and Addressee/Addr.

## 2 Why *wenn/if* can never be a discourse marker: the necessary requirement<sup>4</sup>

It may seem difficult to tell whether English *if* still has the status of a subjunction in sentences like (1). After all, (1) has a fully saturated communicative meaning even as long as devoid of an appropriate matrix clause. Furthermore, it appears tough, or maybe even impossible, to find a corresponding matrix clause that, even if elided elliptically, may posit the kind of dependency required for (1).

It is argued here that any subjunction requiring finite V to stay in clause-last position in German is at the bottom of the explanation why *if*-clauses may receive independent status as an exclamative speech act. This is a necessary, a *sine qua non* condition on this phenomenon. However, as is obvious from comparing *if*-independents with imperatives and *yes-no* questions independents, an extra condition must be added to come to terms with the difference between independent *if* and equally independent, but *V-to-C* motivated imperatives and sentence questions. Let us first develop the first step in our line of argumentation as illustrated in Table 1. Notice that we assume the illocutive force of the exclamative to derive from FIN ('C(omp)' in syntactic terminology). In other words, V-final-move-to-C is the necessary, albeit insufficient requirement for IS. The main body of our illustrations comes from German, which has an extended literature on sentence types (Altmann 1988; Altmann, Batliner & Oppenrieder 1989; Meibauer, Steinbach & Altmann 2013) and has variously addressed the problem of sentence type interpretation (cf. Kaiser & Struckmeier 2015). Notice that all cases of IS show an empty prefield indicative of a strong intervention effect (of feature alignment) between C(omp/FIN), the left predicate bracket, and the prefield/SpecCP. See Table 1.

[The clausal representation in terms of the five sentence fields in Table 1 simplify structural bracketing or a corresponding tree representation. IS abbreviates the classification as 'in subordinate subordination'. German is a strict OV language due to its

---

<sup>4</sup> To achieve discourse and textually coherent representations, we advocate a bottom-up strategy. Thus, we stand behind arguments in favor of the Uniformity Hypothesis, i.e. the hypothesis that discourse representations and the respective requirements (e.g. text coherence) can, and need to, extend clause syntax dependencies without conflicting with them. As a consequence, any top-down strategy for the representation of text and discourse coherence as the ones dominating the literature are seen to fail for the mere reason that the generative prerequisites for sentences are necessary (although not sufficient) also for discourse and text structure and consequently cannot be dismissed.

1 Subjunctions as discourse markers?

principled left directed valence force as in *Mutter<sub>DAT</sub> ein Geschenk<sub>ACC</sub> bringen* ‘bring mother a present’.]

Table 1: German sentence structure based on clause type, speech act, and subordination (caps for (contrastive) stress).

Prefield	FIN/C(omp)	Left MF	Right MF	Lexical V	IS
<i>Er</i> he	<i>wäre</i> would	–	<i>heute gerne</i> today	<i>hier</i> here	–
<i>Es wäre schön</i> it would be nice	<i>wenn</i> if	<i>er</i> he	<i>bloß heute</i> only today	<i>hier wäre</i> here were	–
–	<b>Wenn</b> if	<b>er *(bloß)</b> he only	<b>heute</b> today	<b>hier wäre!</b> here were	+
<i>Er fragt sich</i> he wonders	<i>was</i> what	<i>das nun</i> this now	<i>wieder</i> again	<i>ist</i> is	–
<b>Was</b>	–	<b>das nun</b>	<b>wieder</b>	<b>ist!?</b>	+
<i>Das</i> this	<i>geht</i> goes	–	<i>heute noch</i> today still	–	–
–	<b>Dass</b>	<b>das</b>	<b>heute noch</b>	<b>geht!?</b>	+
<i>Es ist undeutlich</i> it is not evident	<i>ob</i> if	<i>das</i> this	<i>heute noch</i> today still	<i>geht?</i> goes	–
–	<b>Ob</b>	<b>das</b>	<b>heute noch</b>	<b>geht?</b>	+
<i>Er weiß nicht</i> he knows not	<i>dass</i> that	<i>das</i> this	<i>heute noch</i> today still	<i>geht</i> goes	–
<i>Das</i> this	<i>ist</i> is	<i>nun</i> now	<i>etwas zu viel</i> a little too much	–	–
<b>Das</b>	<b>*ist/IST</b>	<b>nun</b>	<b>wieder etwas!</b>	–	+
<i>Wer</i> who	<i>ist</i> is	<i>das</i> that	<i>*nur?</i> only	–	–
<b>Wer</b>	<b>*ist/IST</b>	<b>das</b>	<b>nur!</b>	–	+

English syntax will typically provide the following structures for the three (four including negation) different sentences as in Table 2.

Quite generally, discourse markers appear outside of CP in English (for example as

Table 2: English sentence structure with respect to clause type, speech act, and subordination.

SpecCP	FIN-AUX	VP	Postfield	IS	
<i>He</i>	<i>would</i>	<i>rather</i>	<i>be here</i>	<i>today</i>	-
<i>It would be nice</i>	<i>if</i>	<i>he</i>	<i>were here</i>	<i>today</i>	-
<i>He</i>	<i>can-</i>	<i>-not</i>	<i>be here</i>	<i>today</i>	-
<i>If *(only) he</i>	<i>would</i>		<i>be here</i>	<i>today!</i>	+

interjections or theticals). Given such status for *if* in Table 2, however, it would have to be the combination *if only* since *if* by itself would not be grammatical. The latter fact yields even more pressure to the semantic result: both *if* and *only* contribute substantively to the entire meaning of the clause with *if* retaining its conditional force and *only* adding some of the notion as a temporal adverb. Discourse markers have as their typical meaning something that is very independent of the semantics and syntax of the CP it refers to (either left dislocated or parenthetically linearized). We conclude that *if* in Table 2 cannot be a discourse marker.

The comparison with Table 1 on German makes this conclusion irrevocable. Independent and dependent clauses are structurally and linearly different. *Wenn* ‘if’ appears in Comp thereby forcing the finite predicate to get stuck in the V-final position. Thus, it cannot be assigned independent status by syntactic criteria.

### 3 Why *wenn/if* can never be a discourse marker: the prosodic disentangler as a sufficient criterion.

While each of the embedded sentences in Table 1 and Table 2 has rising prosody due to its question speech act or falling prosody given its declarative status, the exclamations in question have the opposite tone development, namely high. Phonetic studies have proposed high boundary tones as an intonational cue for “non-final” clauses in a wide range of languages (Gussenhoven & Chen 2000; Beckman et al. 2002). Following this finding, it is argued here that, while elliptical constructions have rising intonation patterns, syntactic independent clauses show falling patterns. Therefore, we propose prosody as an acoustic cue for the level of the syntactic (in)dependency of the construction in question (cf. Elvira-García 2015). This may appear to be the end of the story. Sentences, then, simply will have to be listed under two main parameters: linear form and speech act as necessary conditions and, possibly, intonation and corrective focus/CF as sufficient ones. We shall opt for exactly this confluence of licensing factors. Notice that it is not made clear at what stage of the derivation sentence prosodic tone comes in and whether it is of any sentence type determining effect. See Table 3 for the additional tone and focus characterization.

English syntax will typically provide the following sentence structures for the

1 Subjunctions as discourse markers?

Table 3: German sentence structure wrt. clause type, speech act, and subordination [VF=verum focus, CF=corrective focus].

Prefield	FIN	Left MF	Right MF	Lexical V	Tone	S-type
<i>Er</i> he	<i>wäre</i> were	–	<i>heute gerne</i> today	<i>hier</i> here	high-low ↓	Declarative independent
<i>Schön</i> nice	<i>wenn</i> if	<i>er</i> he	<i>bloß heute</i> just today	<i>hier wäre</i> here were	high-low ↓	Conditional dependent
–	<b><i>Wenn</i></b> if	<b><i>er *(bloß)</i></b> he only	<b><i>heute</i></b> today	<b><i>hier wäre!</i></b> here were	high-low ↓	‘Conditional’ independent
<i>Er fragt sich</i>	<i>was</i>	<i>DAS nun</i>	<i>wieder</i>	<i>ist</i>	CF-low ↓	Indirect Q dependent
<b><i>Was</i></b>	–	<b><i>DAS nun</i></b>	<b><i>wieder</i></b>	<b><i>ist!?</i></b>	CF-low	Indirect Q independent
<i>Das</i>	<i>geht</i>	–	<i>heute noch</i>	–	high-low ↓	Declarative independent
–	<b><i>Dass</i></b>	<b><i>das</i></b>	<b><i>heute noch</i></b>	<b><i>geht!?</i></b>	low-high	Indirect Q independent
<i>S’ist unklar</i>	<i>ob</i>	<i>das</i>	<i>heute noch</i>	<i>geht</i>	high-low ↓	Indirect Q dependent
–	<b><i>Ob</i></b>	<b><i>das</i></b>	<b><i>heute noch</i></b>	<b><i>geht?</i></b>	low-high	Indirect Q independent
<i>Er weiß nicht</i>	<i>dass</i>	<i>das</i>	<i>noch heute</i>	<i>geht</i>	high-low ↑↓	S- complement dependent
<i>Das</i>	<i>geht</i>	–	<i>heute noch</i>	–	high-low ↓	Declarative independent
–	<b><i>Dass</i></b>	<b><i>das</i></b>	<b><i>heute noch</i></b>	<b><i>geht!!</i></b>	low-high ↑	S- complement independent
<i>Das</i>	<i>ist</i>	<i>nun</i>	<i>etwas zu viel</i>	–	high-low ↑↓	Declarative independent
<b><i>Das</i></b>	<b><i>*ist/IST</i></b>	<b><i>nun</i></b>	<b><i>wieder etwas!</i></b>	–	VF-low ↓	Declarative Verum focus
<i>Wer</i>	<i>ist</i>	<i>DAS</i>	<i>(*nur)?</i>	–	low-high	w-question
<b><i>Wer</i></b>	<b><i>*ist/IST</i></b>	<b><i>DAS</i></b>	<b><i>(nur)!</i></b>	–	high-low ↓	w- QUESTION VERUM FOC.

three different sentences together with the tone assignments. See Table 4.

Table 4: English sentence structure based on clause type and speech act (no special form of subordination).

SpecCP	FIN-AUX	VP	Postfield	Tone	
<i>He</i>	<i>would</i>	<i>rather</i>	<i>be here</i>	<i>today</i>	L-↑↓
<i>It would be nice</i>	<i>if</i>	<i>he</i>	<i>were here</i>	<i>today</i>	L-↑↓
<i>If *(only) he</i>	<i>would</i>		<i>be here</i>	<i>today!</i>	L-↑

## 4 Clause types and alleged insubordination

In the recent past, there has been a renewed interest in sentence types/STs in German. ST classifications are based on sentence type criteria, such as word order, predicate bracket in German including V-final in subordinate sentences, the discourse based distribution of clause parts inside the predicate bracket, the choice of material in the prefield/SpecCP, the insertion of modal particles in the middle field, etc. See 4.1. below for an intensional definition of STs by Lohnstein (2000; 2007, based on Groenendijk & Stokhof 1982; 1997; Higginbotham 1996; Meibauer, Steinbach & Altmann 2013; Thurmair 2013) relating the interaction between syntactic form and ST. We will first recapitulate Lohnstein’s definitions and then apply them to the type of sentences that are autonomous, but nevertheless are dependents in German (as V-final sentences and headed by subjunctors).

### 4.1 Classic and non-classic sentence types

According to Lohnstein (2000, 2007; Abraham 2014; Struckmeier 2010), the following definitions interrelate syntactic form and ST. Evaluation under some world of discussion means that the clause is independent (matrix) and that a truth value can be assigned. No such evaluation yields a dependent sentence. Further structural prerequisite: Any verb occupying the left bracket/Comp has to be finite. If, and only if, a verb or its the auxiliary part occupies the left bracket/Comp does a sentence have illocutionary force (a specific speech act type) on its own. Sentences are semantically objects assigned to sets of worlds or an individual world. Let us see which semantic types of world assignable objects we can distinguish.

### 4.2 Types of semantic objects

**A CLAUSE DEPENDENCY – TYPICAL V-FINAL:** If both the lexical and the auxiliary (*haben, sein, werden*; modal verbs) part of the predicative verb, either finite or non-finite, remains in the right bracket (in ‘V-final’), the proposition denoted by the sentence will not be evaluated against a world under discussion. Such a clause

is not asserted and does not carry illocutive force in its own right, but it does that only in the context of a matrix clause (taking over the illocutive force of the matrix clause). The clause under discussion is thus presupposed and does not carry illocutive force in its own right. Dependent upon the type of matrix predicate (factive vs. non-factive), such dependent clauses do not allow the insertion of MPs (*ja, denn, eben*), since such illocution establishing grammatical material is nonsensical in presuppositions. If embedded under non-factive predicates (*say, believe, think*) and allowing bridge constructions, MP insertion is possible since there is no status of dependency in the first place. In canonic terminology: SUBORDINATE CLAUSE (PRESUPPOSED).

**B** CLAUSE INDEPENDENCY – TYPICAL V2: If, and only if, the finite part of the predicate (either as an auxiliary, modal verb, copula or lexical verb) occupies the left bracket (Comp/C°) and the prefield/SpecCP is not filled with a question word (*what, who(m), where, when* etc.) does a sentence have illocutionary force on its own and is assertive. The proposition denoted by the sentence will be evaluated against a world under discussion (common plus subjective grounds). In canonic terminology: MAIN or MATRIX CLAUSE (ASSERTION).

**C** WORD INTERROGATIVE CLAUSE – V2: If a [+wh] phrase (*who(m), where, when, what*) occupies the prefield and the finite (part of the) predicate stays in the right bracket (V-final), the derived semantic object consists of the partition of all possible worlds into a set of sets. In canonic terminology: WORD INTERROGATIVE/W-QUESTION CLAUSE.

**D** Yes-no INTERROGATIVE CLAUSE – V1: The sentence begins with the finite predicate (auxiliary or lexical) and leaves the prefield unoccupied. Thus, it fills the left clause bracket, Comp. Placing the predicate (verb) in the left bracket/Comp and leaving the prefield empty entails a semantic object that conforms to a bipartition of possible worlds. The meaning of such a clause is undecided between two sets of worlds: one where the denoted proposition is true and another one where the denoted proposition is false. In canonic terminology: INTERROGATIVE/YES-NO-QUESTION CLAUSE.

**E** IMPERATIVE CLAUSE – V1: The predicate (verb) is placed in the left bracket/Comp and leaves the prefield unoccupied. The predicative bracket is preserved (in the case that Comp takes the finite auxiliary or modal verb), while the non-finite part is in the right bracket/V-final. This derives a semantic object that conforms to change of the propositional to the discourse force of one individual world and only that, determined in the mind of the speaker and leaving no other choice on the side of the addressee. In canonic terminology: IMPERATIVE.

**F** Illocutive force is a prerequisite for syntax-semantic autonomy (i.e. that a sentence can be interpreted with reference to a set of accessible worlds and thus becomes a semantic object. Among all types of syntactic subordination, non-factive complements as well as logical adverbial sentences are exceptions to the extent that they project illocutive force independent of that for the matrix clause. MPs (of the German and Dutch type) are unmistakable signals for illocutive force. To the extent that they generate illocution as well as due to the fact that they occur independently from syntactic signals of (in)dependence, they serve as classifiers of STs better than

Werner Abraham

any other overt parameter. This dissociates the ST question from the traditional parameters of word order, the occurrence of subjunctions, and, last but not least, finiteness of the predicate).<sup>5</sup>

**G** Given that verbal factivity and temporality/locality on adverbial dependent clauses play the same role in disallowing MPs (Abraham 2013; 2014) and thus waive separate illocutivity, the question arises what is the common feature of the two disperse phenomena. We will come back to that.

## 5 Conclusion

We have argued that ‘in subordinate subordination’ is nothing else but illocutionary autonomy of dependent forms. No new category of discourse marker has to be introduced. The clue is that the emerging sentential form, the exclamative (importing the speech act of unexpectedness) receives sentential autonomy (main clause status) despite its dependent indexation (the Comp slot occupied by a subjunctor and Vfinal in German and Dutch). As such, Evans’ (2007) grammaticalizing derivation through matrix ellipsis is not based on the exhaustion of empirical material in the main pertaining language (German). The gist of our mainly descriptive essay is that the sentence type of exclamative has full illocutionary force despite its clearly dependent form. We have argued elsewhere that this seeming syntactic deviation can be accounted for in terms of featural import in syntactic positions (notably in Comp).

## References

- Abraham, Werner. 2013. Zur grammatischen Grundlegung von Modalität – semantisch-syntaktische Affinitäten zu nominaler Referenz, Aspekt und Quantifikation. In Werner Abraham & E. Leiss (eds.), *Funktionen von Modalität* (Linguistik – Impulse & Tendenzen/LIT 55), 25–76. Berlin: de Gruyter.
- Abraham, Werner. 2014. Strong modality and truth disposability in syntactic subordination: what is the locus of the phase edge validating modal adverbials? *Studia Linguistica* 69 (3). 1–41.

---

<sup>5</sup> Apparently, this is in line with the fact that imperative functions can go along with infinitives (*Arbeiten!* ‘to work’) or (erstwhile) past participles appear to match with the function of predicative attributes (*Gescheit!* ‘smart!’) or hearsay evidentials (*schon gehört?* ‘already heard of’) (cf. Evans 2007: 430). However, the question of finiteness is only of subordinate relevance for determining ST status and so is assertivity and, as a consequence, presuppositionality. Whether this is the result of diachronic processes of elliptical reduction of syntactic matrix portions or idiomaticization appears to be irrelevant for at least two reasons: If the present position is correct, intonation and accent parameters have the force of unmistakable disambiguation (e.g. exclamative and question coded under the identical form, but singled out through different prosody markers). This, in turn, is hard to come by in ancient written documents. Given the alternative which position of this alternative outweighs the other, we give preference to what we witness on today’s phenomena. It is relevant to see what follows from this: for example, on the question whether the phenomena under discussion are a sort of backwash of the conservative view on the direction of grammaticalization.

- Altmann, Hans (ed.). 1988. *Intonationsforschung* (Linguistische Arbeiten 200). Tübingen: Niemeyer.
- Altmann, Hans, Anton Batliner & Wilhelm Oppenrieder (eds.). 1989. *Zur Intonation von Modus und Fokus im Deutschen* (Linguistische Arbeiten 234). Tübingen: Niemeyer.
- Beckman, M., M. Díaz-Campos, J. T. McGory & T. A. Morgan. 2002. Intonation across Spanish, in the Tones and Break Indices framework. *Probus* 14. 9–36.
- Elvira-García, Wendy. 2015. Prosody as a phonological cue for differentiating between elliptical and insubordinated constructions in Spanish. In *SLE 2015 Leiden Book of Abstracts*.
- Evans, Nicholas. 2007. Insubordination and its uses. In Irina Nikolaeva (ed.), *Finiteness: theoretical and empirical foundations*, 366–431. Oxford: OUP.
- Groenendijk, Jeroen & Martin Stokhof. 1982. Semantic analysis of WH-complements. *Linguistics and Philosophy* 5. 175–233.
- Groenendijk, Jeroen & Martin Stokhof. 1997. Questions. In Johan van Benthem & Alice ter Meulen (eds.), *Handbook of logic and language*, 1055–1124. Amsterdam, North-Holland.
- Gunlogson, Christine. 2003. *True to form: rising and falling declaratives as questions in English*. New York: Routledge.
- Gussenhoven, Carlos & Aoju Chen. 2000. Universal and language-specific effects in the perception of question intonation. In *INTERSPEECH/ICSLP 2000 (6)*, 91–94. Beijing.
- Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9. 183–221.
- Higginbotham, J. 1996. The semantics of questions. In S. Lappin (ed.), *The handbook of contemporary semantic theory*. Oxford: Blackwell.
- Kaiser, Sebastian & Volker Struckmeier. 2015. *When insubordination is an artefact (of sentence type theories)*. Talk and Handout SLE-Leiden 2015.
- Lohnstein, H. 2000. *Satzmodus – kompositionell. Zur Parametrisierung der Modusphrase im Deutschen* (Studia Grammatica 49). Berlin, New York: Akademie Verlag.
- Lohnstein, H. 2007. On clause types and sentential force. *Linguistische Berichte* 209. 63–86.
- Meibauer, Jörg, Markus Steinbach & Hans Altmann (eds.). 2013. *Satztypen des Deutschen*. Berlin: De Gruyter.
- Struckmeier, Volker. 2010. Ja doch wohl C. Modal particles in German as C-related elements. *Studia Linguistica* 68 (1). 16–48.
- Thurmair, Maria. 2013. Satztyp und Modalpartikeln. In Jörg Meibauer, Markus Steinbach & Hans Altmann (eds.), *Satztypen des deutschen*. Berlin: De Gruyter.



## Chapter 2

# Two-layer networks, non-linear separation, and human learning

R. Harald Baayen

University of Tübingen

Peter Hendrix

University of Tübingen

Ever since the criticism of the perceptron by Minsky & Papert (1969), two-layer networks have been regarded as far too restricted for classification tasks requiring more than the simplest linear separation. We discuss an example of a classification task that in  $\mathbb{R} \times \mathbb{R}$  is not only not linearly separable, but also not non-linearly separable. Yet, this classification task can be carried out with error-free performance. To do so, it is mandatory to step outside the box of  $\mathbb{R} \times \mathbb{R}$ , and we discuss how several state-of-the-art methods from machine learning achieve this. We also show that a two-layer network that makes use of the learning rule of Rescorla and Wagner (1972) can solve this classification task, with different degrees of success (up to 100% accuracy) depending on the representations chosen for the input units. The excellent classificatory performance of our two-layer network helps explain why wide learning with two-layer networks with thousand and even tens of thousands of input and output units is so successful in predicting aspects of human implicit learning, including the consequences of trial-by-trial learning for response latencies in the visual lexical decision task.

## 1 Introduction

Several computational modeling studies suggest that two-layer networks with connection weights estimated with the learning rule of Rescorla & Wagner (1972) capture non-trivial aspects of lexical processing. In what follows, these networks will be referred to as ‘wide learning’ networks, as they typically comprise just two layers with, however, many tens of thousands of units.

Baayen et al. (2011) observed that the activations of lexical output units (conceptualized as pointers to semantic vectors in Milin et al. (2016)) in wide learning networks with sublexical input units (e.g., letter bigrams) closely mirrored reaction times in the

visual lexical decision task. Regression models, one fitted to the reaction times, and one fitted to the reciprocally transformed activations, produced very similar results. The same predictors reached significance, with very similar relative effect sizes. Even though the network did not have any form units for morphemes or words, it correctly predicted the facilitatory effects of constituent frequency and word frequency typically observed for English. When the same kind of network was trained on Vietnamese, it correctly predicted the inhibitory effect of constituent frequency that surprisingly characterizes compound reading in this language (Pham & Baayen 2015). These results show that wide networks with sublexical input units capture frequency effects that, depending on low-level orthographic distributional structure, can work out in very different ways, facilitatory in English but inhibitory in Vietnamese. Wide learning networks have also been found to provide superior prediction for the brain's electrophysiological response to linguistic stimuli (Hendrix, Bolger & Baayen 2016) and the details of eye-movements during reading (Hendrix 2015).

In classical accounts of lexical processing, the presence of a frequency effect is treated as a litmus test for the existence of form representations. For instance, a frequency effect for complex words is taken as proof of the existence in the mind of form representations for complex words. Typically, such representations are associated with resting activation levels that are assumed to depend on frequency of use, and that are assumed to underlie the frequency effects observed in tasks tapping into lexical processing. However, theories that posit such form units have to explain how such units are accessed. This question seldom is reflected on, probably because we are so familiar with being able to look up words in a dictionary, or to search for patterns in files, that we take for granted that accessing units is trivial. However, models such as the interactive activation model (McClelland & Rumelhart 1981) were developed precisely because human look-up has all kinds of properties that are foreign to look-up with the algorithms implemented on our computers. Wide learning networks offer an alternative algorithm that, like the interactive activation model, targets an algorithmic approximation of human look-up. Importantly, frequency effects (and also similarity effects) come for free with wide learning, and arise as a consequence of continuous error-driven optimization of lexical discrimination. There no longer is a need for positing counters in the head such as resting activation levels.

Current research is revealing that a range of quantitative measures derived from the connection weights of wide learning networks also generate precise predictions about trial-to-trial learning in the visual lexical decision task. Figure 1 presents three measures of goodness of fit, the AIC, the ML score, and the (adjusted)  $R$ -squared, for generalized additive models fitted to the data of a participant contributing responses to the British Lexicon Project (Keuleers et al. 2012). Goodness of fit is worst for a classical model with as predictors the conventional measures of frequency of occurrence and neighborhood density. Model predictions improve when these classical measures are replaced by measures derived from the  $15,106 \times 30,117$  weight matrix of a network trained on the British National Corpus. Performance is best when this network is allowed to continue learning as it is presented with words and nonwords, made available in exactly the same order as in the lexical decision experiment.

2 Two-layer networks, non-linear separation, and human learning

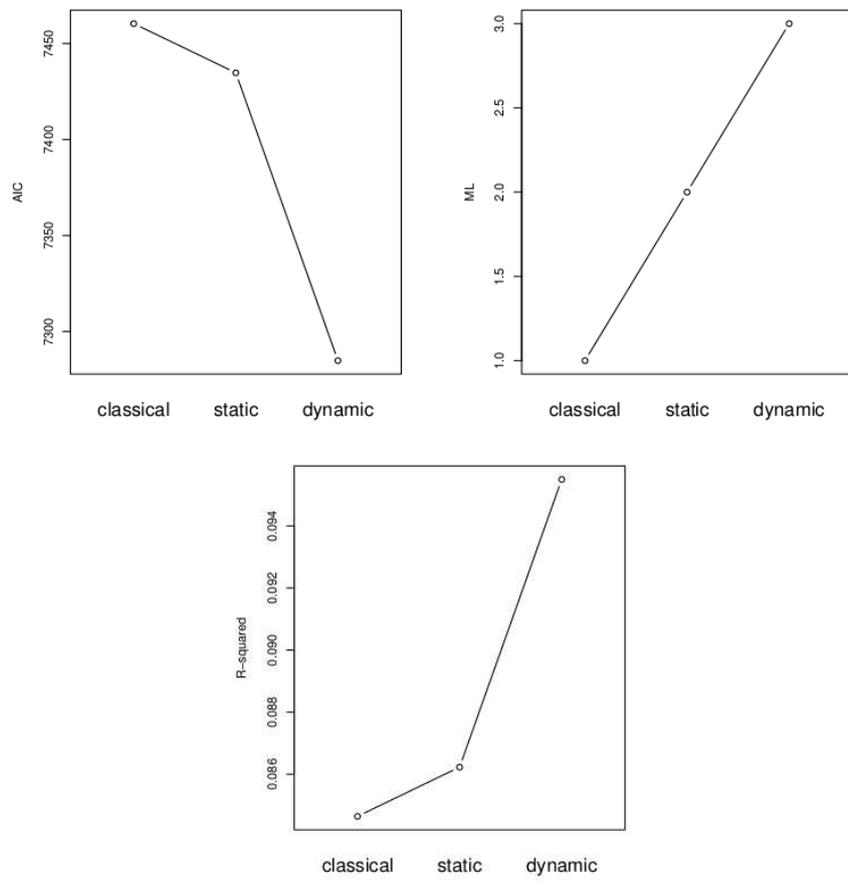


Figure 1: Three measures of goodness for the GAMs fitted to lexical decision reaction times.

These results indicate that wide networks with the Rescorla-Wagner learning rule provide a useful computational window on human lexical processing, complementing the strong support for this learning rule in the literature on animal learning (Siegel & Allan 1996) and more recently also in computational evolutionary biology (Trimmer et al. 2012).

However, in the light of the criticism by Minsky & Papert (1969) of simple two-layer perceptrons as being incapable of approximating a wide range of useful functions, the excellent performance of wide learning networks is surprising. Would performance have been better with deep learning, or with Bayesian updating? Is a two-layer network, however wide, actually too simple to be taken seriously for the computational modeling of lexical processing?

In what follows, this issue is addressed by investigating a simple but non-trivial classification problem and comparing the performance of wide learning with three state-of-the-art classifiers: support vector machines, deep learning, and gradient boosting machines.

## 2 A non-linearly separable classification problem

The left panel of Figure 2 presents a classification problem that is not only not linearly separable, but also not non-linearly separable. In a grid of  $50 \times 50$  pixels, 260 (highlighted in black) belong to class *A* and the remaining 2240 (in gray) belong to class *B*. In the  $\mathbb{R} \times \mathbb{R}$  input space, this classification problem cannot be solved by means of linear or non-linear boundary functions. There is no straight line that separates the grey dots from the black dots, nor is there a sensible curve that would achieve separation.

Three machine learning techniques were applied to this classification task. Deep learning, using the `h2o` package (Fu et al. 2015), which provides an adaptive learning rate per neuron as well as regularization through shrinkage and dropout, with 1 layer of 100 hidden units, reached an accuracy of 99.4%. A gradient boosting machine (fit with 20 trees with a maximum tree depth of 20, using `xgboost` package (Chen, He & Benesty 2015)) provided perfect classification with only minor deterioration under 10-fold cross-validation. A support vector machine (using `svm` in the `e1071` package (Meyer et al. 2015), with a second-order polynomial as a kernel) performed quite well on the full data, but performance dropped below that of the other two methods under 10-fold cross-validation (cf. Table 1). The success of the support vector machine indicates that there exists a transformation of the  $\mathbb{R} \times \mathbb{R}$  input space in which the two classes of data points are to a considerable extent linearly separable.

A very different transformation of the input space is achieved by moving from coordinates in  $\mathbb{R} \times \mathbb{R}$  to one-hot encoding for rows and columns, resulting in two sets of 50 units representing row and column identifiers. Moving to this binary 100-dimensional representation (henceforth  $\mathbb{B}^{100}$ ) allows all three abovementioned machine learning techniques to achieve perfect classification on the full data set, and to retain a high accuracy under 10-fold cross-validation (cf. Table 1).

A wide learning network with as cues the row and column identifiers performs,

## 2 Two-layer networks, non-linear separation, and human learning

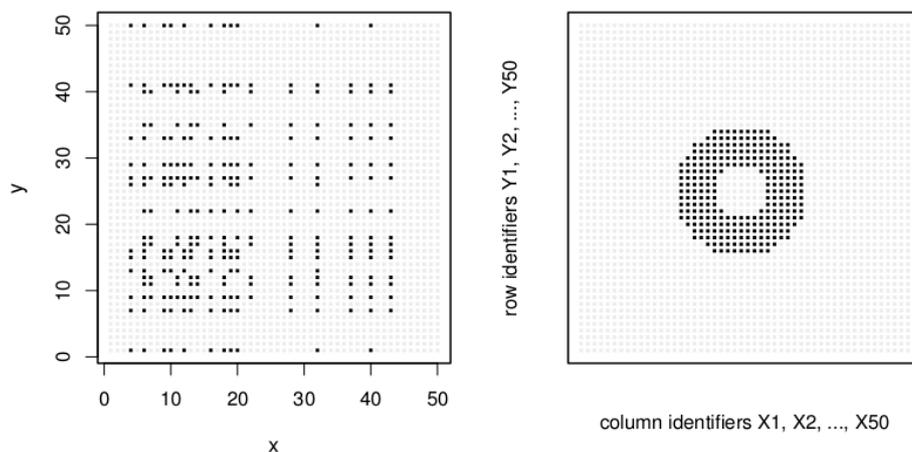


Figure 2: A non-linear classification problem. Left panel: points in a Cartesian grid ( $x = 1, 2, \dots, 50; y = 1, 2, \dots, 50$ ). Right panel: the same points in a 100-dimensional space using one-hot encoding, with re-arranged rows and columns.

Table 1: Accuracy for four algorithms, applied to the complete data set and with 10-fold cross-validation, with and without re-ordering of rows and columns. deep: deep learning; gbm: gradient boosting machine; svm: support vector machine; wide: 2-layer wide learning network.

Accuracy		Method	Input Space
complete	cv-10		
0.994	0.986	deep	$\mathbb{R}^2$
1.000	0.994	gbm	$\mathbb{R}^2$
0.982	0.896	svm	$\mathbb{R}^2$
1.000	0.994	deep	$\mathbb{B}^{100}$
1.000	0.994	gbm	$\mathbb{B}^{100}$
1.000	0.949	svm	$\mathbb{B}^{100}$
0.960	0.950	wide	$\mathbb{B}^{100}$
1.000	0.989	wide	hub features

with a single pass through the data, with 96.0% accuracy ( $F$ -score 0.81, precision and recall both 0.81), with the expected decrease in performance under 10-fold cross-validation (accuracy: 95.0%;  $F$ -score 0.71, precision 0.88, recall 0.61).<sup>1</sup> Although

<sup>1</sup> As wide learning does not make use of nonlinear activation functions at the output layer to obtain a

clearly lagging behind the gradient boosting machine and the deep learning network, it is not the case that wide learning is a total failure — to the contrary, it gets quite far, performing better under 10-fold cross-validation than the support vector machine.

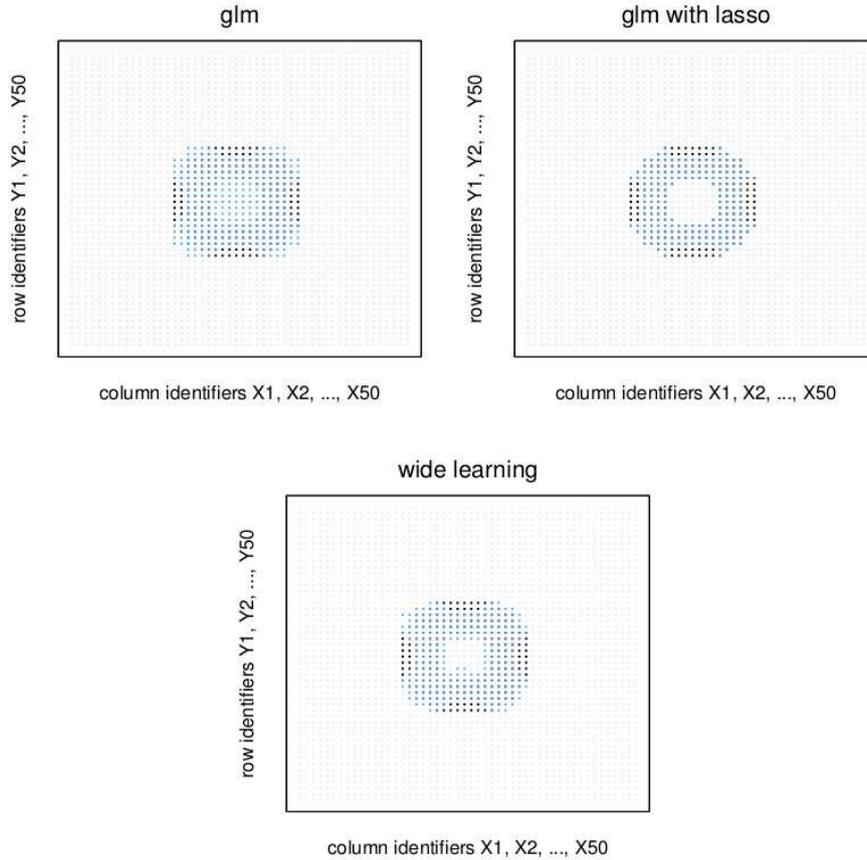


Figure 3: Classification performance for the generalized linear model (left), the generalized linear model with lasso (right), and wide learning (bottom) predicting class from row and column identifiers.

To appreciate better what wide learning achieves, first note that the move from  $\mathbb{R} \times \mathbb{R}$  to  $\mathbb{B}^{100}$  renders the classification problem invariant to exchanges of pairs of rows, and to exchanges of pairs of columns. One re-arrangement of rows and

---

firing versus not firing response, evaluation of model performance proceeded by collecting the activations for all 2500 pixels, and setting a threshold such that the  $k$  pixels with the highest support for class  $A$ , where  $k$  is the cardinality of  $A$ , are assigned to class  $A$ . The same threshold was used under cross-validation.

## 2 Two-layer networks, non-linear separation, and human learning

columns results in a configuration with all points of class  $A$  arranged in a circular band, as shown in the right panel of Figure 2. This rearrangement, possible thanks to ‘domain knowledge’, shows that there is considerably more structure in the data than is apparent to the eye in the scatter in the left panel of Figure 2.

Now consider Figure 3. The top left panel presents the predictions (in blue) of a generalized linear model (top left), a generalized linear model with lasso correction (top right), and a wide learning network (bottom). Each model was asked to predict the class of a data point from its row and column identifier. A logistic generalized *linear* model correctly detected that the points belonging to class  $A$  are located within a *circle*, but failed to exclude the points in its center, and lacked precision at the four outer edges of the circular band. Importantly, this linear model achieves considerable separation of the two classes in  $\mathbb{B}^{100}$  that, if transformed back into  $\mathbb{R} \times \mathbb{R}$ , would classify as non-linear.

Improved classification accuracy can be obtained by shrinking the  $\beta$  coefficients of the GLM through lasso ( $\ell_1$ -norm) regularization (using `glmnet`, Friedman, Hastie & Tibshirani 2010, run with `maxit = 300`). The center panel shows that all points in the inner disk are now correctly assigned to class  $B$ . Yet, the model remains somewhat imprecise at the edges.

The bottom panel of Figure 3 illustrates the performance of wide learning, which succeeds in correctly assigning most points in the inner circle to class  $B$ , while reducing slightly the imprecision at the edges that characterizes the GLM with lasso regularization. Again, we see that a technique that is known to be restricted to linear separation in  $\mathbb{R} \times \mathbb{R}$  achieves good separation in  $\mathbb{B}^{100}$ .

Of special interest is that this classification performance is achieved without knowledge of the topology of the circular band. All that is available to the models is, for each data point, the identifiers for its row and column. In other words, we can re-arrange the rows and columns back into the scatter of the left panel of Figure 2, and a majority of data points would still be correctly classified. This shows that the representation of the problem in  $\mathbb{R} \times \mathbb{R}$  is a highly specific one that is restricted to a unique configuration of data points, whereas the representation in  $\mathbb{B}^{100}$  covers the full set of  $50! \times 50!$  permutations of rows and columns. The classification accuracy of the GLM and of the wide learning network is exactly the same for all these alternative configurations.

However, the accuracy of wide learning can be improved considerably by making use of the fact that the re-arrangements share underlyingly the topology of the circular band. This topology makes it possible to do error-free classification with just four features. Let a data point be a hub if all of its eight surrounding data points belong to class  $A$ , and let a data point be a hub neighbor if at least one of the eight surrounding data points is a hub. We now define four features, IS A HUB, IS NOT A HUB, IS A HUB NEIGHBOR, and IS NOT A HUB NEIGHBOR. When each data point is characterized by the values of these four features, a wide learning network yields error-free classification performance with a single pass through the data. Under leave-one-out cross-validation performance remains error-free. Ten-fold cross-validation with hub features requires special care, as missing data make it impossible to maintain the cri-

terion that a hub should have exactly 8 neighbors. When the neighbor count for a hub is relaxed to 7 during training and to 4 during testing, accuracy remains at 99% ( $F$ -score 0.95, precision 0.94, recall 0.96).

Figure 4 presents two further classification tasks for which wide learning with hub features performs with a very high accuracy. The pattern in the left panel was explicitly characterized by Minsky & Papert (1969) as impossible for perceptrons to classify, which is correct when the problem is formulated in  $\mathbb{R} \times \mathbb{R}$ , but not necessarily true when the problem is reformulated in other spaces. A wide learning network with hub features solves this classification problem in its stride, with error free performance also under leave-one-out cross-validation, but just the representation in  $\mathbb{B}^{100}$  already allows for a classification accuracy no less than 96.9% ( $F$ -score 0.88, precision 0.89, recall 0.87).

Interesting is also the ‘open cross’ task in the right panel of Figure 4. We failed to obtain sensible classification performance for  $\mathbb{R} \times \mathbb{R}$  and for  $\mathbb{B}^{100}$  under cross-validation with gradient boosting machines and support vector machines (all points assigned to the ‘baseline’ class  $B$ ). Deep learning on  $\mathbb{R} \times \mathbb{R}$  with a two layers of hidden layer, the first with 100 units and the second with four units, performed much better (accuracy 94.2%, 93.4% under 10-fold cross-validation), but upon inspection systematically assigned all class  $B$  data points within the open squares that build the cross to class  $A$  under cross-validation. Deep learning on  $\mathbb{B}^{100}$  was a total failure ( $F$ -score = 0.28 under 10-fold cross-validation). Wide learning in  $\mathbb{B}^{100}$  failed miserably as well ( $F$ -score 0.19), but wide learning with hub features was highly effective, with accuracy above 99% both for the full data set, as well as under leave-one-out cross-validation.

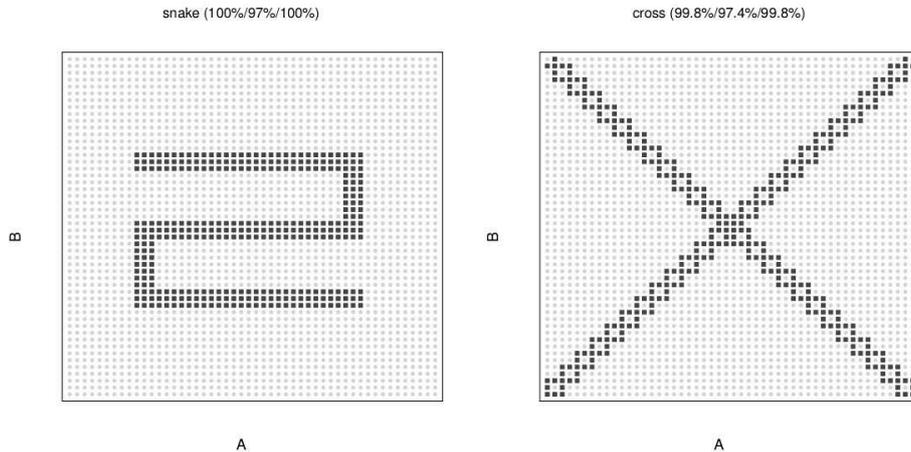


Figure 4: Two further non-linear classification set-ups. Accuracies in parentheses for wide learning with hub features for the full data set, for 10-fold cross-validation, and for leave-one-out cross-validation.

### 3 Discussion

To solve classification problems in  $\mathbb{R} \times \mathbb{R}$  that are not linearly separable and also not non-linearly separable, it is crucial to step outside the box. Gradient boosting machines achieve this by sidestepping the problem of finding a boundary function, and let classification trees do the best they can with local splits, while letting later trees deal with the classification errors of earlier trees. Support vector machines step outside the box by projecting the data points into a higher-dimensional space in which the classes are linearly separable. With a hundred hidden units, deep learning also finds a solution, thanks to regularization through shrinkage and dropout. (A three-layer network with 100 hidden units trained simply with backpropagation fails to assign any datapoint to the *A* class.) These 100 hidden units constitute a new space that re-represents the original  $\mathbb{R} \times \mathbb{R}$  space in such a way that the last two layers of the three-layer network can achieve excellent (linear) separation of the two classes.

Moving from a representation in  $\mathbb{R} \times \mathbb{R}$  to a representation with row and column identifiers is yet another way of stepping outside the box. For wide learning with the Rescorla-Wagner rule, this re-representation is a necessary step because input units are restricted to discrete feature detectors that are either on or off. Given this re-representation, wide learning can achieve considerable separation of data points that are not even non-linearly separable, and in this mirrors the performance of a logistic generalized linear model. But deep learning and the support vector machine also thrive with this re-representation, reaching 100% accuracy on the full data and improved cross-validation scores. Because this re-representation is invariant to order, re-arranging row and column identifiers allows the underlying topology to emerge, which in turn makes an even simpler re-representation with hub features possible.

The present results clarify that it does not make much sense to suppose that a non-trivial wide learning network (such as the abovementioned network with 15,106 input units and 30,117 output units) is handicapped by being limited to ‘linear separation’. This handicap holds for  $\mathbb{R}^n$ , but by re-representing the classification problem in some higher factorial space  $\mathbb{B}^{n+m}$ ,  $m \gg n$ , a wide learning network can achieve separation that, albeit linear in  $\mathbb{B}^{n+m}$ , would count as non-linear when projected back into  $\mathbb{R}^n$ . It follows that what a two-layer wide learning network can or cannot achieve depends crucially on the input representations. Deep learning networks can discover good input representations at their (final) hidden layer, but hand-crafted representations building on domain knowledge can also be highly effective.

Given the strong support for the Rescorla-Wagner learning rule in the literature on animal learning (Siegel & Allan 1996) and evolutionary biology (Trimmer et al. 2012) and its success in predicting details of human lexical processing with input and output features that have a clear and transparent linguistic interpretation, we think it makes sense to delve deeper into the benefits of going wide for understanding human error-driven learning.

## References

- Baayen, R. H., P. Milin, D. Filipović Durdević, P. Hendrix & M. Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.
- Chen, T., T. He & M. Benesty. 2015. *xgboost: eXtreme Gradient Boosting*. R package version 0.4-0. <https://github.com/dmlc/xgboost>.
- Friedman, J., T. Hastie & R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1). 1–22.
- Fu, A., S. Aiello, A. Rao, A. Wang, T. Kraljevic & P. Maj. 2015. *h2o: H2O R Interface*. R package version 2.8.4.4. <https://CRAN.R-project.org/package=h2o>.
- Hendrix, P. 2015. *Experimental explorations of a discrimination learning approach to language processing*. University of Tübingen PhD thesis.
- Hendrix, P., P. Bolger & R. H. Baayen. 2016. Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle & M. Brysbaert. 2012. The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods* 44(1). 287–304.
- McClelland, J. L. & D. E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: part I. An account of the basic findings. *Psychological Review* 88. 375–407.
- Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel & F. Leisch. 2015. *e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien*. R package version 1.6-7. <https://CRAN.R-project.org/package=e1071>.
- Milin, P., M. Ramscar, R. H. Baayen & Laurie Beth Feldman. 2016. Discrimination in lexical decision. *Manuscript submitted for publication*.
- Minsky, M. & S. Papert. 1969. *Perceptrons: An introduction to computational geometry*. Cambridge, MA.
- Pham, H. & R. H. Baayen. 2015. Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience* 30(9). 1077–1095.
- Rescorla, R. A. & A. R. Wagner. 1972. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (eds.), *Classical conditioning II: current research and theory*, 64–99. New York: Appleton Century Crofts.
- Siegel, S. & L. G. Allan. 1996. The widespread influence of the rescorla-wagner model. *Psychonomic Bulletin & Review* 3(3). 314–321.
- Trimmer, P. C., J. M. McNamara, A. I. Houston & J. A. R. Marshall. 2012. Does natural selection favour the Rescorla-Wagner rule? *Journal of Theoretical Biology* 302. 39–52.

## Chapter 3

# John's car repaired. Variation in the position of past participles in the verbal cluster in Dutch

Sjef Barbiers

University of Leiden

Hans Bennis

Meertens Institute; University of Amsterdam

Lotte Dros-Hendriks

Meertens Institute

Although sentence final verbal clusters in dialects of Dutch demonstrate a large amount of variation in the order of verbs, we argue that this is only apparently so. We take each dialect to allow just one order of verbs in three-verb clusters with a past participle. In the north of the Dutch language area, the order is descending (V3-V2-V1) and the rest of the dialects show an ascending order (V1-V2-V3). The large amount of apparent counterexamples will be explained by independently motivated, interfering properties. First, participles might be V-type or A-type. Only V-type participles occur in V-positions in the verbal cluster. Secondly, non-verbal elements (such as A-type participles) may interrupt a verbal cluster. We will show that the distribution of the different orders in dialects of Dutch strongly supports such a restrictive approach. We thus take this to be an argument that a structural approach to dialectology is required to gain insight in the properties of the formation of verbal clusters in Dutch.

## 1 Introduction<sup>1</sup>

In a famous article from 1954, Weinreich poses the question: is a structural dialectology possible? In this article we answer that question positively with the discussion

---

<sup>1</sup> We have presented parts of this paper at conferences in Amsterdam, Utrecht and Gent. We thank the audiences for their useful comments. We also like to thank Erik Tjong-Kim-Sang for his assistance with Map 2.

of a construction in which a structuralist approach will provide insights in the properties of dialects themselves; moreover, such an approach to dialects will lead us to answers to more general questions with respect to the organization of the syntactic system as a whole. We will concentrate our discussion on the phenomenon of participle placement in verbal clusters. We will demonstrate that a structural dialectological approach will provide us with a new perspective on the syntax of past participles. Consequently, it is not only the case that structural dialectology is possible, it turns out that structural dialectology is necessary to understand the grammar of Dutch.

In a recent paper (Barbiers, Bennis & Dros-Hendriks submitted) we argue that verbal clusters in varieties of Dutch show either a strictly descending or ascending order of verbs. Northern dialects (Friesland, Groningen) have a descending order, whereas the other varieties of Dutch, both in the Netherlands and Flanders only allow ascending orders. Let us illustrate this with the following sentence from the Syntactic Atlas of the Dutch Dialects (SAND; Barbiers et al. (2008)).

- (1) John weet dat hij voor drie uur de wagen moet hebben gemaakt.  
John knows that he before three o'clock the car must<sub>1</sub> have<sub>2</sub> made<sub>3</sub>

The verbs at the end of this sentence appear in the ascending order: the hierarchically highest verb comes first and the lowest comes last. We argue that this order is the only available order of verbs in Dutch varieties, except in the northern ones. The northern varieties have a descending order, as demonstrated in (2).

- (2) John weet dat hij voor drie uur de wagen gemaakt hebben moet.  
John knows that he before three o'clock the car made<sub>3</sub> have<sub>2</sub> must<sub>1</sub>

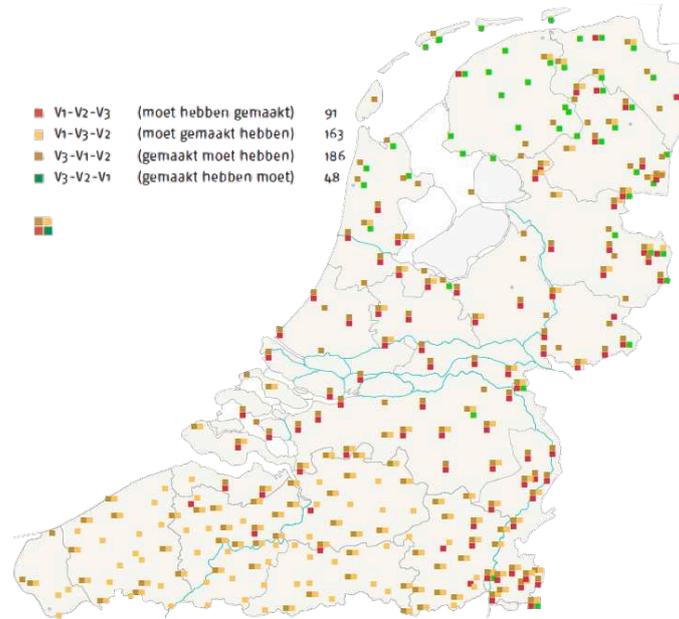
The problem with this perspective is that linguistic reality does not appear to support our theoretically motivated, structurally based hypothesis. It seems to be the case that a structural approach and a dialectological approach diverge. This is shown on the map of these sentences in Map 1.

Looking at the distribution of orders in the verbal cluster we draw the conclusions given in (3).

- (3) i. V2-V1-V3 is absent;  
ii. V2-V3-V1 is absent as well;  
iii. V1-V3-V2 is the dominant order in the Belgian part of the language area;  
iv. V3-V2-V1 is the typical order in the northern part of the language area;  
v. V1-V2-V3 is restricted to the Netherlands part of the language area. It is never the only order in a particular location;  
vi. V3-V1-V2 is found in the whole language area except Friesland. It is the most frequent order and often occurs as the only order in specific dialects.

There is a large gap between our hypothesis and the distribution of orders that are found on Map 1. In 139 instances (91x V1-V2-V3 and 48x V3-V2-V1), the facts support our theory. However, in 349 instances (163x V1-V3-V2 and 186x V3-V1-V2)

### 3 Variation in Dutch past participle position



Map 1: SAND-II map 17b.

our hypothesis is not corroborated. It implies that 72% of the data contradict our theory. Nevertheless we will argue in the remainder of this paper that our ascending/descending hypothesis is supported by the data and that the hypothesis provides new insights into the theory of verbal clustering.

## 2 The categorial status of past participles

We know that participles are ambiguous with respect to their categorial status. They show up in verbal or adjectival contexts. Participles appear in attributive position in noun phrases, in contrast to infinitival verbs. The noun phrase *de verslagen vijand* ‘the beaten enemy’ is perfectly fine, but the noun phrase *de verslaan vijand* ‘the beat enemy’ is strongly ungrammatical.<sup>2</sup> In some cases there is an interpretative difference between adjectival and verbal participles (Kraak & Klooster 1968). A participle such as *geopend* can be interpreted as ‘open’ or ‘has been opened’. In a verb cluster as in (4a, [V2-V1]), the participle indeed allows both meanings of *geopend*. However, in the other order (4b, [V1-V2]) the participle can only be interpreted as verbal, with

<sup>2</sup> Similarly, participles do appear in adverbial position, as in *de vijand zat verslagen op de grond* ‘the enemy sat on the ground beaten’, but infinitives do not. As opposed to bare infinitives, to-infinitives do occur in attributive positions in Dutch, as in *de te bellen kandidaten* ‘lit. the to call candidates, the candidates that need to/can be called’ showing that the presence of the infinitival marker *te* ‘to’ may correspond to a categorial difference (cf. van Riemsdijk 1982; Bennis 1990).

the interpretation ‘has been opened’.

- (4) a. John zag dat de deur geopend<sub>2</sub> is<sub>1</sub>.  
John saw that the door opened is  
‘John saw that the door has been opened / is open.’
- b. John zag dat de deur is<sub>1</sub> geopend<sub>2</sub>.  
John saw that the door is opened  
‘John saw that the door has been opened / \*is open.’
- c. de geopende deur  
the opened door  
‘the door that has been opened / the open door’

Participles in attributive position within nominal phrases allow both interpretations, as is demonstrated in (4c). Apparently the adjectival position of the participle in (4c) allows a verbal, passive interpretation (‘has been opened’) and an adjectival, stative (‘open’) interpretation.<sup>3</sup> We thus conclude that there are two types of past participles: A-type participles that show up in adjectival position and allow both a verbal and an adjectival interpretation, and V-type participles that are exclusively verbal, both in position and in interpretation.

The fact that only the passive interpretation is available in (4b) can now be accounted for by assuming that the participle in (4b) is a V-type participle rather than an A-type one, thereby excluding the stative interpretation (‘open’). Given that both interpretations are available in (4a), we conclude that the participle in cluster-initial position is an A-type participle, just as the participle in (4c). The difference in interpretation between (4a) and (4b) is thus related to a categorial difference. In (4a) the participle is or may be an A-type, whereas it has to be a V-type in (4b). This is supported by restrictions on modification, e.g. the durative adverbial *de hele dag* ‘the whole day’ is possible with the A-type variant of (4a) but not easily with the V-type variant in (4b).<sup>4</sup>

If we analyse participles as being ambiguous between a V-type and an A-type, we are in a position to provide an answer to the fact that V3-V1-V2 occurs frequently in the Dutch language area although this particular cluster is theoretically predicted not to occur. Non-verbal elements generally appear to the left of the verb in Dutch clauses since Dutch shows an OV-order. Given their (partly) non-verbal properties, we take A-type participles to be non-verbal, and thus to occur to the left of a verbal cluster, just as other non-verbal elements.

The occurrence of a participle in front of the auxiliary verb is possible in the whole language area in two verb constructions (*participle-V*; cf. SAND-II, map 16). Apparently, A-type status of the participle is a common phenomenon in Dutch dialects. This would then lead us to expect that the order *participle-V1-V2* will show up in the whole language area as well. This is indeed the case with the exception of the

<sup>3</sup> This stative interpretation is known in the literature as a target state. Cf. Koeneman, Lekakou & Barbiers (2011) for recent discussion, diagnostics and references.

<sup>4</sup> More precisely, there is coercion such that *de hele dag* in (4b) has a repetitive, not a durative interpretation, as expected.

northern part of the language area. Moreover, it can be observed on Map 1 that the order V1-V2-V3 is accompanied by a participle-initial order in all locations.<sup>5</sup> In order to have both interpretive possibilities for the participle, the initial position must be available. As we have seen, the V-type status reduces the interpretive possibilities of the participle. Consequently, clusters with a participle in a cluster-final, verbal position are expected to constitute a subset of clusters with participles in a non-verbal position.

We thus analyse the V3-V1-V2 order in this construction as an instance of the *participle<sub>A</sub>-V1-V2* order, and this order is consequently no longer a problem for the theory. If the participle is a V-type, it will show up in the V1-V2-V3 order as the rightmost element. We thus have eliminated the problem of V3-V1-V2 orders as counterexamples to our hypothesis. There are no V3-V1-V2 clusters. In those cases the ascending V1-V2 cluster is preceded by an A-type participle.

If we now turn to the geographical distribution of these sentences, we observe that a participle can have an A-type status in the whole language area. It is rather the question where it may show up as V-type in the order V1-V2-V3. It is clear that there exists a strong preference for A-type participles in the Belgian part of the language area, whereas the Dutch part shows an ambiguity in categorial status. For the northern area it is difficult to determine what the status of the participle is. In the order *participle-V2-V1* the participle can be A-type, as is a possibility in the rest of the language area, but it may also be V-type since the northern part of the language area has a descending strategy in verbal clusters.

### 3 Cluster interruption

We are now in a position to turn to the verb cluster order V1-V3-V2. We have argued that participles in Dutch are ambiguous in having a V-type or A-type status. We argued that a V-type participle would give rise to the order V1-V2-*participle<sub>V3</sub>* in (5a), whereas an A-type participle would be ordered to the left of the verbs, and thus leads to the order *participle<sub>A</sub>-V1-V2* as in (5b). The northern order is *participle-V2-V1* in (5c). However, the order V1-*participle-V2* in (5d) occurs quite often (n = 163) as well, especially in the southern part of the language area (see map 1).

- (5) a. ...dat John de wagen voor drie uur *moet hebben gemaakt*. [V1-V2-pcp] (V-type)  
 b. ...dat John de wagen voor drie uur *gemaakt moet hebben*. [pcp-V1-V2] (A-type)  
 c. ...dat John de wagen voor drie uur *gemaakt hebben moet*. [pcp-V2-V1] (V-type or A-type in northern varieties)  
 d. ...dat John de wagen voor drie uur *moet gemaakt hebben*. [V1-pcp-V2]  
 ‘...that John the car before three o’clock must have made.’

<sup>5</sup> V3-V2-V1 in the north, V3-V1-V2 in the rest of the language area. These orders can be analyzed as occurrences of A-type participles.

Given the fact that the participle can be A-type or V-type, there are two ways to analyse the occurrence of the order *V1-participle-V2* in (5d). If the participle would be V-type, we face a problem for our approach since we predict the order *V1-V3-V2* not to occur since it involves a non-uniform order, i.e. not a strictly ascending or descending order. Alternatively, we may take the participle to be of the A-type and argue that A-type participles may show up in a cluster in between two verbs. In order to further support our order hypothesis, it will be clear that we will take the latter approach.

The fact that verb clusters can be interrupted by non-verbal material has received a lot of attention in the literature. Varieties of Dutch differ with respect to the amount and the nature of the material they allow to appear within a verb cluster. Most varieties allow verb particles to appear in the cluster, as is shown in (6). These particles may be prepositional, adjectival or adverbial in nature.

- (6) a. Ik vind dat John Marie *moet OP bellen*. [part = P]  
 I find that John Marie must up call  
 'I think that John should call Mary.'
- b. Ik vind dat John die mug *moet DOOD slaan*. [part = Adj]  
 I find that John that mosquito must dead beat  
 'I think that John should kill that mosquito.'
- c. Ik vind dat John die valse hond *moet WEG jagen*. [part = Adv]  
 I find that John that mean dog must away chase  
 'I think that John should chase away that mean dog.'

The capitalized elements are generally called *verb particles*. This label is just a way to describe a class of elements that together with the main verb constitute a complex verb, or rather a verbal predicate. There is no evidence for a syntactic category of the type Particle. There is no compelling evidence to consider particles as verbal prefixes either. Particles can be separated from the main verb in verb-cluster constructions (7a) and must be separated in clauses with Verb Second (7b). Moreover, they appear outside verbal inflection, as in the case of participles in which the particle shows up in front of the inflectional prefix *ge-* (7c).

- (7) a. Ik vind dat John Marie *OP moet bellen*.  
 I find that John Marie up must call
- b. John belt Marie *OP*. (\*John *OP*belt Marie)  
 John calls Mary up
- c. Ik vind dat John Marie *moet hebben OPgebeld*. (\**geOP*beld)  
 I find that John Marie must have called

The literature on Dutch particles is vast. The analyses can roughly be divided into lexical approaches in which verb and particle are part of a lexical verb (a.o. Neeleman & Weerman 1993; Neeleman 1994), syntactic approaches in which particles are generated as separate items in the VP (a.o. Hoekstra, Lansu & Westerduin 1987; Bennis 1991;

### 3 Variation in Dutch past participle position

den Dikken 1995), and hybrid proposals in which the particle-verb combination constitutes a syntactically complex word (Booij 2002; Blom 2005). All three approaches have theoretical and empirical problems. We will not enter into a detailed discussion of particles in this article. We just establish that this type of particle may easily be incorporated in a verb cluster in all Dutch dialects. This is also evident from the SAND (SAND II, maps 31a/b).

In this paper we will not discuss the properties of cluster formation either (see Barbiers, Bennis & Dros-Hendriks (submitted) for an elaborate discussion of this issue in terms of the structure building process Merge). The crucial fact is that non-verbal material of the particle-type may appear in between verbs within a verb cluster. In that respect these cases are in our view structurally similar to the occurrence of participles of the A-type within the verb cluster.

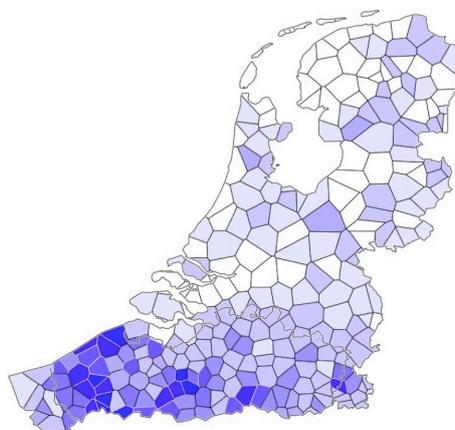
Not only particles and participles may appear in verb clusters. A whole range of other phrases show up in verb clusters in dialects of Dutch as well. A number of cases is given below.

- (8)
- |    |                                                                                                                                         |                        |
|----|-----------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| a. | Ik vind dat John <i>moet brood eten</i> .<br>I find that John must bread eat<br>'I think that John should eat bread.'                   | [bare noun / object]   |
| b. | Ik vind dat John <i>moet klein schrijven</i> .<br>I find that John must small write<br>'I think that John should write small.'          | [VP-adverb]            |
| c. | Ik vind dat John <i>moet boeken lezen</i> .<br>I find that John must books read<br>'I think that John should read books.'               | [plural noun / object] |
| d. | Ik vind dat John <i>moet een auto wassen</i> .<br>I find that John must a car wash<br>'I think that John should clean a car.'           | [indefinite object]    |
| e. | Ik vind dat John <i>moet op tafel staan</i> .<br>I find that John must on table stand<br>'I think that John should stand on the table.' | [prepositional phrase] |
| f. | Ik vind dat John <i>moet de meisjes zoenen</i> .<br>I find that John must the girls kiss<br>'I think that John should kiss the girls.'  | [definite object]      |

If interruption occurs, it is always optional. In the whole language area, all the sentences in (8) are perfectly fine in the order in which the interrupter precedes the cluster. It is clear that cluster interruption by non-verbal material of the construction types in (8) is basically confined to the Flemish area (West- and East-Flanders and the province of Flemish Brabant). We find some cases in which a bare noun occurs in the cluster-medial position in the Netherlands, but for reasons we do not yet understand, the remainder of constructions is geographically very much restricted to the southern part of the language area. However, the fact that the cluster interruption in (8) is predominantly a southern phenomenon ties in with the fact that we

have observed above that cluster interruption by type-A participles is the dominant order in verb clusters with participles in southern varieties of Dutch as well, as was evident from Map 1.

On the other hand, we have shown that cluster interruption is a general phenomenon of Dutch verb clusters given that particles are found within the verbal cluster in most varieties of Dutch and that interruption of A-type participles is quite often found in the Netherlands as well. It is striking that the possibilities to allow interruption slowly increase geographically in moving to the southwest (West-Flanders) of the language area. This is shown on Map 2.



Map 2: Cluster interruption - synthesis (= SAND-II map 30b + particles + participles).

On map 2, the color is getting darker the more interruption types (maximum is 8) a particular area accepts. The map shows that cluster interruption is increasing from north (the Frisian area) to south-west (the West-Flemish area).

From these data we conclude that interruption of a verb cluster is a possibility in almost all varieties of Dutch with the exception of the northern dialects. The extent to which interruption occurs is determined by two tendencies:

- 1) the more predicative the non-verbal element is, the more readily it appears as part of the verb cluster;
- 2) going in the direction of the southwest of the language area the preference for inclusion of non-verbal parts within the verb cluster increases.

The first tendency implies that non-predicative elements such as subjects, clitics, and sentence adverbials are generally not acceptable within verbal clusters and this is indeed the case. These tendencies do not seem to be determined by structural principles because (i) all descending varieties do allow cluster interruption in principle, (ii) there are no clear geographic borders between dialects that allow interruption and those that do not.

## 4 Conclusion

From the discussion above, we can conclude that the hypothesis that Dutch dialects have either a descending (northern varieties) or an ascending (rest of the language area) order in the verb cluster is supported notwithstanding superficial evidence to the contrary. The apparent counterexamples are due to two independent factors: participles can be either A-type or V-type, and non-verbal predicative constituents may interrupt the verbal cluster. The fact that the V2-V1-V3 order and the V2-V3-V1 order are lacking in the construction under discussion is further support for a structural analysis of the phenomenon of cluster formation. These orders are not ascending/descending and there are no independent structural factors that interfere. We have thus formulated a very restrictive theory which makes quite precise predictions on the occurrence of different orders and the geographic correlations between different instantiations of occurring orders. We have provided a clear example of structural dialectology. We thus answer Weinreich's question positively.

## References

- Barbiers, Sjef, Hans Bennis & Lotte Dros-Hendriks. Submitted. Merging verb cluster variation.
- Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther de Vogelaer & Margreet van der Ham. 2008. *The syntactic atlas of the Dutch dialects*. Vol. 2. Amsterdam University Press.
- Bennis, Hans. 1990. TI: a note on *modal passives*. In J. Mascaro & M. Nesport (eds.), *Grammar in progress*, 33–40. Foris Publication.
- Bennis, Hans. 1991. Theoretische aspecten van partikelvooropplaatsing II. *TABU* 21.
- Blom, Corrien. 2005. *Complex predicates in Dutch: synchrony and diachrony*. LOT Utrecht.
- Booij, Geert Evert. 2002. Constructional idioms, morphology, and the Dutch lexicon. *Journal of Germanic linguistics* 14(04). 301–329.
- den Dikken, Marcel. 1995. *Particles: on the syntax of verb-particle, triadic, and causative constructions*. Oxford University Press.
- Hoekstra, Teun, Monic Lansu & Marion Westerduin. 1987. Complexe verba. *Glott* 10. 61–78.
- Koeneman, Olaf, Marika Lekakou & Sjef Barbiers. 2011. Perfect doubling. *Linguistic Variation* 11(1). 35–75.
- Kraak, A. & W. G. Klooster. 1968. *Syntaxis*. Culemborg / Keulen: Stam-Kemperman.
- Neeleman, Ad. 1994. *Complex predicates*. Utrecht University (OTS) PhD thesis.
- Neeleman, Ad & Fred Weerman. 1993. The balance between syntax and morphology: Dutch particles and resultatives. *Natural Language & Linguistic Theory* 11(3). 433–475.
- van Riemsdijk, Henk. 1982. A note on case absorption. *Wiener Linguistische Gazette* 28–29. 72–83.
- Weinreich, Uriel. 1954. Is a structural dialectology possible? *Word* 10(2-3). 388–400.



## Chapter 4

# Perception of word stress

Leonor van der Bij

University of Groningen

Dicky Gilbers

University of Groningen

Wolfgang Kehrein

University of Groningen

### 1 Introduction

“When the rain washés you clean, you’ll know” sings Fleetwood Mac’s Stevie Nicks in the song *Dreams* (Nicks 1977) and it sounds weird to all speakers of stress-timed languages. The weak second syllable in the trochaic word *washes* is assigned to a downbeat in the music and its musical note is longer and higher than the one on the strong initial syllable of the same word. This lyric set to music violates the Stress to Beat Matching Principle (Halle & Lerdahl 1993): stressed syllables in words should be assigned to strong beats in music in stress-timed languages (cf. also Beckman, 1986; Proto, 2015).

The second example is less clear: “is er leven op Pluto?” (‘is there life on Pluto?’) is a phrase from the Dutch song *België* (Temming & Westbroek 1982) by Het Goede Doel. Again, the way in which the trochaic word *Pluto* is set to music violates the Stress to Beat Matching Principle in that the weak syllable *to* is on the downbeat. *To* is also longer than *plu*, but in contrast to *washes* the higher pitch in *Pluto* is on the *first* (strong) syllable. As a consequence, there seems to be no consensus among native Dutch listeners as to whether the sung word is *Plúto* or *Plutó*: a preliminary investigation amongst first-year phonology students for the past 25 years by one of the authors reveals that approximately half of the group of roughly 100 students does not hear anything strange in the text setting, whereas the other half does. This clear

division did not change over the years, and it suggests that listeners may be sensitive to different acoustic cues in their perception of word stress.

Word stress is the emphasis given to a certain syllable in a word. The difference between stressed and unstressed syllables can be expressed by means of pitch, loudness, duration, and/or articulatory effort (Laver 1994), but languages differ widely as to which of these components (and to which extent) they apply. This has led to traditional typological classifications, such as the one into languages with intensity-based *dynamic accent* (also: *stress accent*) and those with pitch-based *melodic accent* (or *pitch accent*; see, e.g., Beckman, 1986, Ladd, 2008),<sup>1</sup> or the rhythm-based classification into syllable-timed and stress-timed languages (Pike 1945), with duration playing a prominent role to tell stressed from unstressed syllables in languages of the latter type.<sup>2</sup>

In this paper, we report the results of a pilot study testing how speakers from typologically different languages identify word stress in the trisyllabic nonsense word *tatata*. Our research question is whether participants from different stress-timed and syllable-timed languages have different preferences for pitch or duration as stress cues.

## 2 Method

We created seven variants of the nonsense word *tatata*. First, the non-word *ta* was recorded, copied and put twice behind the original fragment to create a trisyllabic nonsense word with phonetically identical syllables (called NNN in Table 1). This stimulus was used as a control item to check for language-specific positional preferences. The other six items were created from NNN by manipulating two of the three syllables, one for pitch (P in Table 1) and one for duration (D in Table 1). F0 was raised by 2.3% in P-syllables and duration was increased by 30% in D-syllables. The manipulation was performed in Adobe Audition (version 3.0).<sup>3</sup> Each of the seven different stimuli, shown in Table 1, was presented five times in isolation, i.e. listeners had to judge 35 items in total. Two lists were created in which the items were

---

<sup>1</sup> Stress accent languages do show pitch movements on stressed syllables, but pitch is not a property of stress in these languages, but a correlate of accent, i.e. pitch has a pragmatic function (such as focus marking) which happens to be realized on stressed syllables.

<sup>2</sup> Due to lack of space, we can only give a very simplified picture here of both the phonetics of stress and the different accent-based typologies. We refer the reader to Grabe & Low (2002) and Ortega-Llebaria & Prieto (2010) for more thorough and detailed discussions of these issues, including recent work on typologies.

<sup>3</sup> The manipulation can be based on just noticeable differences (JND). Of course, JND is listener-specific. Rietveld & van Heuven (2001: 201) claim that the JND for pitch is between 0,3 and 2,5% and for duration at least 10%. Jusczyk, Cutler & Redanz (1993) and Gut (2013), on the other hand, claim that in English the pitch of stressed syllables is 10% higher than the pitch of unstressed ones and the duration of stressed syllables is twice as long as in unstressed syllables. Because of this difference, a pilot study was performed in which different manipulated items were used to find out which minimal manipulation was noticeable for the participants. The results of this pilot study were that differences in pitch of 2,3% and differences in duration of 30% were still audible. Therefore, these norms were being used for the current study.

pseudo-randomized.

Table 1: The seven stimuli created for this study.

1	Neutral-Neutral-Neutral	NNN
2	Neutral-Pitch-Duration	NPD
3	Neutral-Duration-Pitch	NDP
4	Pitch-Duration-Neutral	PDN
5	Pitch-Neutral-Duration	PND
6	Duration-Pitch-Neutral	DPN
7	Duration-Neutral-Pitch	DNP

The experiment was conducted in a quiet room at the University of Groningen. The items were presented with PowerPoint 2013 on a laptop screen and audibly presented through headphones. All items could be repeated as often as possible. The participants were asked to identify one syllable in each word as the stressed one and write down their answers on an answer sheet.

### 3 The participants and their native languages

Six participants, aged between 22 and 27 years (mean age: 25), took part in the experiment. They were all exchange students from the University of Groningen with English as their second language (L2), but different native languages (L1): two stress-timed European languages (Dutch, German), two syllable-timed European languages (Spanish, Bosnian-Croatian-Serbian), and two (presumably) syllable-timed “Asian” languages (Mandarin Chinese, Singapore English). We give brief descriptions of the phonetic and phonological properties of stress in these languages below.

DUTCH is a stress-timed language with dynamic word stress (Collins & Mees 1984). Sluijter & van Heuven (1996) find duration to be the most reliable correlate of stress in Dutch; overall intensity and vowel quality appear to be weaker cues.<sup>4</sup> Phonologically, stress in Dutch can fall on any of the last three syllables in a word; stress may not fall on the antepenultimate syllable, though, if the penultimate syllable is closed (Kager 1989; van Oostendorp 2012).

GERMAN is very similar to Dutch in both phonetic and phonological respects: like Dutch, German is a stress-timed language with dynamic stress, and like Dutch, duration appears as the most important cue for stress perception, followed by pitch,

<sup>4</sup> Intensity in the higher frequency regions of the spectrum (so called ‘spectral balance’) fares better than overall intensity. Notice that Sluijter & van Heuven do not take pitch (F0) into account “since we take the view that pitch movements are the correlate of accent rather than of stress.” (Sluijter & van Heuven 1996: 2473).

intensity, and vowel quality (Jessen et al. 1995; Dogil & Williams 1999). Finally, German words are stressed on one of the last three syllables, with the restriction that the antepenultima may not be stressed if the penultima is closed (Wiese 2000).

SPANISH is the prototypical language with a syllable-timed rhythm (e.g. Pike, 1945), and we would thus expect duration to play a minor role for stress at best. Indeed, Quilis (1971) and Llisterri et al. (2003) report pitch as the most prominent cue for the perception of stress in Spanish. Ortega-Llebaria (2006), however, identifies duration as another important cue. For stress in unaccented positions (i.e. without a pitch accent), duration turns out to be the most important cue, a finding that is in line with reports from Sluijter & van Heuven (1996) on Dutch, and Jessen et al. (1995) and Dogil & Williams (1999) on German. As in Dutch and German, the position of stress in Spanish is restricted to one of the last three-syllables: in non-verbs, stress typically falls on the final syllable if closed by a consonant (other than *-s*, *-n*), otherwise on the penultimate syllable (Harris 1992).

BOSNIAN-CROATIAN-SERBIAN (BCS) is usually classified as a syllable-timed language with a melodic accent (Josipović 1994). Yet, Lehiste & Ivić (1986) report increased relative duration as the most reliable phonetic correlate of stress. Pitch is less reliable, presumably because BCS is a pitch accent language (also called tonal accent language or restrictive tone language) and thus pitch has a distinctive function: all words in BCS have one of two melodies, one falling, the other rising, aligned with the stressed syllable<sup>5</sup> (Lehiste & Ivić 1986). So called falling accents reach their tonal peak on the stressed syllable, while rising accents reach them not before the poststressed syllable. As far as phonology is concerned, BCS stress can fall on any syllable but the last. Falling accents only occur on the first (or only) syllable, rising accents on any syllable but the last (i.e. they do not occur in monosyllabic words). Contrasts of falling and rising accents are thereby confined to the first syllable of polysyllabic words (Lehiste & Ivić 1986; Inkelas & Zec 1988).

MANDARIN CHINESE is a tone language with four lexical tones (Duanmu 2000). The language is said to have a melodic accent (Chao 1968) and a syllable-timed rhythm (Grabe & Low 2002; Lin & Wang 2007). Yet, acoustic studies (Moore 1993; Shen 1993) show that pitch, duration, intensity, and segmental quality all play a role in distinguishing stressed from unstressed syllables in Mandarin: stressed syllables are produced with a higher pitch range (i.e. raised F<sub>0</sub> in high-toned syllables, lowered F<sub>0</sub> in low-toned syllables, and so on), they are significantly longer, have a greater amplitude, and more peripheral vowels. Moreover, Shen (1993) reports that Mandarin stress can be identified even in the absence of pitch cues, with duration being more important than intensity. As for position, Duanmu (2000) analyzes Mandarin as a language with initial stress and syllabic trochees, built from left to right.

SINGAPORE ENGLISH (SE) is classified as a syllable-timed language, by Ling, Grabe & Nolan (2000), mainly because vowels in unstressed syllables in SE are much less reduced in both duration and quality compared to stress-timed British English. Still, intensity and duration seem to be the most important phonetic cues to stress percep-

---

<sup>5</sup> In fact, most researchers assume that tone is primary to stress in BCS, i.e. the position of stress depends on the position of lexical tones (Inkelas & Zec 1988; Zec 1999).

tion, at least for the Chinese ethnic group of SE speakers (Tan 2002), the group our participant belongs to. The position of stress in SE seems to be largely restricted to one of the first two syllables in a word.

Table 2 summarizes the most important stress properties of our six participants' native languages according to the literature.

Table 2: Summary of important stress properties of the six languages.

Language	Stress- syllable- timed	vs.	Dynamic vs. melodic stress	Important cues for stress	Position of the stressed syllable
Dutch	Stress-timed		Dynamic stress	Duration and intensity	One of the last three syllables
German	Stress-timed		Dynamic stress	Duration and intensity	One of the last three syllables
Spanish	Syllable- timed		Melodic stress	Pitch and du- ration	One of the last three syllables
BCS	Syllable- timed		Melodic stress	Duration	Any syllable but the last
Mandarin	Syllable- timed		Dynamic stress?	Duration, in- tensity, pitch	First syllable
Singapore English	Syllable- timed		Dynamic stress?	Duration and intensity	First or sec- ond syllable

## 4 Results

The major results of the experiment are summarized in Table 3, Table 4 and Figure 1. Table 3 reports how often our participants perceived which syllable type (lengthened, raised pitch, or neutral) as stressed. As can be seen, the German speaker shows by far the strongest preference for duration (80%), followed by, in descending order, the L1 speakers of Spanish, Singapore English, and Mandarin. The Dutch speaker shows only a slight preference for duration (53.3%, as against 40% for pitch). The BCS speaker is the only participant with a preference for pitch; but at the same time, this preference is also the most distinct of all (93.3%).

The very small share of neutral syllables perceived as stressed proves that our participants did hear a difference; and the fact that the BCS speaker identified higher

pitched syllables as stressed in almost all items proves that raising F0 by only 2.3% is perceptible.

Table 3: Perceived stress as a function of different syllable types.

L1	Duration	Pitch	Neutral
Dutch	16 (53.3%)	12 (40%)	2 (6.7%)
German	24 (80%)	6 (20%)	0 (0%)
Spanish	20 (66.7%)	5 (16.7%)	5 (16.7%)
BCS	2 (6.7%)	28 (93.3%)	0 (0%)
Mandarin	18 (60%)	10 (33.3%)	2 (6.7%)
Singapore English	19 (63.3%)	9 (30%)	2 (6.7%)

Positional preferences for stress can be deduced from the speakers' judgments on the five non-manipulated items (NNN), displayed in Table 4.

Table 4: Perceived stress as a function of position (non-manipulated items).

L1	$\sigma 1$	$\sigma 2$	$\sigma 3$
Dutch	1	4	-
German	2	-	3
Spanish	1	4	-
BCS	2	2	1
Mandarin	3	-	2
Singapore English	5	-	-

The SE speaker perceived these items consistently with initial stress, the Dutch and Spanish speakers show a strong preference for the second (= penultimate) syllable, our German and Mandarin speakers picked the first or last syllable as stressed, and the BCS speaker shows no preference for a particular position.

The full picture of stress judgments ordered by stimuli and speakers is shown in Figure 1. Starting again with the Dutch speaker (top on page 39), we see that the slight preference for duration over pitch (see Table 3) is unevenly spread over the different stimuli due to an effect of position. Thus, the lengthened syllable is most often perceived as stressed if (a) in second position (third and fifth bar) or (b) in first position with the neutral syllable in second position (second bar). Something similar applies to the higher pitched syllables (see bars 1 and 4), with the notable exception of NPD (bar 6), where a higher pitched  $\sigma 2$  is outranked by a lengthened  $\sigma 3$ .

For the German speaker (mid on page 39), duration outranks both pitch and position by far, again with NPD (bar 6) as the one notable exception: contrary to the general trend ('duration rules') and default positions ('stress the first or last syllable'),

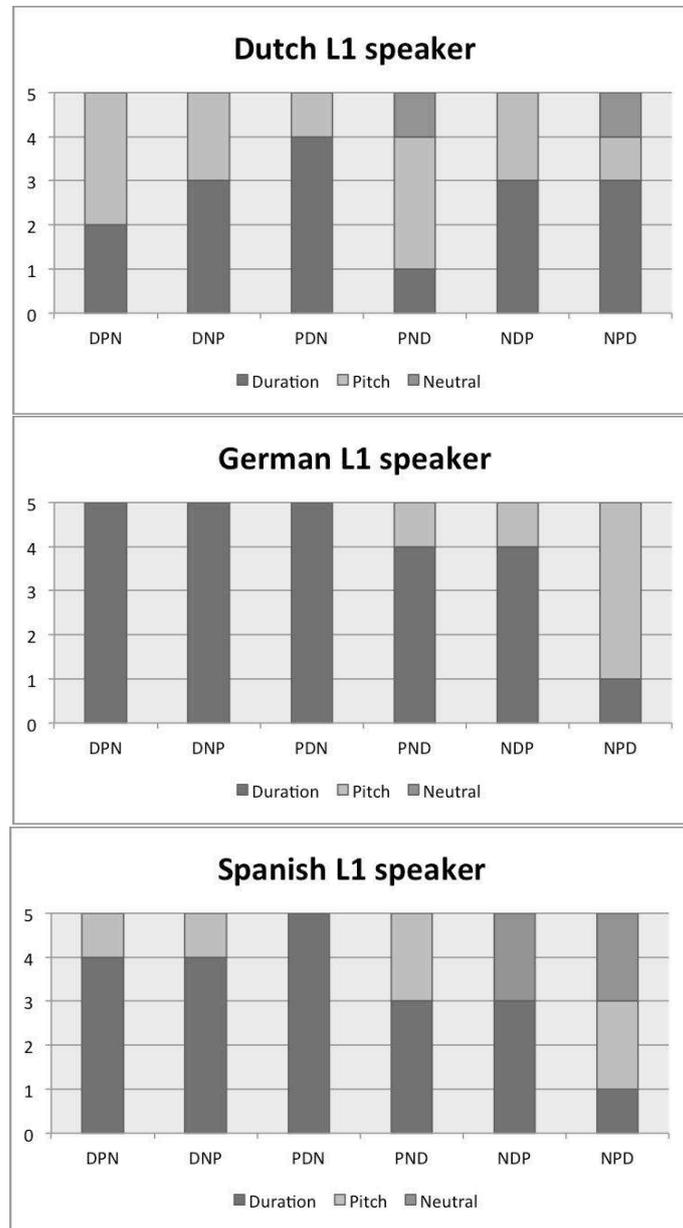


Figure 1: (Continued on page 40.)

our speaker perceived the higher pitched  $\sigma_2$  in NPD (and only there, cf. the first bar) as stressed in four out of five cases.

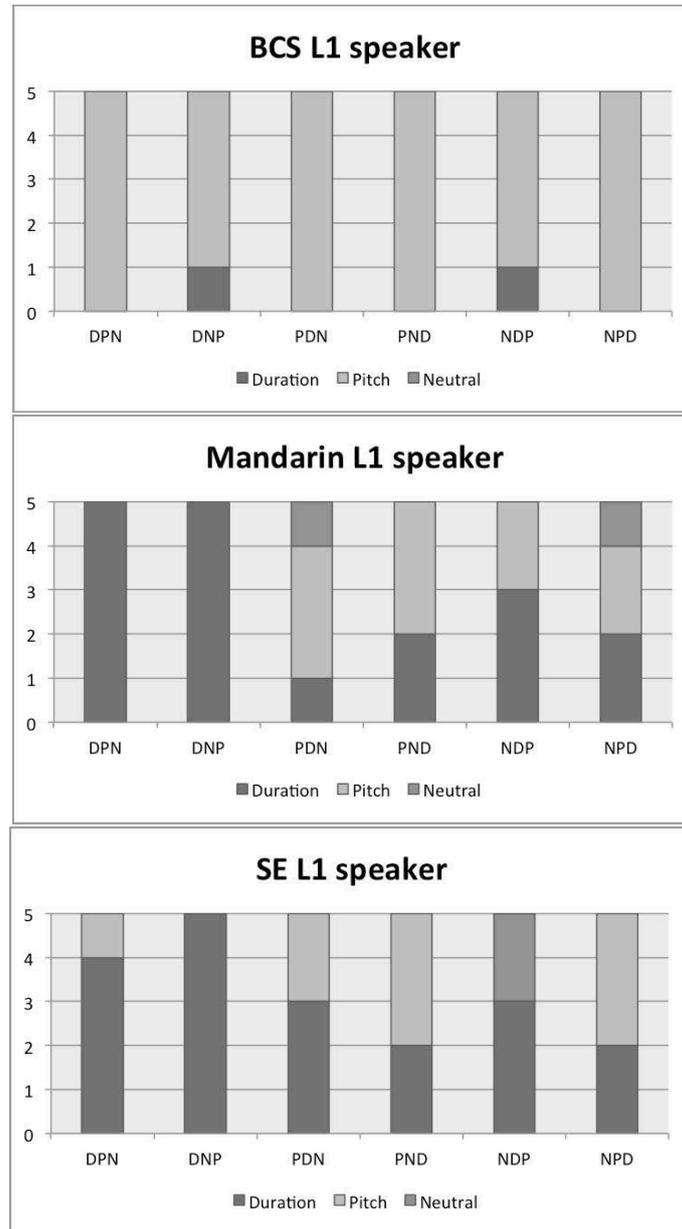


Figure 1: (Continued from page 39.) Perceived stress as a function of syllable types and position.

For the Spanish speaker (bottom on page 39), position shows an effect on the general preference for duration over pitch: lengthened syllables in first position (bars 1 and 2) and second position (bars 3 and 5) are perceived as stressed in 80% of the cases, as against only 40% in final position (bars 4 and 6). Again, NPD (bar 6) differs from the rest in that all three syllable types are perceived as stressed at least once. Finally, the Spanish speaker shows the largest share of neutral syllables (see Table 3); but rather than being randomly distributed between the different items, they only occur with N in word initial position (bars 5 and 6).

The BCS speaker (top on page 40) shows a very strong preference for pitch. If anything, we might infer a very slight effect of position from bars 2 and 5, where our speaker identifies the lengthened syllable as stressed (though only once per category), presumably due to a bias against final stress.

Apart from a solid preference for duration over pitch, our Mandarin speaker (mid on page 40) shows a marked preference for word-initial stress. Thus, lengthened syllables are perceived as stressed to 100% if word-initially (bars 1 and 2), but only to 40% in non-initial positions. Similarly, higher pitched syllables are perceived as stressed to 60% in initial position (bars 3 and 4), but only to 20% non-initially. Notice that position ( $\sigma_1$ ) and acoustic cue (duration or pitch) have to coincide in order to achieve high scores, in other words: word-initial *neutral* syllables (bars 5 and 6) are barely perceived as stressed.

Finally, the SE speaker (bottom on page 40) shows a pattern similar to the Mandarin speaker: a preference for duration over pitch, and another favoring initial stress ( $\sigma_1 > \sigma_2 > \sigma_3$ ). Compared to Mandarin, duration seems to be a slightly stronger cue in SE, in the sense that lengthened second syllables are still perceived as stressed in 60% of the cases (bars 3 and 5). Pitch scores of 60% are only found with duration in final position (bars 4 and 6).

## 5 Discussion

We expected speakers of stress-timed languages to perceive stress primarily by means of duration, and speakers of syllable-timed languages to rely mainly on pitch differences (because stressed and unstressed syllables should not differ much in duration). German and BCS seem to confirm this hypothesis, but stressed-timed Dutch does not, and neither do syllable-timed Spanish, Mandarin, and SE. As a matter of fact, our brief discussion of some phonetic work in Section 3 has already unmasked the traditional isochrony of stress-timed and syllable-timed languages as an oversimplification; and the same holds for the classification of languages into those with dynamic stress and those with melodic stress.

As regards the actual phonetic properties of stress in our participants' L1, our results confirm reports from the literature on the prominent role of duration for stress in German (Dogil & Williams 1999), Spanish (Ortega-Llebaria 2006), Mandarin (Shen 1993), and SE (Tan 2002). The relatively even distribution between duration and pitch of the Dutch speaker is less in line with the literature (Sluijter & van Heuven 1996), but matches well with the variable reactions to the stress pattern of the word *Pluto*

in “is er leven op Pluto?”, mentioned in the introduction. Finally, the BCS speakers’ strong preference for pitch goes directly against (Lehiste & Ivić 1986), who found duration to be the most reliable phonetic correlate of stress. We have no explanation for this mismatch. Notice that the equation of higher pitch with stress on the second and third syllable is particularly surprising, given that in BCS higher pitch on a non-initial syllable does *not* coincide with stress, but rather indicates stress on the *preceding* syllable.

The strength of a particular phonetic cue (duration, pitch) for the perception of stress can also be seen in the way it competes with positional preferences. To start with the two extremes, duration in German and pitch in BCS, these cues overrule position by far: words in BCS, for instance, are never stressed on their final syllable, and yet our BCS speaker perceives final syllables with higher pitch as stressed to 80%. The effect of position on phonetic cues is moderate for our Dutch and Spanish participants, and strong for the Mandarin and the SE speaker.

Finally, stress judgments on NNN words (Table 4) show similarities between Dutch and Spanish on the one hand, and German and Mandarin on the other. The marked difference between Dutch and German might come as a surprise; after all, the two languages are closely related and their stress systems are typically analyzed as very similar. Yet, both languages allow (in principle) stress on all three light syllables in a trisyllabic word (Dutch: *Cánada*, *pyjáma*, *chocolá*; German: *Kánada*, *Bikíni*, *Melodíé*), and it is thus possible that the Dutch speaker prefers a foot structure with one final syllabic trochee L(́LL) while the German speaker parses [tatata] into two trochees (LL)(L), with either the first or the second carrying main stress: (́LL)(L) or (LL)(́L). It is also possible that the Dutch speaker interprets [a] as a long vowel /a:/ (because short /a/ has a back quality [ɑ] in Dutch), which would increase the likeliness of stress on the penultimate syllable (Gilbers & Jansen 1996).

## 6 Conclusion

We conducted a pilot study to examine how speakers from typologically different languages (Dutch, German, Spanish, Bosnian-Croatian-Serbian, Mandarin Chinese and Singapore English) perceive word stress by means of pitch and duration. Our German speaker relied mainly on duration, the speaker of Bosnian-Croatian-Serbian used pitch almost exclusively. The other participants showed a slight or moderate preference for duration. Our results are (mostly) in line with the phonetic properties of stress in our participants’ L1, but *not* with traditional classifications into stressed-timed and syllable-timed languages, thereby confirming earlier criticism of such clear-cut typological categories (Roach 1982; Cauldwell 2002; among many others). Notice, however, that not all our results can be ascribed to the native language of our participants. Since we tested only one speaker of each language, we cannot rule out from the outset that (some of) the differences are a result of *individual* (and thus L1-independent) preferences for one cue over the other. In other words: if speakers of Dutch can be divided between duration and pitch if they receive conflicting cues (as in *Pluto* above), so can speakers of other languages. Future research with more

speakers from each language will help to detangle systematic (i.e. L1-related) effects from possible individual preferences.

## References

- Beckman, Mary E. 1986. *Stress and non-stress accent*. Dordrecht: Foris Publications.
- Cauldwell, R. 2002. The functional irrythmicality of spontaneous speech: a discourse view of speech rhythms. *Journal of Applied Language Studies* 2. 1–24.
- Chao, Y.-R. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Collins, Beverley & Inger M. Mees. 1984. *The sounds of English and Dutch*. Leiden: Leiden University Press.
- Dogil, Grzegorz & Briony Williams. 1999. The phonetic manifestation of word stress. In Harry van der Hulst (ed.), *Word prosodic systems in the languages of Europe*, 273–334. Berlin/New York: Mouton de Gruyter.
- Duanmu, S. 2000. *The phonology of standard Chinese*. Oxford: Oxford University Press.
- Gilbers, Dicky & W. Jansen. 1996. Klemtoon en ritme in Optimality Theory. *TABU* 2. 53–101.
- Grabe, Esther & Ee L. Low. 2002. Durational variability in speech and the rhythm class hypothesis. In N. Warner & C. Gussenhoven (eds.), *Papers in laboratory phonology 7*, 515–546. Berlin: Mouton de Gruyter.
- Gut, U. 2013. Analysing phonetic and phonological variation on the suprasegmental level. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*. Cambridge: Cambridge University Press.
- Halle, John & Fred Lerdahl. 1993. A generative textsetting model. *Current Musicology* 55. 3–23.
- Harris, John W. 1992. *Spanish stress: the extrametricality issue*. Washington: IULC Publication.
- Inkelas, S. & D. Zec. 1988. Serbo-Croatian pitch accent: the interaction of tone, stress, and intonation. *Language* 64. 227–248.
- Jessen, Michael, Krzysztof Marasek, Katrin Schneider & Kathrin Classen. 1995. Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German. In *Proceedings 13th ICPhS (Stockholm)*, vol. 4, 428–431.
- Josipović, V. 1994. English and Croatian in the typology of rhythmic systems. *Studia Romanica et Anglica Zagabiensia* 39. 25–37.
- Jusczyk, P. W., A. Cutler & N. J. Redanz. 1993. Infants' preference for the predominant stress patterns of English words. *Child Development* 64. 675–687.
- Kager, R. 1989. *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris Publications.
- Ladd, R. 2008. *Intonational phonology*. Second edition. Cambridge: Cambridge University Press.
- Laver, John. 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.
- Lehiste, I. & P. Ivić. 1986. *Word and sentence prosody in Serbocroatian*. Cambridge: MIT Press.

- Lin, H. & Q. Wang. 2007. Mandarin rhythm: an acoustic study. *Journal of Chinese Linguistics and Computing* 17. 127–140.
- Ling, L. E., Esther Grabe & F. Nolan. 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43. 377–401.
- Llisterri, Joaquim, María Machuca, Carme de la Mota, Montserrat Riera & Antonio Ríos. 2003. The perception of lexical stress in Spanish. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, vol. 39, 2023–2026.
- Moore, C. B. 1993. Phonetic observations on stress and tones in Mandarin Chinese. *Working Papers of the Cornell Phonetics Laboratory* 8. 87–117.
- Nicks, S. 1977. Dreams. On: *Fleetwood Mac – Rumours* (Warner Brothers WB 56 344).
- Ortega-Llebaria, Marta. 2006. Phonetic cues to stress and accent in Spanish. In M. Diaz-Campos (ed.), *Selected proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology*, 104–118. Somerville, MA: Cascadilla Press.
- Ortega-Llebaria, Marta & Pilar Prieto. 2010. Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech* 54 (1). 73–97.
- Pike, K. L. 1945. The intonation of American English. In D. Bolinger (ed.), *Intonation*, 53–83. Harmondsworth: Penguin.
- Proto, Teresa. 2015. Prosody, melody and rhythm in vocal music: the problem of textsetting in a linguistic perspective. In Björn Köhnlén & Jenny Audring (eds.), *Linguistics in the Netherlands 2015 (AVT 32)*. Amsterdam: Benjamins.
- Quilis, Antonio. 1971. Caracterización fonética del acento español. *Travaux de Linguistique et de Littérature* 9. 53–72.
- Rietveld, A. C. M. & Vincent J. van Heuven. 2001. *Algemene fonetiek*. Bussum: Uitgeverij Coutinho.
- Roach, P. 1982. On the distinction between “stress-timed” and “syllable-timed” languages. In D. Crystal (ed.), *Linguistic controversies: essays in linguistic theory and practice in honour of F. R. Palmer*, 73–79. London: Edward Arnold.
- Shen, X. S. 1993. Relative duration as a perceptual cue to stress in Mandarin. *Language and Speech* 36. 415–433.
- Sluijter, Agaath M. C. & Vincent J. van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical society of America* 100. 2471–2485.
- Tan, Y. Y. 2002. *Acoustic and perceptual properties of stress in the ethnic subvarieties of Singapore English*. National University of Singapore PhD thesis.
- Temming, H. & H. Westbroek. 1982. *België*. 45 rpm single Het Goede Doel - CNR 161.013.
- van Oostendorp, Marc. 2012. Quantity and the three-syllable window in Dutch word stress. *Language and Linguistics Compass* 6. 343–358.
- Wiese, R. 2000. *The phonology of German*. Oxford: Oxford University Press.
- Zec, D. 1999. Footed tones and tonal feet: rhythmic constituency in a pitch-accent language. *Phonology* 16. 225–264.

## Chapter 5

# Empirical evidence for discourse markers at the lexical level

Jelke Bloem

University of Amsterdam

I use a discourse-annotated corpus to demonstrate a new method for identifying potential discourse makers. Discourse markers are often identified manually, but particularly for natural language processing purposes, it is useful to have a more objective, data-driven method of identification. I link this task to the task of identifying co-occurrences of words and constructions, a task where statistical association measures are often used to compute association strengths. I then apply a statistical association measure to the task of discourse marker identification, and present results for several discourse relation types. While the results are noisy due to the limited availability of corpus data, they appear usable after manual correction or as a feature in a classifier. Furthermore, the results highlight a few types of lexical discourse relation cues that are not traditionally considered discourse makers, but still have a clear association with particular discourse relation types.

## 1 Introduction

The coherence structure of texts is one of many aspects of language that can be studied computationally to gather empirical evidence for previously formulated theories or categorizations. When such theories or categorizations can be automatically mapped to language usage data, this may also provide natural language processing (NLP) systems with more information about a text. Automatically determining the coherence structure of a given text allows for higher-level semantic analysis that is useful in many NLP applications. This structure, consisting of relationships between clauses, is generally described in terms of coherence relations, such as *ELABORATION*.

There are various frameworks that formalize these relations. One widely used model is Rhetorical Structure Theory (RST) (Mann & Thompson 1988), which provides detailed, tree-shaped structures that cover everything from elementary discourse units to long text spans. At the lowest level of a RST tree structure, discourse relations hold between minimal units of discourse, or spans. These units are generally defined at the level of clauses. Spans can be more central (a nucleus) or secondary to

the relation (a satellite). Relations between nuclei and satellites allows us to define the fact that one span fulfills a specific role in the discourse, relative to the other. Multinuclear relations are also possible, such as *SEQUENCE*. In this case, no single nuclear span can be defined, they are all equally central to the relation. Higher up in a RST tree, these relations also apply to larger spans that cover multiple elementary discourse units. Other examples of formal discourse models are Discourse Representation Theory (DRT) (Kamp 1981; Kamp, van Genabith & Reyle 2011) and the Cognitive Approach to Coherence Relations (CCR) (Sanders, Spooren & Noordman 1992), in which coherence relations consist of features.

Coherence relations in text can be explicitly marked, as in example (1) from the RST Discourse Treebank (Carlson, Okurowski & Marcu 2002), which shows a *RESULT* relation:

- (1) (...) it is better positioned than most companies for the coming overcapacity, because its individual mills can make more than one grade of paper.

Non-lexical markers, such as intonation, are also possible, but not relevant in written text. Coherence relations are often left implicit as well, such as in the following example by Webber (2004):

- (2) John is stubborn ( $C_1$ ). His sister is stubborn ( $C_2$ ). His parents are stubborn ( $C_3$ ). So, they are continually arguing ( $C_4$ ).

The *RESULT* relation between  $C_4$  and the other elements is explicitly marked by ‘so’, but the relations between  $C_1$ ,  $C_2$  and  $C_3$  are left implicit (in RST, this is a *LIST* relation).

Explicit lexical markers can be used as cues for systems that attempt to automatically determine discourse relations. There are a limited number of markers known for each kind of discourse relation, so the markers are often identified intuitively and little empirical research has been done to verify their association with the discourse relation(s) they mark. In addition, there might be more markers that are not as obvious, which could be missed in this way.

In this work, I present a method for detecting potential discourse relation markers in a more objective and empirical way. Based on a statistical measure of association, the method used can bring up candidate markers from real-world annotated corpus data purely on the basis of an objective statistical measure. The method will not guarantee a clean list of discourse markers, but it can serve as a first step in detecting them and could have applications as a feature in a larger system for discourse relation detection.

The problem of finding associations between discourse markers and discourse relations can be related to the general issue of discovering whether words co-occur with some other linguistic structure or element. In corpus linguistics, this type of issue is commonly solved using statistical measures of association, such as the  $\chi^2$ -test. Collocations, words that co-occur more often than would be expected by random chance, are generally analyzed using such methods. Discourse markers and the relations they mark can be viewed from the same perspective – one would expect a

discourse marker to be strongly associated with the discourse relation they mark. First I will discuss some related work, and then explain the method and what data is required to apply it to. I will then discuss the results of applying this method to some English-language corpus data, and draw conclusions.

## 2 Related work

Various studies discuss the identification of discourse markers. Knott (1996) manually examined a set of academic texts to gather a corpus of about 350 discourse markers. To test whether words are discourse markers, he employed a linguistic test. The test involves isolating clauses that contain the markers and checking whether they appear complete that way. If they seem incomplete, then they are considered relational, since they have to be in a relation with another clause to be coherent. This test is argued to be reasonably objective. However, one needs to manually identify candidate discourse markers in advance to be able to perform it, based on intuition. Recent work on discourse connectives in discourse processing also use a pre-defined set of discourse markers, such as Pitler & Nenkova (2009), who use the annotated explicit connectives in the Penn Discourse Treebank and present a method of disambiguating them when they can mark multiple relations.

Timmerman (2007) developed an automatic recognizer for Dutch that relies on discourse markers to generate the RST structure of texts in the medical domain. In his analysis, he produces manually collected lists of Dutch discourse markers, and the relations they signal. He also considers the class of domain-specific markers, in this case medical words, which is interesting since it indicates that (topic) domains are a consideration and that discourse markers are an open class for which new members can be found.

Statistical measure of association have already been applied in discourse structure research, but only for disambiguation purposes. Spenader & Lobanova (2009) have taken two specific relations, CONTRAST and CAUSE-EFFECT, and used the  $\chi^2$ -test of association strength to determine if intuitively selected markers can reliably distinguish the two. They measure the association between the occurrence of a specific marker in the CONTRAST relation versus the CAUSE-EFFECT relation. However, this is only possible if there is a hypothesis about what relation types a potential marker might belong to, and if only a few relation types are being studied. They surprisingly find that the marker *however*, normally considered a marker of contrast, doesn't help to distinguish CONTRAST and CAUSE-EFFECT. They also find some novel discourse markers. This research shows that statistical methods can result in new findings about discourse markers that intuition does not provide, and do so in a more objective way.

Khazaei, Xiao & Mercer (2015) also present a method of selecting potential lexical coherence markers for RST relations involving the use of n-grams from corpora, to be used in a discourse relation classifier for CIRCUMSTANCE relations. This selection task is similar to the task we describe, but a different method is used. The RST Discourse Treebank (Carlson, Okurowski & Marcu 2002) is used for the extraction

of n-grams (up to trigrams). A modified TF-IDF metric is applied to data extracted from the corpus. The relations from the corpus are sorted into documents, one document per relation type, and TF-IDF is then used to select potentially relevant cues that identify a particular document (i.e. relation type). Their study focuses on the CIRCUMSTANCE relation, and results are reported only for this relation. The fact that Khazaei, Xiao & Mercer (2015) use their lexical markers to classify CIRCUMSTANCE relations also shows that automatically detected lexical coherence markers can indeed be used for discourse relation classification, although in this work, only one relation type was classified.

### 3 Method

In this section, I will describe my use of statistical methods of association for detecting potential discourse markers for a given relation type. The basic idea is to find words that occur in a given discourse relation type more often than would be expected by chance, i.e. if the words were randomly distributed over the different relation types. These words can then be considered to be associated with that discourse relation, which indicates that they may be markers for the relation. We also distinguish between nuclei and satellite spans of relations. For example, in a RESULT relation, the thing in the nucleus span is what caused (or might have caused) the thing described in the satellite span. Since different words may be associated with each of these functions, it seems better to distinguish them. However, Khazaei, Xiao & Mercer (2015) do not make this distinction.

This task requires a sufficiently large corpus of discourse-annotated data, from which the relations and the words in their spans should be extracted. One can then calculate the association between these words and the relation nucleus or satellite. Since some markers can consist of multiple words, I included bigrams of the words as well. Trigrams could also be considered, in case there are multi-word potential discourse markers, but this would likely make the results too noisy.

#### 3.1 Use of Context Ratio

The proposed method using measures of statistical association is inspired by the use of association strength in other areas of linguistics. Lichte & Soehn (2007) used the method to discover *negative polarity items* (NPIs). These are lexical constructions that can only occur in the scope of *downward entailing* (DE) operators. By measuring the association strength between words (potential NPIs) and clauses in the scope of DE operators, new NPIs could be found. The task is similar to finding discourse markers, in that there is some class of lexical items that is only partly known, associated with a particular kind of context. The authors of this study used the Context Ratio measure, a very basic measure of association. However, the NPIs that they are looking for are expected to occur in DE contexts most of the time, while discourse markers don't only occur in the discourse relation they mark. For our purposes, it would be better

to use a measure that also takes into account how often the marker word appears in other contexts, such as Fisher's Exact Test.

### 3.2 Use of Fisher's Exact Test

Another example of the use of association strength in linguistics is the method of collocation analysis, first described by Stefanowitsch & Gries (2003). It is concerned with linguistic constructions, such as [N *waiting to happen*], where N is an open slot that can be filled with a noun. They found that some words are more strongly associated with such a construction than others, providing clues about the construction's meaning. They calculate the association between a word and a construction using Fisher's exact test. This test is particularly well suited to sparse data, a common occurrence in linguistics. Many measures of association make distributional assumptions that are not valid for linguistic data. The occurrence of words in language is not normally distributed, some words occur quite often while many words occur rarely.

Fisher's Exact Test provides a  $p$ -value as well as a measure of effect size (Maximum Likelihood Estimate odds ratios), which provides a threshold that makes it possible to state whether an association is statistically significant. It is also an exact computation, while many other association measures are estimations, an important point for this investigation since sparse data is likely. So I have decided to follow Stefanowitsch & Gries (2003) and use Fisher's Exact Test focusing on  $p$ -values. However, an empirical validation to determine the optimal association measure for a particular task could show another measure to be better. This was done by Wiechmann (2008) for the task of collocation analysis. The use of a measure of effect size such as odds ratios is also a viable alternative, when one considers the size of the associations between words and relations to be more important than defining a threshold.

### 3.3 Use of TF-IDF

The choice for a measure of association, such as Fisher's Exact Test, can be contrasted with Khazaei, Xiao & Mercer's (2015) approach using TF-IDF (Term Frequency - Inverse Document Frequency), which is more commonly used in NLP tasks. This measure is normally used to measure the importance of words in documents, i.e. in the NLP tasks of keyword identification or text summarization. Khazaei, Xiao & Mercer (2015) apply it to the task of discourse marker identification by compiling all spans of a particular relation, such as CIRCUMSTANCE, into a single document, resulting in one document per relation type. Methods that operate on documents, such as keyword identification, can then be employed. The method works by taking the term frequency (TF) (the frequency of a particular word throughout the whole set of documents, i.e. the corpus) and dividing it by the inverse document frequency (IDF). The IDF represents the number of documents that contain the word. In terms of Khazaei, Xiao & Mercer's (2015) study, it represents the number of discourse relation types that contain the word. Therefore, the frequency of a word in the corpus is weighted by the number of different relation types it appears in. Statistically, this is very similar to the way Fisher's Exact Test is computed, however, there is one factor that is

involved in the computation of measures of association, but not in computing TF-IDF: the frequency of the word (or term) in a particular relation type. TF-IDF only takes into account whether a term occurs in a relation type, but not how often. It can be argued that important information is lost in this way – a word that occurs only once in a RESULT nucleus is probably less important as a marker of results than a word that occurs in RESULT nuclei many times. Furthermore TF-IDF cannot provide  $p$ -values.

## 4 Data

This section will discuss my source of discourse-annotated data, and show what information should be extracted from it in order to use Fisher's Exact Test to compute association strengths between the discourse relations and the words that are potential markers of these relations. The task requires a sufficiently large corpus of discourse-annotated data, meaning that the text is annotated for discourse units, the relations between them (or at least the low-level relations), and the relation type. Discourse markers don't need to be annotated, since they are what the method is trying to find. I chose to use the RST Discourse Treebank<sup>1</sup> (Carlson, Marcu & Okurowski 2001), a manually annotated treebank with over 178,000 words and over 21,000 discourse units, as my data source. This treebank was also used by Khazaei, Xiao & Mercer (2015) for their discourse marker identification study. While it is smaller than the Penn Discourse Treebank (PDTB), it is based on the widely used Rhetorical Structure Theory (RST) approach, rather than being theory-neutral like the PDTB. The RST annotators claim not to have been influenced by the presence of discourse markers during their analysis (Williams & Reiter 2003). More importantly, the PDTB does not distinguish between nuclei and satellites of relations, an important concept in RST. Instead, relations have a first and a second argument. Discourse markers are more likely to be related to the function of a discourse unit than to its position, so the nucleus-satellite distinction is an important one.

To apply measures of association strength to this data, we gather frequency data from the corpus on discourse types, words that occur in them, and their co-occurrence.  $p$ -values are computed from this, where a lower  $p$ -value indicates a stronger association. I only consider words in the first, second or last position, or bigrams of the first and second position in case there are two-word markers, because this is where discourse markers tend to appear. For the relations, I only consider the ones between minimal units of discourse. To test all of the discourse relations, I run the test multiple times, once for each relation type.

## 5 Results

In this section, I will discuss the results of applying this method to the chosen corpus for a few selected types of discourse relation. There are 57 types in the corpus, so I

---

<sup>1</sup> <http://www.isi.edu/~marcu/discourse/Corpora.html>.

## 5 Empirical evidence for discourse markers at the lexical level

cannot report on all of them for reasons of space. The following results have been calculated with Fisher’s Exact Test, without any frequency cutoffs.

Table 1: Rankings of top discourse marker candidates for nuclei and satellites of CONSEQUENCE-N relations. The values are the  $p$ -values that represent association strength.

<i>because</i>	8.71658e-027	<i>and</i>	1.53749e-005
<i>because of</i>	5.62176e-012	<i>the dollar</i>	1.54064e-005
<i>of</i>	2.02458e-007	<i>share</i>	4.21597e-005
<i>largely because</i>	2.15549e-005	<i>loss</i>	9.29785e-005
<i>when</i>	2.31501e-005	<i>dollar</i>	9.29785e-005

(a) CONSEQUENCE-N satellites
(b) CONSEQUENCE-N nuclei

Table 1a shows the five words that are most strongly associated with satellites of CONSEQUENCE-N relations, the same type used in the example contingency table above. *Because* tops the rankings, along with some bigrams involving this word. While I wasn’t able to find a list of consequence markers (this relation type is not used in all versions of RST), the definition of *because*, ‘for the reason that’,<sup>2</sup> seems to imply consequence. The word *of* is not a discourse marker, but is likely listed due to the common collocation ‘because of’, which also appears in the list and can be considered a multiword discourse marker. The word *when*, appearing 5th in the ranks, tends to occur in this relation when the satellite comes first, in the form ‘When  $x$  happens, there is a consequence’. Therefore, it can be considered a marker of consequence, even though it doesn’t seem to be known as such.

Table 1b shows the top five for nuclei of the same relation type. It contains a stopword and various financial terms, and nothing that could be considered a discourse marker. Yet, the candidate markers that are found still have  $p$ -values below common thresholds of statistical significance, such as  $\alpha < 0.001$ . The financial terms are domain-specific noise, since the RST treebank is based on Wall Street Journal articles. Domain-specific discourse markers also exist (Timmerman 2007), but we do not find them here. It seems that this type of relation is generally marked in the satellite. An example of such a relation helps to illustrate this:

- (3) Nuc: *Lockheed reported a \$32 million third-quarter net loss,*  
 Sat: *largely because of cost overruns on fixed-price military contracts.*

The relation here is marked by *largely because* in the satellite, while the nucleus simply reports the consequence (of the cost overruns) without further marking.

The PURPOSE relation is regarded as one of the most commonly marked ones (Taboada 2006). By examining such a relation, there should be less noise among the most strongly associated words. Table 2a shows the result for PURPOSE satellites,

<sup>2</sup> <http://www.merriam-webster.com/dictionary/because>.

Table 2: Rankings of top discourse marker candidates for satellites of PURPOSE and CONCESSION relations, extracted using the described method.

<i>to</i>	0	<i>even</i>	2.73407e-036
<i>to yield</i>	2.31430e-033	<i>though</i>	4.36469e-030
<i>yield</i>	3.68036e-032	<i>although</i>	7.37035e-026
<i>build</i>	5.59667e-017	<i>despite</i>	1.76170e-021
<i>to build</i>	7.65295e-014	<i>even if</i>	1.81139e-013
<i>accommodate</i>	9.69231e-014	<i>despite the</i>	1.08417e-012
<i>to accommodate</i>	9.69231e-014	<i>even though</i>	7.25280e-010
<i>keep</i>	2.09227e-013	<i>if</i>	7.74321e-007

(a) PURPOSE satellites

(b) CONCESSION satellites

which is where the relation is normally marked. It shows that *to* is very strongly associated with PURPOSE satellites, which is also one of the markers identified by Taboada in spoken dialogue. Most of the other candidate markers in the top part of the list are bigrams of *to* and some verb, and these verbs also occur by themselves in the list (such as *yield*). They are not normally considered discourse markers, but clearly they have a strong association with PURPOSE satellites anyway. Intuitively they do seem to be semantically related, which is actually a claim of the collocation analysis method that inspired the method used in the present study – one can find out about the meaning of a construction by looking at the words it occurs with. Perhaps the same can be said of discourse relation types.

In the 11th position (not visible in the table) there is also the bigram *in order*, likely used as part of *in order to*, which indicates purpose. Apparently there are some potential discourse markers that will be missed by not including trigrams in the analysis. However, introducing more n-grams also introduces noise. This can be seen in the results of CONCESSION satellites in Table 2b, where we find the bigram *despite the* as a potential discourse marker.

In their discourse marker identification study, Khazaei, Xiao & Mercer (2015) focused on the CIRCUMSTANCE relation, identifying the following potential lexical cues: *When, after, on, before, with, out, as*. It is difficult to compare these studies directly due to the different choices made in selecting and filtering the data (for example, their study does not distinguish between nucleus and satellite spans), but I will try to provide a comparison here. In my results, the cue *when* can be found as the top ranked cue for CIRCUMSTANCE satellites. *as* is ranked second, *after* is ranked fourth. *With* is ranked 15th for satellites, and 40th for nuclei. *On* is ranked 57th for satellites. *Before* is the third ranked cue for CIRCUMSTANCE nuclei. *Out* is only ranked as the 264th cue for nuclei, however, *without* is the 6th ranked cue for satellites of CIRCUMSTANCE relations. Other cues that are highly ranked in my results but are not mentioned by Khazaei, Xiao & Mercer (2015) include bigrams involving *when*, some high-frequency words such as *the*, and some domain-specific words such as *prices*

as a cue for circumstance nuclei. There is clearly some degree of overlap between the results of the two methods, but some of the markers they found were not ranked highly by my method, even though the same corpus was used.

## 6 Discussion

The results show that statistical association measures can be used to identify potential discourse makers, but that the results are noisy. Taking all of the markers that are statistically significant would result in a far too large list. The significance threshold could be made more strict by applying a correction for performing multiple tests, such as the Bonferroni-correction, but that would only be informative if the data was less noisy. Alternatively, it would be better to use a method of association that is not focused on identifying a particular class of statistically significant discourse markers, but rather on ranking them by their association strength.

The data appears to be noisy due to the use of a large number of domain-specific terms in the corpus, i.e. related to the topics typically discussed in the Wall Street Journal. Khazaei, Xiao & Mercer (2015) deal with domain-specific lexical cues by filtering the lists of candidate discourse markers against another discourse-annotated corpus from a different domain: they remove all cues that fail to successfully identify the same relations in the secondary corpus. However, this may not always be desirable. It is quite likely that there are domain-specific discourse markers or lexical cues to discourse relations. This didn't come up in the results I presented, but in a previous pilot study with Dutch fundraising letters, a possible example was found: the marker 'fill' for filling in a form occurred in satellites of *ENABLEMENT* relations. Other domain-specific lexical cues that are strongly associated with the discourse relation, such as the purpose verbs in table 2a, may not necessarily be functional markers of text cohesion, but could still be interesting for natural language processing systems that aim to detect discourse relations in the domain in question. Some of the noise may also come from the small size of the corpus. Perhaps such concerns could be alleviated by combining certain similar discourse relations, or not distinguishing nucleus and satellite spans of relations, but this may make the results less interesting or detailed.

The present results would require further manual processing, for example using Knott's (1996) test, to be useful as lists of actual words with a function of marking coherence. Alternatively, they could be compared to existing, manually compiled lists of discourse markers for that relation, but such lists don't seem to be commonly available, and are made more difficult by the different theories on discourse relations.

Lastly, it would be interesting to investigate why some lexical cues that are not traditionally discourse markers are strongly associated with certain discourse relations. We noted that some of the lexical cues appear to have strong semantic relations to the discourse relation type that they appear in. If speakers use these semantically related elements in interpreting coherence relations, this implies that there might be networks of discourse markers and words with discourse-marking properties. This would provide an interesting parallel to construction grammar theories of language.

Experimental studies could investigate whether speakers of a language can use constructions, rather than just single words, as cues for processing coherence relations between clauses.

## Acknowledgements

An older version of this paper was written for John Nerbonne's 2011 Seminar in Methodology and Statistics. I am grateful for his helpful comments on this work.

## References

- Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 1–10.
- Carlson, Lynn, Mary Ellen Okurowski & Daniel Marcu. 2002. *RST Discourse Treebank*. Linguistic Data Consortium, University of Pennsylvania.
- Kamp, Hans. 1981. A theory of truth and semantic representation. *Formal Methods in the Study of Language*.
- Kamp, Hans, Josef van Genabith & Uwe Reyle. 2011. Discourse representation theory. In *Handbook of philosophical logic*, 125–394. Springer.
- Khazaei, Taraneh, Lu Xiao & Robert E. Mercer. 2015. Identification and disambiguation of lexical cues of rhetorical relations across different text genres. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, 54.
- Knott, A. 1996. *A data-driven methodology for motivating a set of coherence relations*. Department of Artificial Intelligence, University of Edinburgh PhD thesis.
- Lichte, T. & J. P. Soehn. 2007. The retrieval and classification of negative polarity items using statistical profiles. *Roots: linguistics in search of its evidential base*. 249–266.
- Mann, W. C. & S. A. Thompson. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3). 243–281.
- Pitler, Emily & Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 13–16. Association for Computational Linguistics.
- Sanders, Ted J. M., Wilbert P. M. Spooren & Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes* 15(1). 1–35.
- Spender, J. & A. Lobanova. 2009. Reliable discourse markers for contrast relations. In *Proceedings of the eighth International Conference on Computational Semantics*, 210–221. Association for Computational Linguistics.
- Stefanowitsch, A. & S. T. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.
- Taboada, M. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics* 38(4). 567–592.

### *5 Empirical evidence for discourse markers at the lexical level*

- Timmerman, S. 2007. *Automatic recognition of structural relations in Dutch text*. MA thesis, University of Twente PhD thesis.
- Webber, Bonnie. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28(5). 751–779.
- Wiechmann, D. 2008. On the computation of collocation strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.
- Williams, S. & E. Reiter. 2003. A corpus analysis of discourse relations for natural language generation.



## Chapter 6

# Verb phrase ellipsis and sloppy identity: a corpus-based investigation

Johan Bos

University of Groningen

In a corpus-based confrontation between sloppy and strict identity in elliptical contexts, the former beats the latter with a striking 9–0. Whether this results is representative for verb phrase ellipsis in general is a question of debate. Perhaps the sloppy players had home advantage, and strict players perform better in corpora other than the Wall Street Journal.

## 1 Introduction

In this article<sup>1</sup> I will present corpus instances of Verb Phrases Ellipsis (VPE), a linguistic phenomenon that manifests itself in the English language when an auxiliary verb is used to refer to a complete verb phrase mentioned elsewhere in the linguistic context, as in *Bill wrote a paper, and John did too*. Recently, Jennifer Spenser and myself annotated a large corpus of English newspaper text (parts of the Wall Street Journal) on occurrences of VPE (Bos & Spenser 2011). We undertook this effort because, until then, detailed annotation work of VPE carried out on a large scale did not exist, with the exception of Hardt (1997) and Nielsen (2005). The primary aim of this enterprise was to develop benchmark tools for automated ellipsis recognition and resolution in the context of natural language processing. However, the results can also be used to study the distribution and frequency of the various types of VPE and problems they trigger known from the rich linguistic literature on ellipsis.

Some of the findings of Bos & Spenser (2011) were expected, and some unexpected. Not surprising was the relative rarity of the phenomenon of VPE in newswire: on average they found only one instance of VPE in every 109 sentences (Bos & Spenser 2011). However, much to my surprise, is the lack of overlap of types of VPE found in

---

<sup>1</sup> This paper is dedicated to John Nerbonne, who introduced me more than twenty years ago to the world of computational semantics (Nerbonne 1996). He supervised my master's thesis work with an enormous amount of enthusiasm and expertise. This resulted in my first international publication (Bos 1993). I am proud to say that I was John's very first graduate student in Groningen.

*Johan Bos*

the Wall Street Journal corpus with the classical examples found in the theoretical ellipsis literature. Well-studied phenomena such as pseudo-gapping, split antecedents, cascaded ellipsis, antecedent-contained deletion are scarce in the newspaper genre. So are VPE that give rise to what Dahl (1973) called *sloppy identity*.<sup>2</sup> It is this latter phenomenon that I will closer inspect in this article. Its statistical presence in real corpora was unknown, and it still is. In this article I will look at the occurrence of sloppy identity in the well-known Wall Street Journal corpus.

## 2 Sloppy identity

Many linguists are fascinated by the notorious strict/sloppy ambiguity that manifests itself in VPE (Dahl 1973; Sag 1976; Klein 1987). It occurs when the subjects of the source and target clause denote different entities, and a pronoun appears in the source clause and co-refers with the subject of the same clause. An example is *John likes his mother, and Bill does too*, where the strict interpretation gives rise to the reading where Bill likes John's mother and the sloppy interpretation yields the reading where Bill likes his own mother. Another example is *John has never read a Russian novel he disliked. But Bill has. It was War and Peace*,<sup>3</sup> taken from Gawron, Nerbonne & Peters (1991), where the strict interpretation is implausible because that would imply that John disliked a Russian novel that he never read.

This is without any doubt an interesting kind of ambiguity, and many computational solutions have been proposed for it (Dalrymple, Shieber & Pereira 1991; Bos 1994; Crouch 1995; Bos 2012). The question is how important this phenomenon is from a language technology point of view. To answer this question, I think it is good to look at naturally occurring data rather than examples invented by theoretical linguists. From the 554 cases of VPE that Bos & Spenader (2011) annotated in their one-million-word corpus, only *nine* show a *potential* ambiguity between a strict and sloppy interpretation. That is very little, perhaps even disappointingly little given the amount of theoretical work on the topic. What I am going to do in this article is to carefully study the behaviour of these nine cases. The main goal is to see whether the so-called sloppy identity is the rule or rather the exception.

## 3 Sloppy vs. strict in the Wall Street Journal corpus

Before I present the instances of VPE found by Bos & Spenader (2011), let me first introduce some notational conventions that I will use. For each occurrence of VPE, the antecedent VP is marked by square brackets, and the auxiliary verb triggering the elliptical VP is set in bold face. The pronoun causing the potential strict/sloppy ambiguity is underlined>. Co-referential phrases are indicated by printing the same indices (*i*, *j*) in subscript. The cases are listed in the order they appear in the Wall

---

<sup>2</sup> Dahl attributes the origin and name of the problem to J. R. Ross.

<sup>3</sup> Incidentally, this example is an instance of the *Missing Antecedent Problem*, because the antecedent of the neuter pronoun is not explicitly available in the discourse.

Street Journal corpus. All of them are presented with a reference to the location in the corpus, which is the name of the raw file as distributed by the Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993).

### 3.1 Carlos Saul Menem (sloppy vs. strict: 1–0)

The first instance we look at shows an odd kind of elliptical construction for a couple of reasons.<sup>4</sup> First, we have a present participle form of *do*, which is rather unusual for elided VPs. Second, there is a semantic mismatch between the parallel elements of the source and target clause: The subject of the source clause is a country (Brazil), whereas the subject of the target clause is a person (the President of Argentina). Here it is:

**Case 1: Carlos Saul Menem <wsj\_0415>**

If Brazil<sub>i</sub> devises an economic strategy allowing it to resume growth and service debt, this could lead it<sub>i</sub> to [VP open up and deregulate its<sub>i,j</sub> sheltered economy], analysts say, just as Argentinian President Carlos Saul Menem<sub>j</sub> has been **doing** even though he was elected on a populist platform.

The presence of the possessive pronoun *its* in the source clause, referring to Brazil, causes the potential strict/sloppy ambiguity. However, the target clause certainly doesn't mean that the Argentinian president opened up Brazil's economy – that would be highly unlikely – hence a strict interpretation is out of the question. The sloppy interpretation of *its* yields Argentina as antecedent, of course. Note however, that this antecedent isn't overtly expressed in the text, which is a further reason why this case is interesting.

### 3.2 IBM (sloppy vs. strict: 2–0)

This is very much like a standard textbook occurrence of VPE, where the source and target are connected via a temporal adverb:

**Case 2: IBM <wsj\_0445>**

IBM<sub>i</sub>, though long a leader in the Japanese mainframe business, didn't [VP introduce its<sub>i,j</sub> first PC in Japan] until five years after NEC<sub>j</sub> **did**, and that wasn't compatible even with the U.S. IBM standard.

The source clause contains the possessive pronoun *its* that co-refers with the subject, *IBM*. Hence, there is a potential ambiguity in the target clause. The strict variant

---

<sup>4</sup> Daniel Hardt, in email correspondence on 17-04-2007, says the following about this example: "This is a variant of VPE that has not been much studied as far as I know. I think that the *as* binds a variable standing for a VP meaning, very much like a wh-operator, as you could have for example in a variant of the above ...*which*<sub>i</sub> Argentinian President Carlos Saul Menem has been **doing**<sub>i</sub> even though he was elected on a populist platform. I think this would suggest that the missing VP should be linked up to the VP which *as* is modifying, in this case *open up and deregulate its sheltered economy*. And that is the reading I get."

Johan Bos

would be paraphrased as *until five years after NEC introduced IBM's PC in Japan*, and the sloppy one as *until five years after NEC introduced NEC's PC in Japan*. Of course, only the latter, sloppy interpretation makes sense – after all, why would one introduce a product of one's competitor?

### 3.3 Mr. Engelken (sloppy vs. strict: 3–0)

Here we have an instance of a *do-the-same* type of VP anaphor. Interesting here is that the subject of the source clause denotes a plural entity, whereas the subject of the target clause is a singular noun phrase:

Case 3: Mr. Engelken <wsj\_0758>

Some 34,320 fans<sub>*i*</sub> jammed the stands, and [<sub>VP</sub> shouted at the top of their<sub>*i,j*</sub> lungs]. Mr. Engelken<sub>*j*</sub> was **doing the same** across the Hudson River in New Jersey, where, with his nose pressed against the front window of the Passaic-Clifton National Bank, he watched the duel on a television set the bank set up for the event.

The potential strict/sloppy ambiguity here is caused by the plural possessive pronoun *their*. It is physically impossible to shout at the top of someone else's respiratory organs, at least in the preferred, non-literal sense of 'at the top of someone's lungs' that is obviously used here. Hence, there is no ambiguity here at all and the pronoun needs to be sloppily interpreted to get the desired reading. Note that the mismatch in number of the parallel subjects doesn't seem to matter at all to get the interpretation that Mr. Engelken was shouting at the top of his lungs.

### 3.4 Mr. Lawson (sloppy vs. strict: 4–0)

In order to fully comprehend the following case it might help to provide some context. It's 1989, we're in the UK. John Major has just been appointed Chancellor of the Exchequer, succeeding Nigel Lawson. Margaret Thatcher was Prime Minister at the time, as well as leader of the Conservative Party. Now consider:

Case 4: Mr. Lawson <wsj\_0883>

Neil Kinnock, Labor Party leader, dubbed the 46-year-old Mr. Major<sub>*i*</sub> a "lap dog" unlikely to [<sub>VP</sub>veer from his<sub>*i,j*</sub> boss's strongly held views], as Mr. Lawson<sub>*j*</sub> sometimes **did**.

Again we have a possessive pronoun, *his*, in the source clause, that co-refers with *Mr. Major*.<sup>5</sup> So Mr. Major isn't likely to veer from the views of his boss, Margaret Thatcher. And Mr. Lawson? He sometimes *did* veer from his boss's views, but of

<sup>5</sup> We note in passing that this is, in fact, an interesting control construction, as the subject of the source clause isn't explicitly expressed.

course the only sensible way this makes sense is that this be his own boss, not Mr. Major's. Now, it turns out the case that Mr. Major and Mr. Lawson had in fact the same boss, namely Margaret Thatcher. Hence, extensionally speaking, the strict and sloppy reading would both lead to the same interpretation anyway.

### 3.5 Mr. Turner (sloppy vs. strict: 5–0)

This case features the broadcasting company Comsat Video wishing to contract the Denver Nuggets Basketball team. Comsat Video happens to be a rival of Turner Broadcasting System Inc. Here is the example:

Case 5: Mr. Turner <wsj\_1461>

Comsat Video<sub>i</sub>, which distributes pay-per-view programs to hotel rooms, plans to [VP add Nuggets games to their<sub>i,j</sub> offerings], as Mr. Turner<sub>j</sub> did successfully with his Atlanta Hawks and Braves sports teams.

Once more we have a possessive pronoun causing a potential ambiguity. Obviously, Mr. Turner didn't add the Atlanta Hawks and Braves to the offerings of Comsat Video, but to his own company. Hence, only the sloppy interpretation makes sense here.

We are now halfway in discussing the potentially ambiguous VPE cases. So far they have been all sloppy – is there a chance for a strict reading? Let's see...

### 3.6 Americans (sloppy vs. strict: 6–0)

This is an example similar in structure and analysis to a case we considered earlier: a temporal adverb connecting the source with the target clause, and a possessive pronoun causing a potential strict/sloppy ambiguity:

Case 6: Americans <wsj\_1591>

“What this means is that Europeans<sub>i</sub> will [VP have these machines in their<sub>i,j</sub> offices] before Americans<sub>j</sub> do,” the spokesman said.

As it is absurd to think that Americans have machines in offices of Europeans, only the sloppy interpretation of the possessive pronoun *their* is available.

### 3.7 Democrats (sloppy vs. strict: 7–0)

Here we have a comparative construction coinciding with VPE. Both GOP senators and Democrats turn back a percentage of their allocated personal staff budgets.

Case 7: Democrats <wsj\_1695>

First, economists James Bennett and Thomas DiLorenzo find that GOP senators<sub>i</sub> [VP turn back roughly 10% more of their<sub>i,j</sub> allocated personal staff budgets] than Democrats<sub>j</sub> do.

Johan Bos

The strict interpretation would yield an interpretation where Democrats turn back allocated personal staff budgets of GOP senators, which isn't a realistic possibility. Hence the sloppy interpretation of the possessive pronoun is the only way to interpret it.

### 3.8 Most magazines (sloppy vs. strict: 8–0)

This is an interesting instance of VPE, because the subject of the source clause isn't overtly expressed. The target clause is recovered as *most magazines spread out ads among its articles*, where *its* co-refers, obviously, with the subject of the target clause, *most magazines*, not to National Geographic. In other words, once more we end up with a sloppy interpretation:

**Case 8: Most magazines** <wsj\_2109>

Another sticking point for advertisers was National Geographic<sub>*i*</sub>'s tradition of lumping its ads together, usually at the beginning or end of the magazine, rather than [<sub>VP</sub> spreading ads out among its<sub>*i,j*</sub> articles], as most magazines<sub>*j*</sub> **do**.

This case comprises, in addition, a complex nominalisation controlling the subject of the coordinated present participle constructions *lumping its ads together* and *spreading ads out among its articles*.

### 3.9 Competitors (sloppy vs. strict: 9–0)

This last case involves another comparative form of VPE, escorted by subject-auxiliary inversion in the target clause. The potential ambiguity is caused by a third-person plural pronoun. It is, in fact, an example of a "lazy" pronoun (Geach 1962), because it stands for a literal repetition of a full definite noun phrase:

**Case 9: Competitors** <wsj\_2109>

But the magazine was slower than its competitors to come up with its regional editions, and until last year [<sub>VP</sub> offered fewer of them] than **did** competitors.

The subject of the source clause is *the magazine*, which refers to National Geographic. The pronoun *them* refers to regional editions of the National Geographic. Hence, the items of comparison are the number of regional editions of the National Geographic offered by the National Geographic, and the number of regional editions of competitors (not the number of regional editions of the National Geographic, that would be ridiculous). Hence, apart from the complexity introduced by the comparative and the lazy pronoun, the analysis shows that, once again, we have a sloppy interpretation on our hands.

## 4 Discussion

The much discussed ambiguity between strict and sloppy interpretation caused by VPE is actually very rare in newswire text. Of more than 500 cases of VPE found in

the Wall Street Journal corpus, only nine cases showed potential sloppy/strict ambiguity. It turns out that all nine of them unequivocally show sloppy identity. It is true that we are working with a relatively small dataset of a restricted domain. Yet it is striking.

In eight out of nine cases a possessive pronoun caused the potential ambiguity. The balance between singular and plural number was equal: four in each case. The remaining case was a lazy pronoun referring to a possessive noun phrase. In the literature on VPE often examples with personal pronouns are given, but we found none in our corpus study.

Interestingly, several cases of surface semantic agreement conflicts were encountered. The Carlos Saul Menem example shows disagreement in the parallel subjects of the source and target clause. The Mr. Engelken example shows a mismatch in number between the source and target interpretation of the elided VP. Three of the VPE instances involve (complex) control constructions.

## 5 Conclusion

One could argue that any conclusion drawn from this dataset isn't significant because of its relatively small size. After all, we're only talking about *nine* examples. For the sake of the argument, let's assume that the distribution of strict and sloppy interpretations would be equally divided in texts. The odds to draw nine instances of VPE with a potential strict/sloppy ambiguity from a large corpus, which then turn out to be all of sloppy identity, are really low. So it is very likely that there is no equal distribution between sloppy and strict interpretation — informal Google searches confirm this claim.

It is certainly true that we need more empirical work and we should inspect larger and other genres of text. I have only looked at a very specific text genre: newswire. Journalistic prose is often associated with competent writing and governed by style guides. Often, texts are heavily edited for the sake of clarity and readability. In the case of the Wall Street Journal, its style guide gives the general advice to avoid ambiguity, however without saying anything in particular on pronouns and ellipsis (Martin 2002). Undeniably, empirical work on ellipsis should be extended to cover other genres of text, including spoken dialogue.

My tentative conclusion is that we don't need sophisticated algorithms in language technology, whose practitioners are content with accuracy figures of 90% or more given the inherent difficulty of the task, to compute all strict and sloppy interpretation for instances of VPE. First of all, because it is an extremely rare phenomenon, and secondly because if one defaults on sloppy identity a high accuracy is achieved already. As a consequence, computational implementations of ellipsis resolution algorithms could be far simpler than assumed so far. However, they could be more complicated with respect to other linguistic aspects, such as coordination, control, and mismatch between parallel elements.

## References

- Bos, Johan. 1993. VP ellipsis in a DRT-implementation. In *Proceedings of the sixth Conference of the European Chapter of the ACL (student session)*, 425–430. Utrecht, Netherlands.
- Bos, Johan. 1994. Presupposition & VP ellipsis. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling 1994)*, 1184–1190. Kyoto, Japan.
- Bos, Johan. 2012. Robust VP ellipsis resolution in DR Theory. In Staffan Larsson & Lars Borin (eds.), *From quantification to conversation*, vol. 19 (Tributes), 145–159. College Publications.
- Bos, Johan & Jennifer Spenader. 2011. An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation* 45(4). 463–494.
- Crouch, Richard. 1995. Ellipsis and quantification: a substitutional approach. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 229–236. Dublin, Ireland.
- Dahl, Östen. 1973. On so-called sloppy identity. *Synthese* 26. 81–112.
- Dalrymple, Mary, Stuart M. Shieber & Fernando C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy* 14. 399–452.
- Gawron, Jean Mark, John Nerbonne & Stanley Peters. 1991. The absorption principle and E-type anaphora. In *Proceedings of the 2nd Conference on Situation Theory and its Applications*. J. Mark Gawron, Gordon Plotkin & Syun Tutiya (eds.). Stanford. 335–362.
- Geach, P. T. 1962. *Reference and generality: an examination of some medieval and modern theories*. Cornell University Press.
- Hardt, Daniel. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 23(4). 525–541.
- Klein, Ewan. 1987. VP ellipsis in DR Theory. *Studies in Discourse Representation Theory and the Theory of Generalised Quantifiers*. Jeroen Groenendijk et al. (eds.).
- Marcus, M. P., B. Santorini & M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2). 313–330.
- Martin, Paul R. 2002. *The Wall Street Journal essential guide to business style and usage*. Wall Street Journal Books.
- Nerbonne, John. 1996. Computational semantics–linguistics and processing. In Shalom Lappin (ed.), *Handbook of contemporary semantic theory*, chap. 17, 461–484. London: Blackwell Publishers.
- Nielsen, Leif Arda. 2005. *A corpus-based study of verb phrase ellipsis identification and resolution*. King’s College London PhD thesis.
- Sag, Ivan. 1976. *Deletion and logical form*. MIT PhD thesis.

## Chapter 7

# *Om*-omission

Gosse Bouma

University of Groningen

The Dutch complementizer *om* is optional if the clause it introduces is a complement. We show that a large part of the variation in the distribution of *om* is accounted for by the governing verb. Syntactic complexity also plays a significant role, as well as semantic properties of the embedded clause.

### 1 Introduction

Dutch *to*-infinitival complement clauses (ICs) can be optionally introduced by the complementizer *om*. We find such ICs as dependents of verbs, nouns, adjectives, and prepositions, but here we will consider verbs only:

- (1) De Indiërs **aarzelen** (*om*) te investeren in Uganda  
The Indians hesitate (COMP) to invest in Uganda  
'The Indians hesitate to invest in Uganda.'

It seems highly unlikely that the presence or absence of *om* in examples like these in actual language use is totally random. For one thing, the governor (i.e. *aarzelen* in (1)) has a very strong effect on the probability that the IC is introduced by *om*. Another factor that might play a role is processing complexity. Processing complexity can be reduced by eliminating (local) ambiguity. The complementizer *om* explicitly marks the start of an IC. Therefore, one potential reason to use *om* is to disambiguate situations where the start of the IC is unclear

More in general, we might expect *om* to be used more often in sentences that are 'complex' in one way or another. Long sentences containing material that could be part of either the matrix clause or the IC, with many words intervening between the verbal governor and the verbal head of the IC, might contain *om* more often than 'simple', short, sentences.

An alternative, semantic, explanation might point to the fact that in (purpose or goal) modifier clauses, *om* is obligatory:

- (2) Omstanders duwden hem in een vijver om af te koelen  
Bystanders pushed him in a pond COMP PRT to cool  
'Bystanders pushed him into a pond to cool off.'

Historically, the use of *om* as a complementizer in modifier clauses precedes that of its use as complement marker (IJbema 2002). If this historical origin is still reflected in the current use of *om*, one expects *om* to be present especially in those ICs that bear some resemblance to purpose and goal modifier clauses. We investigate the role of two features that might be used to distinguish between typical complements of a verb and typical modifier clauses.

Jansen (1987) discusses the fact that prescriptive grammars until recently disapproved of the use of *om* in complement clauses, and also provides some corpus evidence for the fact that *om* is used more often in spoken (informal) language, suggesting that register and genre might play a role.<sup>1</sup> However, there has not been any corpus-based study into the distribution of *om* that investigates the features that influence the presence or absence of *om* in individual sentences. This is in strong contrast with a similar phenomenon in English, i.e. the optional presence of *that* in finite complement clauses, which has been the subject of numerous studies (see, among others, Ferreira & Dell (2000) and Hawkins (2002)). In particular, Roland, Elman & Ferreira (2006) observe that the strongest predictor for complementizer presence is the governing verb. Jaeger (2010) extends this result by showing that this effect can to a large extent be contributed to subcategorization frequency, in particular, the likelihood that a governing verb occurs with a complement clause in general.

## 2 Why add *om*?

There are two considerations that might explain why language users sometimes do and sometimes don't include *om*: processing complexity and semantics. A complementizer explicitly marks the beginning of an infinitival clause, and as such can help to reduce processing complexity. Roland, Elman & Ferreira (2006) observe that in English the verb governing the complement clause (cc) is important for predicting *that*. This in turn can be explained in terms of the probability that the governing verb selects for a cc: if a governing verb occurs with a cc often (i.e. of all occurrences of the verb, a high proportion is with a cc), the complementizer *that* will be omitted more easily. Jaeger (2010) gives a similar but more general account in terms of *information density*. One might argue that choice for the complementizer *om* in Dutch can be explained in a similar way. Furthermore, if reducing syntactic complexity is the driving force for choosing *om*, we expect factors such as length of the IC, distance (in words) between governor and IC, matrix clause type (i.e. verb final or not), and the presence of other complements to play a role as well.

One might also argue for a semantic account. Purpose and goal infinitival modifier clauses obligatorily are introduced by the complementizer *om*. Some verbs that take

<sup>1</sup> A comparison between the Corpus of Spoken Dutch and the newspaper corpus used in this study confirms that *om* is indeed more frequent in spoken language.

*om* as complement express a meaning that makes the complement clause very close in meaning to a purpose or goal clause:

- (3) De EU zal alles in het werk stellen om te helpen  
 The EU will everything in the work put COMP to help  
 ‘The EU will do everything it can to help.’

A semantic account predicts that complement clauses that are close in meaning to a goal or purpose clause, will more likely be introduced by *om*. The opposite idea is to measure how typical a combination of matrix verb and (the head of) an IC is. ICs headed by verbs that are ‘typical’ for a given matrix verb are probably less likely to be introduced by *om*.

### 3 Data

We used an 80 million word subset of the Twente Newspaper corpus (Ordelman et al. 2007) as corpus.<sup>2</sup> For computing semantic association scores, we used the full Twente Newspaper corpus (500 million words). The corpus was parsed automatically using Alpino (van Noord 2006). Using automatically parsed data has the advantage that it allows us to collect a large number of relevant examples quickly, including several features that might be relevant for predicting the distribution of *om*. We took several measures to ensure that the amount of noise is kept to a minimum.

Initially, we selected all sentences containing a TI (*te-infinitival*) or OTI (*om-te-infinitival*) clause functioning as verbal complement, i.e. with grammatical relation label VC in the dependency graph output by the parser.<sup>3</sup> We filter all examples involving governors that did not occur at least 10 times with a TI and at least 10 times with an OTI. We imposed this restriction to make sure that we are indeed considering examples where both forms are possible. We also filtered all cases where the governor (also) had a use as *cross-serial dependency* verb. An example is the verb *besluiten* (*to decide*):

- (4) ...waarna hij zich blijvend in de VS **besloot** te vestigen  
 after-which he himself permanent in the US decided to stay  
 ‘...after-which he decided to stay in the US permanently.’
- (5) ...waarna hij **besloot** (om) zich blijvend in de VS te vestigen

Example (4) exhibits cross-serial dependency word order where insertion of *om* is never possible. In (5), the IC is extraposed and *om* is possible. As the dependency structure of both cases is identical, it is hard to detect cross-serial cases automatically. To avoid confusion about the actual number of (extraposed, non cross-serial) TI cases, we decided not to include cases where the governor allows both word orders.

<sup>2</sup> Consisting of material from *Algemeen Dagblad* and *NRC Handelsblad*, 1994 and 1995.

<sup>3</sup> Subject TI and OTI clauses are rare and were ignored.

From the newspaper corpus, we collected 49,077 relevant sentences, containing an IC and a verbal governor that met the frequency and grammatical properties described above. 11,682 cases contain *om* (23%). 95 different verbal governors occur in the data, with a Zipfian frequency distribution, ranging from 9,287 (*besluit, decide*) to 26 (*beschouw, consider*).

### 4 Variables for predicting TI vs. OTI

In this section we present the various variables that we extract from the data to predict whether *om* is present in a particular sentence containing an IC.

23% of the 49K relevant examples in our corpus contains *om*. There are 95 different verbal governors, 23 of which have a preference for OTI over TI. 11 verbal governors occur with an OTI less than 10% of the time. It is well known that frequency of lexical items can have an effect on processing. If *om* is used to reduce processing complexity, we expect OTIs to occur relatively more often with low frequent governors than with high frequent governors. Figure 1 (left pane) illustrates that such a correlation indeed exists. The y-axis represents the log frequency of the verbal stem in our 80M newspaper corpus, and the x-axis represents the ratio of OTI against TI occurrences with this verbal stem as governor in our dataset. It shows that verbs that occur with *om* relatively often, tend to be low frequent. On the other hand, the right pane of Figure 1 shows that the correlation between the frequency with which a verb occurs with a verbal complement and the probability of complementizer presence (as observed for English) does not hold for our data.

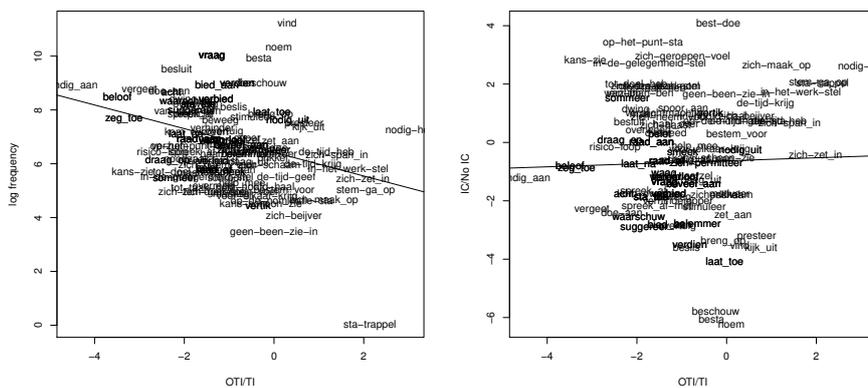


Figure 1: Overall log frequency of a verbal governor against ratio of OTI occurrence (left pane) and ratio of overall IC over non IC occurrence of verbal governors against ratio of OTI occurrence (right pane).

We expect *om* to show up especially in those cases where the start of the IC is hard

to recognize (locally) or where the sentence is just complex. Several features can be used as predictors for syntactic complexity: length of the IC (in number of words), relative position of the *te*-infinitive heading the IC from the start of the IC, syntactic category of the first constituent of the IC (*nominal, adverbial, verbal, or other*).

Table 1 lists the percentage of ICs for various distances between the governor and IC. ICs immediately following the governor have *om* in only in 20% of the cases, whereas for ICs at least two words away, the percentage of *om* is 28% or higher. Surprisingly, the lowest percentage of IC use is found with a distance of 1, i.e. with a single word intervening between the governor and the IC. We speculate that this is due to some peculiarities of Dutch word order, but at the moment have no clear explanation for this fact.

Table 1: Percentage OTI

distance	TI	OTI	% OTI	clause type	TI	OTI	% OTI
0	23,382	5,780	19.8	SMAIN	14,941	5,357	26.4
1	4,843	752	13.4	INF	4,342	1,552	26.3
2	3,437	1,387	28.7	SSUB	5,045	1,609	24.2
3	2,884	1,196	29.3	SV1	523	152	22.6
4	1,367	868	38.8	PPART	13,723	2,947	17.7
5	846	526	38.3	<i>average</i>	38,574	11,617	23.2
≥ 6	1694	1108	40.0				

(a) %OTI for various distances between between governor and start of the IC.

(b) %OTI for different clause types.

We can also look at the category of the clause headed by the verbal governor. If this is a finite main clause, we expect the percentage of *om* to be higher. Table 1b shows that our expectations are confirmed only to a certain extent. The highest percentage of OTIs is indeed found in main clauses, but it is only slightly higher than that for cases where the governor is infinitival or heading a (finite) subordinate clause. The lowest percentage of OTIs is found with participial verbal governors.

Complexity can also be caused by the presence of other complements in the matrix clause. Our data shows that the probability of OTI goes up strongly if an inherent reflexive (45.6% OTI), predicative complement (83.8%), or expletive *het* (50.3%) is present. Expletives are interesting, as they can be seen as placeholder for the IC. The majority of these cases occur with the governor *vinden*, which also selects for a predicative complement. Using binary features that measure the presence of such complements can be an alternative for using valence frames.

Distributional models of semantics determine the association strength between pairs of words, stems, phrases, and other linguistic units by means of statistical measures based on the relative frequency of occurrence of the individual units. For instance, the verb *eat* will occur relatively often with a subject that denotes an animate entity, and with an object that is edible. We can use this technique also to measure

how much a verbal governor is associated with the verbal head of its IC. The assumption is that, if the two are strongly associated, (the event described by) the IC is typical for this governor. In such cases, the need to use *om* might be less. The association score between a governor and the verbal head of its IC is computed as the *pointwise mutual information* (Church & Hanks 1990) between the two (where  $f(W)$  is the relative frequency of  $W$  in the corpus):

$$\text{pmi}(\text{Gov,IC-head}) = \ln \left( \frac{f(\text{Governor,IC-head})}{f(\text{Governor}) \cdot f(\text{IC-head})} \right) \quad (6)$$

Some verbs will occur in modifier OTI purpose clauses much more often than others. Such verbs express an event that is typical for a goal or purpose. If an IC is headed by such a verb, its semantics shares some resemblance with a purpose clause. We expect the probability of *om* to go up in such cases. Again, we use pointwise mutual information to measure the association between the modifier purpose clause and the verbal head:

$$\text{pmi}(\text{PurposeClause,Head}) = \ln \left( \frac{f(\text{PurposeClause,Head})}{f(\text{PurposeClause}) \cdot f(\text{Head})} \right) \quad (7)$$

To obtain the relevant statistics, we assume that all OTI constituents in the corpus that have the dependency relation MOD express a purpose or goal. Verbs and verbal expressions that are ranked high according to this measure are for instance: *kracht bij zetten* ‘to emphasize’, *erger voorkomen* ‘to limit the damage’, *het hoofd bieden (aan)* ‘to cope with’, *voorkomen* ‘to prevent’, *promoten* ‘to promote’, *beschermen tegen* ‘to protect against’.

## 5 Experiments

We describe experiments to determine which properties influence the choice for *om*, and how these properties interact. We used R and *lme4* (Bates, Maechler & Bolker 2011) to perform a linear mixed effects analysis, where verbs are random effects (see Baayen 2008).

We start with the situation that is perhaps most similar to English *that*-deletion, i.e. the distribution of *om* where the governing verb is finite and heading a main clause. In such cases, the governing verb is in second position in the sentence, while the IC is clause final. There are 19,862 relevant cases in our dataset, containing 94 different governors. We use the verb as random effect, where a verb is identified by its stem. As fixed effects, we used various features that might be indicators of syntactic or processing complexity.

The best model according to these assumptions (Table 2, *Main clauses only*) includes distance between governor and IC (*dist*), length of the  $\pi$ , distance between start of the  $\pi$  and the *te*-infinitive verb (*te*), and presence of expletive *het* (*het*). Numeric features were log-normalized and centered.

The negative intercept follows from the fact that the majority of cases do not have *om*. Longer distances between governor and IC, and between the start of the IC and

Table 2: Best model using verbal sense of the governor as random effect and various syntactic complexity features as fixed effects.

$$\text{Model} = \text{outcome} \sim \text{dist} + \text{length} + \text{te} + \text{het}(1 + \text{dist} + \text{length} + \text{te} + \text{het}|\text{sense})$$

	Main clauses only			All clause types		
	effect	std. err	significance	effect	std. err	significance
(Intercept)	-0.90	0.20	***	-0.98	0.13	***
<i>dist</i>	0.13	0.05	*	0.15	0.02	***
<i>length</i>	-0.13	0.03	***	-0.10	0.02	***
<i>te</i>	0.27	0.04	***	0.20	0.02	***
<i>het</i>	0.38	0.19	*	0.49	0.12	***

the *te*-infinitive verb, as well as the presence of expletive *het* all increase the likelihood of *om*. The overall length of the IC has a small negative effect. An anova test shows that the model improves significantly over a baseline model using only sense as random effect (Model AIC<sup>4</sup> = 15,716, Baseline AIC = 16,001,  $\chi^2 = 288.35$ ,  $p < 0.001$ ). Addition of various other potential features such as length and syntactic category of the first constituent of the IC, frequency of the head of the  $\tau_1$ , and presence of other syntactic dependents in the matrix clause (direct object, predicative phrase, reflexive, prepositional complement) did not improve the model significantly.

Next, we consider the complete dataset, i.e. also including cases where the governing verb is nonfinite or where the governor heads a subordinate clause. There are 49,077 cases in this set and 95 different verbal governors. Using the same model as for main clauses, we get the result given in table 2 (*All clause types*). The model outperforms the baseline significantly ( $\chi^2=549.88$ ,  $p < 0.001$ , Model AIC = 40,087, baseline AIC = 40,601). We found that including a categorical feature for clause type was in general not significant as soon as the feature measuring distance between governing verb and IC was also included.

To test our hypothesis that semantics might play a role, we use two features based on pointwise mutual information, as explained in Section 4. A model that uses only these two features as fixed effect is given in Table 3. The model confirms our expectation. If a  $\tau_1$  is headed by a verb that typically occurs in purpose/goal modifier clauses, the likelihood of *om* goes up, whereas if the verb heading the  $\tau_1$  co-occurs with the given governor often, the likelihood of *om* goes down. The model outperforms the baseline (using only the random effect) significantly ( $\chi^2 = 181.64$ ,  $p < 0.001$ , Model AIC = 40,433, baseline AIC = 40,601).

The model does not perform as well as the model using features inspired by syntactic and processing complexity considerations. Thus, complexity seems to play a more dominant role in the choice for *om* than semantics. A model using both complexity

<sup>4</sup> The Akaike Information Criterion is a measure for model fit based on Information Theory. Lower values indicate better model fit.

Table 3: Model and fixed effects for the complete dataset using semantic features.

$$\text{Model} = \text{outcome} \sim \text{complement} + \text{purpose} + (1 + \text{complement} + \text{purpose} | \text{stem})$$

	effect	std. err	significance
(Intercept)	-0.85	0.13	***
<i>complement</i>	-0.07	0.02	***
<i>purpose</i>	0.11	0.02	***

features and semantic features does perform better than the model using complexity features only ( $\chi^2 = 144.27$ ,  $p < 0.001$ , complexity + semantics model AIC = 39,973). The integrated model has a concordance (C) score of 0.809, which indicates that the model has modest predictive qualities.<sup>5</sup> We conclude that complexity and semantic factors both influence the choice for *om*.

## 6 Conclusions

In this paper, we have investigated the distribution of the complementizer *om* in *te*-infinitive complement clauses in Dutch using a large automatically parsed corpus. The matrix verb influences the likelihood of *om* significantly and thus we decided to use a mixed effects model, where the verb is used as random effect. Features that reflect syntactic complexity play a significant role. Semantic features that measure the similarity of the *te*-infinitive to typical complements for the given governor and to typical purpose or goal modifier clauses, play a significant role as well, although their effect is smaller than the ‘complexity’ features. A combination of ‘complexity’ and ‘semantic’ features gives rise to the best model.

We see a number of ways in which this work could be extended: manually corrected treebanks might give rise to more accurate data and stronger effects, medium and genre is likely to play a role,<sup>6</sup> but requires a balanced corpus, and finally, other measures for syntactic complexity (such as local and global sentence ambiguity according to a parser) could be explored.

## References

Baayen, R. H. 2008. *Analyzing linguistic data*. Cambridge University Press.

<sup>5</sup> The concordance scores measures for all pairs of a negative (TI) outcome and a positive (OTI), how often the model predicts a higher log-odds for the positive case. We used the function `somers2` from the `Hmisc` package.

<sup>6</sup> Indeed, we found that including the source newspaper as factor already had an effect.

- Bates, Douglas, Martin Maechler & Ben Bolker. 2011. *lme4: linear mixed-effects models using S4 classes*. R package version 0.999375-39. <http://CRAN.R-project.org/package=lme4>.
- Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics* 16(1). 22–29.
- Ferreira, V. S. & G. S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40(4). 296–340.
- Hawkins, J. A. 2002. Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics* 28(2). 95–150.
- Ijbema, Aniek. 2002. *Grammaticalization and infinitival complements in Dutch*. Leiden University PhD thesis.
- Jaeger, Florian. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* 61. 23–62.
- Jansen, F. 1987. Omtrent de om-trend. *Spektator* 17. 83–98.
- Ordelman, Roeland, Franciska de Jong, Arjan van Hessen & Hendri Hondorp. 2007. TwNC: a multifaceted Dutch news corpus. *ELRA Newsletter* 12(3/4). 4–7.
- Roland, D., J. L. Elman & V. S. Ferreira. 2006. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* 98(3). 245–272.
- van Noord, Gertjan. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister & Patrick Watrin (eds.), *TALN06. Verbum ex machina. Actes de la 13<sup>e</sup> Conference sur le Traitement Automatique des Langues Naturelles*, 20–42.



## Chapter 8

# Neural semantics

Harm Brouwer

Saarland University

Matthew W. Crocker

Saarland University

Noortje J. Venhuizen

Saarland University

The study of language is ultimately about meaning: how can meaning be constructed from linguistic signal, and how can it be represented? The human language comprehension system is highly efficient and accurate at attributing meaning to linguistic input. Hence, in trying to identify computational principles and representations for meaning construction, we should consider how these could be implemented at the neural level in the brain. Here, we introduce a framework for such a *neural semantics*. This framework offers meaning representations that are neurally plausible (can be implemented in neural hardware), expressive (capture negation, quantification, and modality), compositional (capture complex propositional meaning as the sum of its parts), graded (are probabilistic in nature), and inferential (allow for inferences beyond literal propositional content). Moreover, it is shown how these meaning representations can be constructed incrementally, on a word-by-word basis in a neurocomputational model of language processing.

### 1 Introduction

Language is about meaning. The aim of the study of language, therefore, is to capture and represent meaning, as well as to understand how it is constructed from linguistic input. Hence, albeit with different approaches and for different proximate goals, the fields of theoretical linguistics, computational linguistics, and psycholinguistics, all pursue systems that comprehend language. A successful comprehension system requires the representation of grammar, meaning, and knowledge about the world, be it in either a heterogeneous or homogeneous way (Nerbonne 1992). One system that is particularly effective and accurate at attributing meaning to linguistic input is

the human language comprehension system. Crucially, this system is implemented in the neural hardware of the brain. This suggests that, in trying to identify optimal computational principles and representations for meaning derivation, we may want to turn to how those principles and representations are implemented in neural hardware; that is, we may want to understand meaning construction and representation in terms of ‘brain-style computation’ by identifying a *neural semantics*.

A framework for neural semantics should minimally meet the following requirements:

- **neural plausibility**: the assumed computational principles and representations should be implementable at the neural level (cf. Rumelhart 1989);
- **expressivity**: the representations should capture necessary dimensions of meaning, such as negation, quantification, and modality (cf. Frege 1892);
- **compositionality**: the meaning of complex propositions should be derivable from the meaning of its parts (cf. Partee 1984);
- **gradedness**: meaning representations are probabilistic, rather than discrete in nature (cf. Spivey 2008);
- **inferential**: the derivation of utterance meaning entails (direct) inferences that go beyond literal propositional content (cf. Johnson-Laird 1983);
- **incrementality**: as natural language unfolds over time, representations should allow for incremental construction (cf. Tanenhaus et al. 1995).

In the present paper, we will introduce a framework for neural semantics that offers meaning representations that meet these requirements. Moreover, we show how these representations can be used within a neurocomputational model of language processing, to derive them incrementally, on a word-by-word basis for unfolding linguistic input.

## 2 A framework for neural semantics

In order to model story comprehension, Golden & Rumelhart (1993) developed a framework for modeling mental representations as points in a high-dimensional space called “situation-state space” (see also Golden et al. 1994). In their model, there is a one-to-one mapping between dimensions of the situation-state space and propositional meaning. Frank et al. (2003) extended this localist model for story comprehension by incorporating a distributed notion of propositional meaning. In what follows, we will introduce this Distributed Situation-state Space (DSS) model and show that it captures the aforementioned requirements for a neural semantics.

Table 1: Distributed Situation-state Space.

	proposition <sub>1</sub>	proposition <sub>2</sub>	proposition <sub>3</sub>	...	proposition <sub>n</sub>
observation <sub>1</sub>	1	0	0	...	1
observation <sub>2</sub>	0	1	1	...	1
observation <sub>3</sub>	1	1	0	...	0
...	.	.	.	...	.
observation <sub>m</sub>	0	1	0	...	0

## 2.1 Distributed situation-state space

A DSS is an  $m \times n$  matrix that is constituted of a large set of  $m$  observations of states-of-affairs in the world, defined in terms of  $n$  atomic propositions (e.g., *enter(john, restaurant)* and *order(ellen, wine)*)—the smallest discerning units of propositional meaning. Each of the  $m$  observations in this matrix is encoded by setting atomic propositions that are the case in a given observation to 1/True and those that are not to 0/False (see Table 1). The resulting situation-state space matrix is then effectively one big truth table, in which each column represents the *situation vector* for its corresponding atomic proposition—a point in situation-state space on a Euclidean perspective.

Situation vectors encode the meaning of propositions in terms of the observations in which they are the case. As a result, propositions that are the case in a similar set of observations obtain a similar meaning, whereas propositions that mostly occur in different observations obtain a dissimilar meaning. Crucially, the co-occurrence of propositions across the entire set of  $m$  observations in a DSS naturally captures world knowledge; that is, some propositions may never co-occur (hard constraints; e.g., a person can only be a single place at any given time), and some propositions may co-occur more often than others (probabilistic constraints; e.g., a person may prefer certain activities over other).

## 2.2 The DSS model as a neural semantics

The DSS-derived situation vectors inherently satisfy the aforementioned requirements. Firstly, situation vectors are **neurally plausible**. They can be represented at the neural level as firing patterns over neural ensembles, where vector components correspond to either the firing of single neurons or to the collective firing of neural populations.

Secondly, situation vectors are also **expressive** and **compositional**. The meaning of the negation of an atomic proposition  $a$ , for instance, is given by the situation

vector  $\vec{v}(\neg a)$  that assigns a 0 to all observations in which  $a$  is the case, and a 1 otherwise (thus resulting in a maximally different situation vector relative to  $\vec{v}(a)$ ); this vector can be directly derived from  $\vec{v}(a)$ , the situation vector of  $a$ , as follows:  $\vec{v}(\neg a) = 1 - \vec{v}(a)$ . In a similar manner, the meaning of the conjunction between propositions  $a$  and  $b$  will be described by the situation vector that assigns 1 to all observations in which both  $a$  and  $b$  are the case, and 0 otherwise; this vector can be calculated by the pointwise multiplication of the situation vectors of  $a$  and  $b$ :  $\vec{v}(a \wedge b) = \vec{v}(a)\vec{v}(b)$  for  $a \neq b$ , and  $\vec{v}(a \wedge a) = \vec{v}(a)$ .<sup>1</sup> Since the negation and conjunction operators together define a functionally complete system, the meaning of any other logical combination between propositions in situation-state space can be described using these two operations (in particular, the situation vector representing the disjunction between  $p$  and  $q$ ,  $\vec{v}(p \vee q)$ , is defined as  $\vec{v}(\neg(\neg p \wedge \neg q))$ , which assigns a 1 to all observations in which either  $p$  or  $q$  is the case, and a 0 otherwise). Hence, we can combine atomic propositions into complex propositions, which can in turn be combined with other atomic and complex propositions, thus allowing for situation vectors of arbitrary complexity (i.e., both  $a$  and  $b$  can be either atomic or complex propositions in the aforementioned equations). This means that we can *minimally* capture all meanings expressible in propositional logic.

Thirdly, situation vectors constitute **graded** representations; that is, they inherently encode the (co-)occurrence probability of propositions. On the basis of the  $m$  observations in the situation-state space matrix, we can estimate the *prior probability* of the occurrence of each (atomic or complex) proposition  $a$  in the microworld from its situation vector  $\vec{v}(a)$ :  $P(a) = \frac{1}{m} \sum_i \vec{v}_i(a)$ . Indeed, this probability is simply the number of observations in which proposition  $a$  is the case, divided by the total number of observations constituting the situation-state space. Similarly the co-occurrence probability of two propositions  $a$  and  $b$  can also be estimated from their corresponding vectors  $\vec{v}(a)$  and  $\vec{v}(b)$ :  $P(a \wedge b) = \frac{1}{m} \sum_i \vec{v}_i(a)\vec{v}_i(b)$  for  $a \neq b$ , and  $P(a \wedge a) = P(a)$ . Crucially, this means that we can also compute the conditional probability of proposition  $a$  given  $b$ :  $P(a|b) = \frac{P(a \wedge b)}{P(b)}$ . Hence, given a proposition  $b$ , we can **infer** any proposition  $a$  that depends on  $b$ . Taking this one step further, this allows us to define a comprehension score  $cs(a, b)$  that quantifies how much a proposition  $a$  is ‘understood’ from  $b$ : if  $P(a|b) > P(a)$ , then  $cs(a, b) = \frac{P(a|b) - P(a)}{1 - P(a)}$ , otherwise  $cs(a, b) = \frac{P(a|b) - P(a)}{P(a)}$ ; this score yields a value ranging from +1 to -1, where +1 indicates that event  $a$  is perfectly ‘understood’ to be the case from  $b$ , whereas a value of -1 indicates that  $a$  is perfectly ‘understood’ *not* to be the case from  $b$  (Frank, Haselager & van Rooij 2009).

In sum, DSS-derived situation vectors offer meaning representations that are neurally plausible, expressive and compositional, as well as graded and inferential. Hence, the DSS model meets five of the six requirements for a neural semantics. But what about the requirement of **incrementality**?

<sup>1</sup> This is to account for real-valued situation vectors, which may result from applying dimension reduction to a DSS.

### 3 Neural semantics in a neurocomputational model

Natural language unfolds over time, and the human language comprehension system incrementally attributes meaning to this unfolding input (see e.g., Tanenhaus et al. 1995). In what follows, we will show that the DSS-derived meaning representations allow for such incremental meaning construction; that is, we will show how situation vectors for linguistic input can be derived on a word-by-word basis in a neurocomputational model of language processing.

#### 3.1 A neurocomputational model

Our comprehension model is a Simple Recurrent Network (SRN; Elman 1990), consisting of three groups of artificial logistic dot-product neurons: an INPUT layer (22 units), HIDDEN layer (100), and OUTPUT layer (150). Time in the model is discrete, and at each processing time-step  $t$ , activation flows from the INPUT through the HIDDEN layer to the OUTPUT layer. In addition to the activation pattern at the INPUT layer, the HIDDEN layer also receives its own activation pattern at time-step  $t - 1$  as input (effectuated through an additional CONTEXT layer, which receives a copy of the activation pattern at the hidden layer prior to feedforward propagation). The HIDDEN and the OUTPUT layers both receive input from a bias unit. We trained the model using bounded gradient descent (Rohde 2002) to map sequences of localist word representations constituting the words of a sentence, onto a DSS-derived situation vector representing the meaning of that sentence (initial weight range:  $(-.5, +.5)$ ; zero error radius: 0.05; learning rate: 0.1, momentum: 0.9; epochs: 5000). After training, the overall performance of the model was perfect (each output vector has a higher cosine similarity to its target vector than to any other target vector in the training data).

#### 3.2 A microworld approach to DSS construction

The sentences on which the model is trained describe situations in a confined microworld (cf. Frank, Haselager & van Rooij 2009). This microworld is defined in terms of two persons  $P = \{john, ellen\}$ , two places  $X = \{restaurant, bar\}$ , and two types of food  $F = \{pizza, fries\}$  and drinks  $D = \{wine, beer\}$ , which can be combined into 26 atomic propositions using the following 7 predicates: *enter*( $P,X$ ), *ask\_menu*( $P$ ), *order*( $P,F/D$ ), *eat*( $P,F$ ), *drink*( $P,D$ ), *pay*( $P$ ) and *leave*( $P$ ). A DSS was constructed from these atomic propositions by sampling 10K observations (using a non-deterministic inference-based sampling algorithm), while taking into account hard and probabilistic constraints on proposition co-occurrence; for instance, a person can only enter a single place (hard), and *john* prefers to drink *beer* over *wine* (probabilistic). In order to employ situation vectors derived from this DSS in the SRN, we algorithmically selected 150 observations from these 10K that adequately reflected the structure of the world. Situations in the microworld were described using sentences from a microlanguage consisting of 22 words. The grammar of this

microlanguage generates a total of 176 sentences, including simple (NP VP) and coordinated (NP VP And VP) sentences. The sentence-initial NPs may be *john*, *ellen*, *someone*, or *everyone*, and the VPs map onto the aforementioned propositions. The corresponding situation vectors for these sentences were derived using the machinery discussed above. In particular, existentially quantified sentences such as *Someone entered the restaurant and left* map onto a vector corresponding to a disjunctive semantics:  $(enter(john, restaurant) \wedge leave(john)) \vee (enter(ellen, restaurant) \wedge leave(ellen))$ . Universally quantified sentences, in turn, obtain a conjunctive semantics, e.g., *Everyone left* maps onto  $leave(john) \wedge leave(ellen)$ .

### 3.3 Incremental Neural Semantics

On the basis of its linguistic input, the model incrementally constructs a situation vector capturing its meaning; that is, the model effectively navigates DSS on a word-by-word basis. This means that we can study what it ‘understands’ at each word of a sentence by computing comprehension scores for relevant propositions (i.e.,  $cs(a, b)$ ), where  $a$  is the vector of a proposition of interest, and  $b$  the output vector of the SRN. Figure 1 shows the word-by-word comprehension scores for the sentence *John entered the restaurant and ordered wine* with respect to 6 propositions. First of all, this figure shows that by the end of the sentence, the model has understood its meaning:  $enter(john, restaurant) \wedge order(john, wine)$ . What is more, it does so on an incremental basis: at the word *restaurant*, the model commits to the inference  $enter(john, restaurant)$ , which rules out  $enter(john, bar)$  (since these do not co-occur in the world;  $P = 0$ ). At the word *ordered*, the model finds itself in state that is closer to the inference that  $order(john, beer)$  than  $order(john, wine)$  (as John prefers beer over wine;  $(P = 0.63) > (P = 0.26)$ ). However, at the word *wine* this inference is reversed, and the model understands that  $order(john, wine)$  is the case, and that  $order(john, beer)$  is (probably) not the case. In addition, the word *wine* also leads the model to understand  $drink(john, wine)$ , even though this proposition is not explicitly part of the semantics of the sentence (John ordering wine is something that co-occurs relatively often with John drinking wine;  $P = 0.25$ ). Finally, no significant inferences are drawn about the unrelated proposition  $leave(ellen)$ .

## 4 Discussion

We have shown how the DSS model of story comprehension developed by Frank et al. (2003) can serve as a framework for neural semantics. This framework offers neurally plausible, expressive and compositional, as well as graded and inferential meaning representations. Moreover, we have shown how these meaning representations can be derived on a word-by-word basis in a neurocomputational model of language processing (see also Frank, Haselager & van Rooij 2009). Building in this direction, we are currently employing the framework to increase the coverage of a neurocomputational model of the electrophysiology of language comprehension (Brouwer 2014; Brouwer, Hoeks & Crocker 2015), to model script-based surprisal

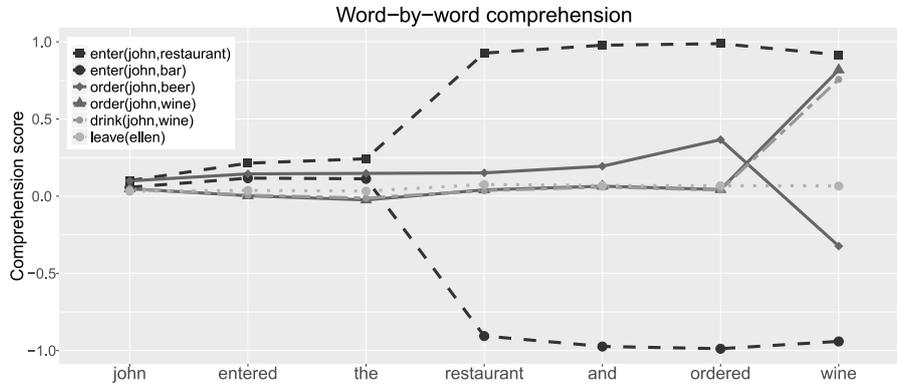


Figure 1: Word-by-word comprehension scores of selected propositions for the sentence *John entered the restaurant and ordered wine* with the semantics:  $enter(john, restaurant) \wedge order(john, wine)$  (see text for details).

(Venhuizen, Brouwer & Crocker 2016), and to model language production (Calvillo, Brouwer & Crocker 2016).

**Scalability.** The meaning representations that we employed in our neurocomputational model were derived from a DSS constituted of observations sampled from a microworld. For cognitive modeling, this microworld-strategy has the advantage that it renders it feasible to make all knowledge about the world available to a cognitive model, which is preferred over omitting or selecting relevant world knowledge. The outlined framework for neural semantics does, however, not hinge upon this microworld-strategy; that is, all machinery naturally scales up to larger DSSs. Crucially, this also true if situation vectors obtain real-valued components when dimension reduction techniques are used to render very large DSSs computationally manageable (i.e., all machinery extends to the domain of fuzzy logic). Hence, it is interesting to see how larger DSSs can be automatically constructed from large corpora of annotated data, such that the framework can be employed in wide-coverage natural language processing (NLP).

**Comparison to Distributional Semantics.** The use of the framework in large scale NLP raises the question how its distributed representations relate to those commonly used in the field of distributional semantics. In representations derived using techniques like Latent Semantic Analysis (LSA; Landauer & Dumais 1997), the representational currency are *words*. In DSSs, by contrast, the representational currency are *propositions*. Thus, instead of defining the meaning of a *word* in terms of the *words* that is co-occurs with, in the DSS model the meaning of a *proposition* is defined in terms of the *propositions* it co-occurs with. As result, the meaning representations naturally capture inferences driven by world knowledge, and are in addition expressive and compositional in nature.

**On the nature of atomic propositions.** In the DSS model, as presented in this paper, the smallest meaning-discerning units are atomic propositions (e.g.,

*order(john, beer)*). However, the DSS model does not enforce these units to be propositional in nature; that is, one may think of these units as the smallest meaning-discerning atoms in any relevant domain. For instance, if one were to model an embodied cognition perspective on language, certain atoms may reflect action-related meaning, others sensory-related meaning, and again others conceptual meaning, the co-occurrence of which encodes embodied meaning. Again, all machinery extends beyond propositional atoms.

## 5 Conclusion

We have described a framework for neural semantics that offers neurally plausible meaning representations. These representations directly reflect experience with the world, in terms of observations over meaning-discerning atoms. Complex meaning can be directly derived from these atoms (offering expressivity and compositionality). Moreover, the resulting meaning representations inherently carry probabilistic information about themselves and their relation to each other (gradedness and inferentiality). Finally, it was shown how these representations can be constructed on a word-by-word basis in a neurocomputational model of language processing (incrementality). This framework—which unifies ideas and techniques from theoretical linguistics, computational linguistics, and psycholinguistics—paves the way for a more comprehensive neural semantics. In future work, we will investigate how the approach can be extended with regard to formal semantic properties, linguistic coverage, and as part of larger neurocomputational models.

## References

- Brouwer, Harm. 2014. *The electrophysiology of language comprehension: a neurocomputational model*. University of Groningen PhD thesis.
- Brouwer, Harm, John C. J. Hoeks & Matthew W. Crocker. 2015. The electrophysiology of language comprehension: a neurocomputational model. In *Society for the neurobiology of language (SNL2015)*.
- Calvillo, Jesús, Harm Brouwer & Matthew W. Crocker. 2016. Connectionist semantic systematicity in language production. In Anna Papafragou, Daniel Grodner, Daniel Mirman & John C. Trueswell (eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2555–3560. Austin, TX.
- Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14(2). 179–211.
- Frank, Stefan L., Willem F. G. Haselager & Iris van Rooij. 2009. Connectionist semantic systematicity. *Cognition* 110(3). 358–379.
- Frank, Stefan L., Mathieu Koppen, Leo G. M. Noordman & Wietske Vonk. 2003. Modeling knowledge-based inferences in story comprehension. *Cognitive Science* 27(6). 875–910.
- Frege, Gottlob. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100. 25–50.

- Golden, Richard M. & David E. Rumelhart. 1993. A parallel distributed processing model of story comprehension and recall. *Discourse Processes* 16(3). 203–237.
- Golden, Richard M., David E. Rumelhart, Joseph Strickland & Alice Ting. 1994. Markov random fields for text comprehension. *Neural networks for knowledge representation and inference*. 283–309.
- Johnson-Laird, Philip N. 1983. *Mental models*. Cambridge University Press.
- Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211–240.
- Nerbonne, John A. 1992. Representing grammar, meaning and knowledge. In Susanne Preuß & Birte Schmitz (eds.), *Proceedings of the Berlin Workshop on Natural Language Processing and Knowledge Representation*. Berlin.
- Partee, Barbara. 1984. Compositionality. *Varieties of formal semantics* 3. 281–311.
- Rohde, Douglas L. T. 2002. *A connectionist model of sentence comprehension and production*. Carnegie Mellon University PhD thesis.
- Rumelhart, David E. 1989. The architecture of mind: a connectionist approach. In Michael I. Posner (ed.), *Foundations of cognitive science*, 133–159. Cambridge, MA: The MIT Press.
- Spivey, Michael J. 2008. *The continuity of mind*. Oxford University Press.
- Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard & Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217). 1632.
- Venhuizen, Noortje J., Harm Brouwer & Matthew W. Crocker. 2016. When the food arrives before the menu: modeling event-driven surprisal in language comprehension. In *Pre-CUNY Workshop on Events in Language and Cognition (ELC 2016)*.



## Chapter 9

# Liberating Dialectology

J. K. Chambers

University of Toronto

Until the 1980s, dialectologists simply accepted the fact that their investigations would produce “a superfluity of data”, as Kretzschmar, Schneider & Johnson (1989: v) pointed out. “Even smaller surveys have had to settle for selective analysis of their data,” they went on to say, “because the wealth of possibilities overran the editors’ time and the human capacity for holding in mind so much information at once.” That changed in the 1980s, when computers offered relief from the sorting problem, and advances in multivariate statistical methods began making inroads into the analysis of complex corpora. One of the disciplines that emerged at the intersection of computation and quantitative analysis was dialectometry, merging the joint progress in both areas.

### 1 The Groningen Initiative

What is sometimes missing in dialectometry is the humanistic interpretation that is so integral to the study of language, this most human of all attributes. Our sophistication in applying quantitative models to large corpora has greatly outpaced our capability for evaluating our results in terms of variable strength and social significance. One of the initiatives of John Nerbonne and his Groningen protégés has been, in Martijn Wieling’s phrase (2012: 3), “increasing the dialectology in dialectometry”. Dialectology in the twenty-first century is quantitative, variationist and social. Dialectometry applies quantitative models to large corpora, often to archival dialect data that were collected years before we had the capability for sorting the profusion of information let alone analyzing it. We now know more about those data than the dialectologists who gathered the data could ever imagine.

The great strength of dialectometry so far has come from the rigor it has been able to impose on the geographical distribution of linguistic variables. For the first hundred years after Georg Wenker inaugurated the systematic study of dialect in Germany in 1876, dialectology was largely qualitative and impressionistic. Its few

*J. K. Chambers*

governing concepts had their roots in common sense rather than empirical testing. A prime example is the “dialect continuum” (Chambers & Trudgill 1998: 5-7), the idea that “dialects on the outer edges of a geographical area may not be mutually intelligible, but they will be linked by a chain of mutual intelligibility”. The concept has its origin in the common observation that people who live near to one another normally speak more similarly than people who live further away. The idea of the dialect continuum was routinely invoked, selectively corroborated, and never problematized. The first empirical test of the concept came decades after it had become common coin. Heeringa & Nerbonne (2001) isolated a string of 27 Dutch towns and villages and calculated the aggregate pronunciation difference from one to the next (the “Levenshtein distance”). The dialect continuum survived their scrutiny but in a greatly nuanced conceptualization, with points of relative coherence and disruption correlated to some extent with cultural boundaries. The simple unexamined assumption of the dialect continuum emerged as a layered concept, and we can now imagine it being further examined in terms of internal forces, social and linguistic explanation, and comparative typology.

The next step, a crucial one surely, will be to rationalize gradations in the continuum by correlating them with social factors such as population, mobility and services. The Groningen initiative actively seeks this better balance. In a state-of-the-art summary, Wieling & Nerbonne (2015: 258) quote Lord Kelvin: “When you can measure, ... you know something; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.” Kelvin’s axiom is nicely illustrated, it seems to me, in applying actual measures to the dialect continuum, which has given us a richer concept than ever before. But Wieling and Nerbonne remind us that there is a further step. They go on to say that Kelvin “never suggested that measuring was sufficient”, a succinct reminder for all dialectometrists. The axiom expresses a crucial step for students of language. Measuring dialect has been a liberating force in dialectology.

## 2 The pioneer

A unique measure of how far we have come is afforded by looking backward to some moments when a few solitary souls recognized that dialectology could be dialectometric – quantifiable (not qualitative), relative (not absolute) and dynamic (not static). I would make the case that Jan Czekanowski (1882–1965) was the father, albeit an estranged one, of contemporary dialectometry. Indeed, he was practising dialectometry almost 45 years before Jan Séguy coined the term *dialectométrie* (1973). I will outline Czekanowski’s accomplishment, and document his momentary influence when his methods were applied by a couple of Americans to a purportedly intractable case of dialect heterogeneity. Czekanowski’s quantitative studies of dialect made little impression in his lifetime, and ultimately his insights were bypassed. Looking back reveals something about the core values of our discipline. Some fifty years after his death, dialectology has been reformed in exactly the terms he envisaged.

Jan Czekanowski is always identified as a Polish anthropologist but he was affli-



Figure 1: Jan Czekanowski.

ated with institutions in many nations (Payne 2014-2015), sometimes without physically moving. He was born in Głuchów, Poland, in 1882, and he began his primary education in Warsaw but completed it in Latvia. In 1902, he attended the University of Zurich and studied under the physical anthropologist Rudolf Martin (Oettinger 1926). Anthropological fieldwork took him to the Royal Museum in Berlin and to the Congo in Africa. He published the materials he collected in Africa in five volumes in 1910, while he was curator of the Ethnology Museum in St. Petersburg, Russia. In 1913, he became professor of anthropology at the University of Lwów, which was in Poland at the time (and in Ukraine from 1939). The photograph is probably from his early years at Lwów. He joined the University of Poznan in 1937 and retired there in 1946, when he was 64.

Czekanowski is remembered heroically because he convinced German “race scientists” in 1942 that the Karaim, a Polish-Lithuanian ethnic group that practised Judaism and used Hebrew as their liturgical language, were Turkic, not Semitic ([https://en.wikipedia.org/wiki/Jan\\_Czekanowski](https://en.wikipedia.org/wiki/Jan_Czekanowski)). The Karaim were accordingly spared from the Holocaust.

Czekanowski spent much of his academic life working on the racial categories of the human species, a predilection of physical anthropologists of his day, and judging by the sources I have found (and cited here) his abiding reputation among anthropologists is based on that work. Among linguists he is not well known nowadays, and the American linguists who noticed his work in the 1950s were affiliated with anthropology departments, a common affiliation at the time. Czekanowski’s achievement in linguistics, in retrospect, was no mere dalliance. He tried to establish the genetic and

cultural ties among branches of language families, that is, between dialect groups, and, against the grain of the times, he sought quantitative evidence for determining the relative strength of those ties. This line of research apparently started for him with inquiries that skirted linguistic matters. In the 1920s, he began with attempts at classifying cultural relatedness of ethnic groups by counting the cultural features they shared. He soon realized that shared linguistic features provide a solid basis. In 1927, he compared Polish dialects based on the morphological features they shared. In 1928, he broadened his purview and compared Indo-European dialects using the same methods. In 1929, he narrowed his view to Slavic dialects by the same method.

Czekanowski's method seems surprisingly sophisticated when we recall that he was working at a time when very few anthropologists and even fewer linguists counted anything at all. He began by compiling a list of linguistic features, and then for each pair of dialects he calculated correlation coefficients using a formula known as Q6, based on four factors: (1) the number of features present in both, (2) the number of features absent in both, (3) the number present in the first but absent in the second, and (4) the number absent in the first but present in the second.

He then arranged the dialects on a matrix according to their correlation coefficients. Figure 2 shows the pairwise comparisons of the Slavic dialects (Czekanowski 1931). It is a marvel of quantitative methods in its anticipation of multivariate statistics and a feat of calligraphy in its anticipation of computer graphics. The black squares show near-perfect correlations ( $> .80$ ) and the groupings of black squares represent branches of the family tree as a kind of intuitive cluster analysis. He uses Old Church Slavonic as the base (= *Starocerkiewnoslow* in Figure 2, an abbreviation of *Starocerkiewnoslowianski*, according to my colleague Alexei Kochetov). By tracing coordinates on the axes of the figure, one sees, for instance, that Slovak and Czech (Slowacki and Czeski) share many features ( $> .80$ ) and that Slovak and Malorussian (= *Małoruski*, a philological name for old Ukraine) share almost none ( $< .10$ ). Despite its apparent complexity, the figure is easy to read and very revealing.

It is tempting to linger over Czekanowski's diagram, admiring its elegance and its perspicacity. Indeed, those were the features that captivated the American anthropologists Alfred L. Kroeber and Douglas Chrétien, who praised it in esthetic terms (1937: 84): "If the symbol values are chosen judiciously, the diagram becomes an exceedingly effective and rapidly grasped representation of the stronger relationships, wherein the salient features of the classification force themselves upon the eye and the mind through the automatic clustering of symbols." But Czekanowski was apparently too far ahead of his time and his work made no immediate impact.

### 3 Spreading the word

Kroeber and Chrétien's advocacy of Czekanowski's methods did, however, make an impression in the more adventurous reaches of dialect study, if only momentarily. As American dialect research moved away from the long-settled Atlantic coast into the heterogeneous interior, the traditional atlas-oriented practitioners found themselves at a loss to identify patterns. Two of those practitioners, Alva Davis and Raven

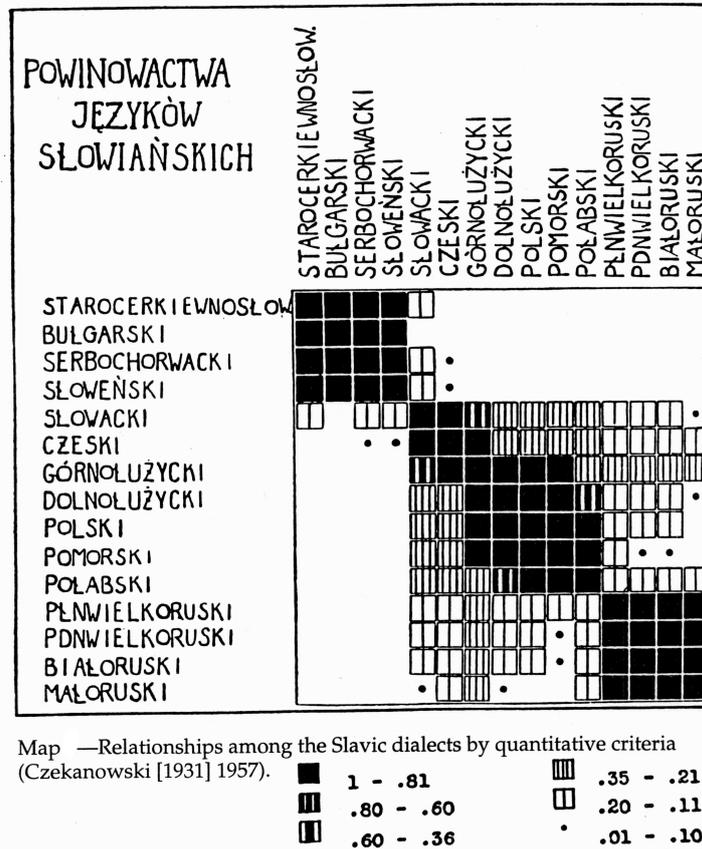


Figure 2: Czekanowski's quantitative arrangement of ancient Slavic dialects (1931).

McDavid, admitted defeat in a widely read article on “transition areas” in the journal *Language*. They wrote, “One is at a loss to give convincing reasons for the restriction of some items and the spreading of others” (1950: 186). In reaction, two of Chrétien’s students, David Reed and John Spicer, offered a rebuttal in the same journal (Reed & Spicer 1952: 348), pointing out that “the speech patterns of transition areas grow much clearer when viewed as quantitative rather than qualitative phenomena”, an axiom that two decades later would become the mantra of sociolinguistics and dialectology.

Reed and Spicer applied Czekanowski’s method to Davis and McDavid’s elicited features from ten subjects in five towns, and compared each pair of speakers using the Q6 formula to derive correlation coefficients. The quantification revealed some rough geographical generalizations, such as: the closer the town, the higher the coefficient, and the two southern towns are most similar to one another. But in the

end their mapping schema was very complicated and unrevealing because they did not choose a base dialect but presented all five as bases, and the variable patterns were hardly discernible because they required making inferences from five-way relationships. Their analysis definitely fails Kroeber's esthetic test: the schema does not provide a "rapidly grasped representation of the stronger relationships" and the "salient features" do not "force themselves upon the eye and the mind". With hindsight, we know that retaining geographical mapping obscured the patterns rather than revealing them, and bivariate statistics, the pairwise comparisons, required too many inferences for impressionistic interpretation.

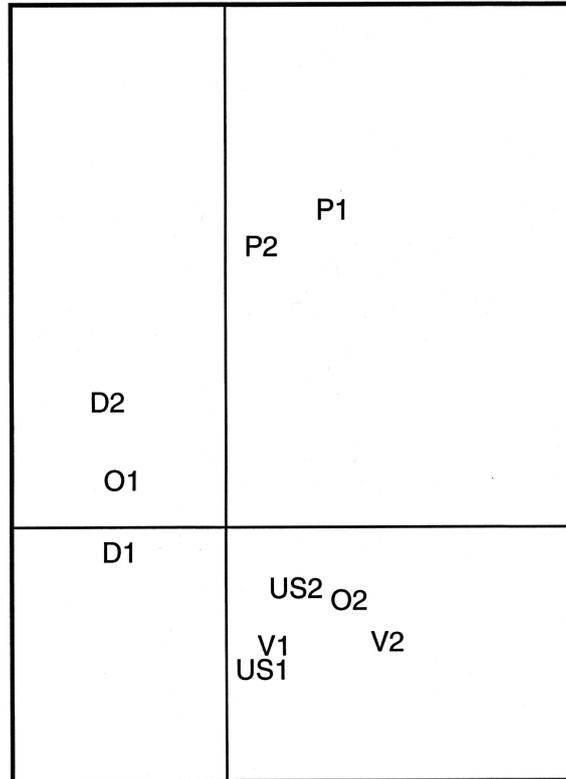
Fifty years later, I applied multivariate correspondence analysis to Davis and McDavid's data matrix and discovered the pattern fairly readily. Figure 3 places the ten speakers in quadrants according to their shared features. They clearly fall into three clusters, with P1 and P2 in one quadrant, five others together in the lower quadrant, and a messy little group of three spread along the left side. Identification of Davis and McDavid's dialect features shows that they are characteristically either Northern or Midland in their point of origin. Speakers in the top cluster (P1, P2) use essentially Northern features, and speakers in the bottom cluster (O2, V1, V2, US1 and US2) use essentially Midland features. The messy group along the side is mixed, a transitional group that uses some Northern and some Midland features.

Although the statistical input for the correspondence analysis encoded no geographic information whatsoever, the subcategories cohere strikingly to their relative locations on the map of northwestern Ohio: speakers P1 and P2, essentially Northern in their dialects, live in the northernmost town, and four of the five essentially Midland speakers live in the two southern towns. The mixed dialect speakers live in the two towns in between.

Why does geographical distance match statistical distance? It is hardly a mystery. As every dialectometrist knows, people who live close together tend to speak more like one another than people who live further away. We have come a long way since Jan Czekanowski, but dialectometrists like John Nerbonne have helped to reshape the study of dialects in exactly the terms that Czekanowski envisioned. Dialect distances, like geographic distances, are measurable things. Having established that beyond a doubt, dialectometrists are now taking the next step and increasing the dialectology in dialectometry.

## References

- Chambers, J. K. & Peter Trudgill. 1998. *Dialectometry*. 2nd edn. Cambridge, UK: Cambridge University Press.
- Czekanowski, Jan. 1931. *Różnicowanie się dialektów prastowiniskich w świetle kryterjum ilocowego [Differentiation of ancient Slavic dialects...]*. Prague: First Congress of Slavic Dialectology, 1928.
- Davis, Alva & Raven I. McDavid. 1950. Northwestern Ohio: a transition area. *Language* 26. 186–89.



Map —Multidimensional scaling of northwestern Ohio informants

Figure 3: Correspondence analysis of ten subjects (P1, P2, etc.) from five towns (P, D, O, US and V) in northwestern Ohio (Chambers & Trudgill 1998: 144-48).

Heeringa, Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13. 375–400.

Kretzschmar, William, Edgar Schneider & Ellen Johnson (eds.). 1989. *Journal of English Linguistics* 22, special edition: *Computer Methods in Dialectology*.

Kroeber, Alfred L. & C. Douglas Chrétien. 1937. Quantitative classification of Indo-European languages. *Language* 13. 83–103.

Oetteking, Bruno. 1926. <http://onlinelibrary.wiley.com/store/10.1525/aa.1926.28.2.02a00070/asset/aa.1926.28.2.02a00070.pdf?v=1&t=iph1ihz6&s=4b5bff98f70837980b38d3c478c4d4aeb73c1229> (June 2016).

Payne, Stephen. 2014-2015. *The Info List — Jan Czekanowski*. <http://www.theinfoлист.com/php/SummaryGet.php?FindGo=Jan%5C%20Czekanowski> (June 2016).

*J. K. Chambers*

- Reed, David W. & John L. Spicer. 1952. Correlation methods of comparing dialects in a transition area. *Language* 28. 348–59.
- Séguy, Jan. 1973. *Atlas linguistique de la Gascogne*. Vol. 6: *Notice explicative*. Paris: Centre national de la recherche scientifique.
- Wieling, Martijn. 2012. *A quantitative approach to social and geographical dialect variation* (Dissertations in Linguistics 103). Groningen: University of Groningen.
- Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.

## Chapter 10

# A new library for construction of automata

Jan Daciuk

Gdańsk University of Technology

We present a new library of functions that construct minimal, acyclic, deterministic, finite-state automata in the same format as the author's `fsa` package, and also accepted by the author's `fadd` library of functions that use finite-state automata as dictionaries in natural language processing.

## 1 Introduction

Finite-state automata (Hopcroft, Motwani & Ullman 2007) are widely used in Natural Language Processing (NLP). Their applications in the domain include morphology tools (e.g. `x`), tagging (Roche & Schabes 1995), approximate parsing (Nederhof 2000), information retrieval (Hobbs et al. 1997), and many more. The most prominent application is their use as dictionaries (Daciuk, Piskorski & Ristov 2010) for spelling correction, morphological analysis and synthesis, speech processing, semantic processing, etc.

Dictionaries in form of automata have been around for decades. The pioneering work has been done in the group of Maurice Gross at Université Marne La Vallée, by Tomasz Kowaltowski, Cláudio Lucchesi, and Jorge Stolfi at Universidade Estadual de Campinas, and by Martin Kay, Ronald Kaplan, Lauri Karttunen, and Kimmo Koskeniemi at Xerox. More software was developed later by other authors. That included our `fsa` package.

## 2 `fsa` package

We started developing `fsa` package during our stay in ISSCO, Geneva, Switzerland in the academic year 1995-1996. We were inspired by a lecture about transducers in NLP given by Lauri Karttunen in Archamps. The package was expanded and modified also afterwards to include new programs, new formats of automata representations, new functions. At present it includes:

*Jan Daciuk*

- `fsa_build` and `fsa_ubuild` – programs for constructing (various) dictionaries in form of finite-state automata. The first program accepts sorted data, the latter one – data in arbitrary order;
- `fsa_spell` – a dictionary-based program for spelling correction;
- `fra_morph` – a dictionary-based program for morphological analysis, and for lemmatization;
- `fsa_guess` – a program for morphological analysis, and for lemmatization of words not present in a lexicon, as well as for guessing morphological descriptions of unknown words so that they could be added to a morphological dictionary;
- `fsa_synth` – a dictionary-based program for morphological synthesis;
- `fsa_accent` – a dictionary-based program for restoring missing diacritics in words;
- `fsa_prefix` – a tool for listing the contents of a dictionary;
- `fsa_visual` – a tool for preparing data for a program that presents a dictionary as a graph (obviously, that makes sense only for tiny dictionaries).

A companion library `fadd` was written during our postdoc at Rijksuniversiteit Groningen from February 2000 to January 2003. It features the same functions as programs of the `fsa` package except for programs `fsa_build` and `fsa_ubuild`. It also contains handling of compressed language models.

The `fsa` package has been successfully used by many people, but there are several problems with it. They are listed in the following subsections.

## 2.1 Difficult maintenance

The first program of the package (`fsa_build`) was written in 1995, and it was our first program in C++. Each function has a comment explaining the purpose of the function, the parameters, the return value, and remarks on possible assumptions, like e.g. that a file must be opened for reading, or that memory must have been allocated before. There are also plenty of comments inside functions. However, functions are very long. The longest ones stretch over 398, 360, 257, 228, 210, 210, and 200 lines, not counting the initial comments. Even though some of those lines are comment lines, understanding such code is very difficult and time consuming.

We wanted to make the code as efficient as possible, at the same time exploring various representations of automata. As differences between particular representations are small, but manifest themselves in different parts of the programs, this resulted in interwoven conditional compilation directives. Some of those directives exist only to handle historical formats that have no advantages in comparison with modern ones. Long functions are difficult to understand, but long functions with conditional compilation directives are an order of magnitude more difficult.

## 2.2 Memory requirements

One of the features that made the package popular was the use of incremental construction algorithms. Contrary to other algorithms used at that time, they had very low memory requirements.

When new features were added to the package, especially the features that had to be introduced during construction, new vectors or new fields in vectors were introduced to handle them. That increased memory requirements of the construction programs `fsa_build` and `fsa_ubuild`. It seemed at that time that the size of possible dictionaries was quite limited, and small increases in vocabulary would be more than compensated by cheap memory of modern computers. One of the main complaints during our Ph.D. defence, that was also voiced many times afterwards at various conferences, was: “Why are you doing this?! Memories are so big and cheap, and they are getting bigger and cheaper, it is easier to write an application for a grant than to learn to use new software.”

It turned out that data grows faster than memories. Researchers at the University of Technology in Brno complained that their dictionaries (based on Wikipedia) were too big to be constructed even in huge virtual memory.

## 2.3 Stand-alone programs

The package was written as a set of stand-alone programs. Library `fadd` written in Groningen at the request of Gertjan van Noord contains functions of most of the programs of the `fsa` package, but the automata construction programs `fsa_build` and `fsa_ubuild` were left out. They were the most difficult to re-implement, and it seemed at that time that those functions were not really needed, as dictionaries are constructed once for a very long time, and then used frequently. That assumption is correct, but we want to write a package for construction of tree automata. The tree automata are to be compressed, including their labels, and the way to do that is to use finite-state automata. It is awkward to call external programs inside other programs, so we need a library.

# 3 New package

The new library called `fsacl` has been in plans for a few years. Since ideas and good will alone are not sufficient to write software, we had to wait till we get a little spare time. The development began in December 2015, stopped after one month, and resumed in mid July 2016.

## 3.1 Requirements

The new library:

- should offer both C++ and C interface in a similar way to `fadd` in order to facilitate its use in various programming languages;

*Jan Daciuk*

- should implement at least two incremental construction algorithms;
- should implement automata representation version 5 from package `fsa` with hash numbers;
- should also implement sparse matrix representation;
- should implement new representation based on (Daciuk & Weiss 2012);
- should use less memory than `fsa_build` and `fsa_ubuild` during construction.

### 3.2 Present status

The library is under development. To test the library, two programs: `nfsa_build` and `nfsa_ubuild` have been developed. They have roughly the same functions as `fsa_build` and `fsa_ubuild`, respectively, from the `fsa` package. The main difference between the new programs and the old ones is that various behavior achieved in old programs by using different compile options is now controlled by different command-line options. Those additional command-line options switch on and off various features, like e.g. sorting transitions on frequency of their labels, or select representation format version. Not all of those options are implemented in the current version of the library.

Construction of automata from sorted and unsorted input has been implemented and tested both with and without option “-0”, i.e. with or without storing some states inside other states. The constructed automaton can be read with programs from the `fsa` package or by the `fadd` library. Only one representation version – version number 5 (list representation of outgoing transitions, with STOP bit flags marking last outgoing transitions of a state, with NEXT bit flag indicating that the target of the present transition is located right after the transition, and without the TAIL bit flag indicating that the remaining transitions of a state are stored in the location specified by the following number). This is the most popular representation in the `fsa` package, and the one that usually gives the best results. It is also the default one. It is also implemented in the `fadd` library.

Constructing guessing automata is partially implemented; compile options `GENERALIZE` and `PRUNE_ARCS` from the `fsa` package are missing. The code is not tested, so it almost certainly contains errors.

Storing numbers for hashing has been implemented and tested. Various non-essential compile options like `PROGRESS` (for showing the progress of construction), `STATISTICS` (for showing statistics on states, transitions, chains of states and transitions, etc.), and `WEIGHTED` (not really for weighted automata, but for producing allegedly better guessing automata) are not implemented. Memory is not de-allocated as it should in a library.

Preliminary results show that `nfsa_build` is significantly faster and less memory-efficient than `fsa_build`. The reason for that must be found by profiling and testing. We have chosen to use linked lists of outgoing transitions. We wanted to achieve a speed-up in adding an outgoing transition to a state, and we did that, but the field

used for linking takes additional memory that is not used in the `fsa` package. Our suspicion is also that use of STL vectors is responsible for additional memory, as the library allocates twice as much memory for a vector when it becomes full.

### 3.3 Additional issues

Good software cannot be written and just left alone. It must be distributed and maintained. It means not only that there is someone who corrects errors when he or she finds them. The software must either be a part of some bigger collections, e.g. become a Linux package in some Linux distribution, or it must be available from a fixed location, with a fixed e-mail address of the author/maintainer.

For a quarter of a century, the `fsa` package, and our other software packages (`utr` – same as `fsa` but implemented with transducers), `fadd` library and other minor pieces of software were available from `ftp.pg.gda.pl` – the FTP server of Gdańsk University of Technology. The author’s address was either `jandac@pg.gda.pl` or `jandac@eti.pg.gda.pl`, with both addresses working.

That pleasant constancy is gone. The FTP server is maintained by an incompetent and even hostile crew. It was switched off without warning in the beginning of 2016 for several months, with complaints about its unavailability quietly ignored. Web pages describing the software are hosted on the faculty web server. That server has recently been taken over by the same crew that maintains the FTP server. It also has new software for managing web pages, which works only partially. The old server is still available, but it will be switched off at the end of August 2016 to force people to use the new sever. Finally, the university decided that it should be located in an edu domain, rather than in a geographical one (`gda` from Gdańsk). To save costs, the old domain (`pg.gda.pl`) will be removed. This will make redirection impossible, and it will affect all addresses: web pages, the ftp server, e-mail addresses.

To provide constant addresses again, we decided to buy a domain and a place on a commercial server. The domain `jandaciuk.pl` is already bought, and our software is downloaded to the FTP server, and the web pages are installed. It turned out that the commercial FTP server does not allow for anonymous FTP connections, so we use the HTTP protocol instead.

### 3.4 Plans for the future

We plan to:

1. Implement missing functions and options (compile options implemented as run-time options). The most urgent among those is a set of options and related functions that deal with construction of guessing automata. Guessing automata can be created using the present version of the library, but they are bigger than those created by programs `fsa_build` and `fsa_ubuild` from the `fsa` package.
2. Significantly reduce memory use. This is one of the objectives for designing this library.

Jan Daciuk

3. Implement representation format 10 of automata in the `fsa` package. This format is the same as format number 5 (already implemented; it uses flags FINAL, STOP, and NEXT) for annotations, and it uses sparse matrix representation for words. At this point, the library will be able to completely replace the construction part of the `fsa` package for two most popular formats.
4. Implement a new representation format that gives much smaller automata (measured in bytes) that can be constructed in shorter time. The representation will be similar to the one described in (Daciuk & Weiss 2012), but with a modification that will make the construction faster.

We also consider moving construction of compressed language models from our `fadd` library to this one.

## References

- Daciuk, Jan, Jakub Piskorski & Strahil Ristov. 2010. Natural language dictionaries implemented as finite automata. In Carlos Martín-Vide (ed.), *Scientific applications of language methods*. Imperial College Press.
- Daciuk, Jan & Dawid Weiss. 2012. Smaller representation of finite-state automata. *Theoretical Computer Science* 450. 10–21.
- Hobbs, Jerry R., Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson. 1997. FASTUS: a cascaded finite-state transducer for extracting information from natural language text. In Emmanuel Roche & Yves Schabes (eds.), *Finite-state language processing*. MIT Press.
- Hopcroft, John E., Rajeev Motwani & Jeffrey D. Ullman. 2007. *Introduction to automata theory, languages and computation*. 3rd edn. Pearson International Edition.
- Nederhof, Mark-Jan. 2000. Practical experiments with regular approximation of context-free languages. *Computational Linguistics*.
- Roche, Emmanuel & Yves Schabes. 1995. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics* 21(2). 227–253.

## Chapter 11

# Generating English paraphrases from logic

Dan Flickinger  
Stanford University

A set of semantic transfer-based paraphrase rules is developed and applied using an existing broad-coverage grammar and efficient generator to implement the automatic production of a wide variety of English sentences from input propositions in quantifier-free first-order logic within the “blocks language” of Tarski’s World, a component of the course Language, Proof and Logic.

### 1 Introduction

A student learning to express propositions in formal logic is typically presented with practice exercises where a proposition is expressed in a natural language such as English, and the task is to construct the corresponding logical form. In an automated instructional system, it would be desirable to be able to construct a new target logical form on demand, and automatically generate from this expression a natural language sentence to present to the student as a new exercise. Similarly, it would be helpful to be able to take the student’s incorrect attempt at a solution, and generate a natural language sentence to show the student what the proposed solution actually says, as one way of prompting repair. Since the mapping between natural language and logic can vary in transparency even for relatively simple logical systems, it would be convenient to generate, in fact, a diverse set of annotated paraphrases so that the instructional software could select a paraphrase appropriate for the student’s current level of proficiency.

One widely used instructional software system for teaching first-order logic (FOL) is provided as part of the textbook *Language, proof and logic* by Barker-Plummer, Barwise & Etchemendy (2011). One component of this software is called Tarski’s World, which defines a “blocks language” used for many of the exercises in the book, with fewer than twenty predicates, varying in arity from one to three arguments, along with boolean and conditional connectives, and quantifiers. Some of these exercises present the student with an English language proposition such as “**b** is a cube” to be

*Dan Flickinger*

translated into this blocks language, where the software system can automatically evaluate the correctness of the student's submitted solution. For the first stage of developing an English generator from FOL propositions in Tarski's World, quantifiers have been excluded, but otherwise the full range of expressive variation illustrated in the textbook is included.<sup>1</sup>

The objective for the present study was to produce an accurate and robust implementation of an English generator which can take as input any well-formed FOL expression (excluding quantifiers) within the blocks language of Tarski's World, and produce a rich set of English paraphrases each of which translates back to the input FOL, and ideally only to that FOL. The next section describes the grammar-based method employed in this implementation, including descriptions of the existing resources adapted for this task. Section 3 presents the set of paraphrase rules and the declarative formalism used to define them, along with examples of output from the generator. The final two sections include a discussion of how the generator's output has been evaluated to date, and provide some larger context for this effort in related work on generation from logic or other formal languages.

## 2 Method

In the approach adopted here, the generation system accepts an input proposition in FOL, converts it to the grammar-specific semantic representation, and then applies a set of paraphrase rules to produce alternate semantics, each of which is then presented to the generator itself to produce a set of English sentences which realize that semantic representation. Each output sentence carries an annotation identifying which if any of the paraphrase rules were applied to produce its semantics, so the instructional system can make a more informed choice about which of the alternative English outputs to present to the student based on properties of the exercise or of the student's proficiency.

### 2.1 Component resources

The FOL-to-English implementation draws on two substantial pre-existing resources developed within the DELPH-IN consortium ([www.delph-in.net](http://www.delph-in.net)): the ACE parser and generator ([mo.in.delph-in.net/AceTop](http://mo.in.delph-in.net/AceTop)) developed by Woodley Packard, and the English Resource Grammar (ERG: Flickinger 2000; 2011), a linguistically rich broad-coverage grammar implementation within the framework of Head-driven Phrase Structure Grammar (HPSG: Pollard & Sag 1994). The ACE engine is a more efficient re-implementation of the chart parser and generator of the LKB (Copestake 2002; Carroll et al. 1999), applying the rules of a grammar such as the ERG either to an input sentence in order to output semantics (parsing), or to an input semantic representation in order to output sentences (generating). Both of these general-purpose

---

<sup>1</sup> The "blocks language" predicates are *Cube*, *Tet*, *Dodec*, *Small*, *Medium*, *Large*, *Smaller*, *Larger*, *LeftOf*, *RightOf*, *BackOf*, *FrontOf*, *SameSize*, *SameShape*, *SameRow*, *SameCol*, *Adjoins*, *=*, *Between*.

resources were used essentially unchanged for this task, apart from minor lexical additions for the Tarski’s World domain.

## 2.2 FOL to MRS

The semantic framework adopted within the ERG is called Minimal Recursion Semantics (MRS: Copestake et al. 2005), so an input FOL proposition for this task is first automatically converted to a ‘skeletal’ MRS by a combination of a simple reformatting utility<sup>2</sup> and then a small set of declarative MRS-to-MRS rules normalize to grammar-internal predicate names, and assign default values for tense/aspect/mood on verbal predications, and person/number plus definiteness for nominal predications. This fully-specified MRS can be given as input to the generator, which will produce one or more English sentences which realize the MRS. Here is a simple example of these three basic steps, before we turn to paraphrases:

FOL:            `Large(a)&Large(b)`

Skeletal MRS:

```

LTOP: h1
INDEX: e1
RELS: < [ "name" LBL: h3 ARG0: x1 CARG: "A" ]
        [ "large" LBL: h4 ARG0: e2 ARG1: x1 ]
        [ "name" LBL: h5 ARG0: x2 CARG: "B" ]
        [ "large" LBL: h6 ARG0: e3 ARG1: x2 ]
        [ "and" LBL: h2 ARG0: e1 L-INDEX: e2 R-INDEX: e3 ] >

```

Full MRS:

```

LTOP: h20
INDEX: e13 [SORT: collective SF: prop TENSE: pres PERF: -]
RELS: <
  [ named LBL: h5 ARG0: x10 [PERS: 3 NUM: sg] CARG: "A" ]
  [ named LBL: h9 ARG0: x11 [PERS: 3 NUM: sg] CARG: "B" ]
  [ proper_q LBL: h2 ARG0: x10 RSTR: h3 BODY: h4 ]
  [ proper_q LBL: h6 ARG0: x11 RSTR: h7 BODY: h8 ]
  [ _large_a_1 LBL: h18 ARG0: e14 [SF: prop TENSE: pres PERF: -]
    ARG1: x10 ]
  [ _large_a_1 LBL: h19 ARG0: e15 [SF: prop TENSE: pres PERF: -]
    ARG1: x11 ]
  [ _and_c LBL: h12 ARG0: e13 L-INDEX: e14 R-INDEX: e15
    L-HNDL: h16 R-HNDL: h17 ] >
HCONS: < h3 qeq h5 h7 qeq h9 h16 qeq h18 h17 qeq h19 >

```

<sup>2</sup> Thanks to Aaron Kalb for the Python script.

*Dan Flickinger*

English realizations from the generator:

*a is large and b is large.*

*a is large, and b is large.*

In this example, the constants *a* and *b* are mapped to “name” elementary predications (EPs) in the ‘skeletal’ MRS; the one-place predicate *Large* is mapped to a one-place EP (plus the inherent ARG0 event variable); and the conjunction symbol is mapped to a two-place EP with its arguments the ARG0 values of the two “large” EPs. The full MRS in this simple example still contains in its RELS list these five EPs normalized with grammar-internal predicate names, with default values filled in for tense, number, etc., and with quantifier EPs added to bind each of the two “named” EP ARG0 instance variables, to satisfy general constraints on MRS well-formedness.<sup>3</sup>

### 2.3 MRS to MRS for paraphrases

In order to produce non-trivial paraphrases of the sentences that correspond to this ‘literal’ mapping from FOL to MRS, a set of MRS-to-MRS mapping rules have been developed, and some or all of these can be applied to the original MRS to produce a (potentially large) set of alternate MRSs, each of which can be given as input to the generator for realization into English sentences. These MRS-to-MRS rules employ a rule specification formalism originally developed by Stephan Oepen for machine translation within the LOGON research project (Lønning et al. 2004), where the formalism allows each ‘translation’ rule to specify one (often partial) MRS expression as *input* and another MRS expression as *output*, along with optional positive and negative constraints on other elements within the full MRS being ‘translated’. For the present task, the rule set serves to map one English MRS to one or more output English MRSs that may be realized by the generator as paraphrases of sentences realized for the input MRS.

To continue with the above example, the paraphrase rules include ones for mapping a conjunction of two clauses with a common subpart into a single clause with either a conjoined noun phrase or a conjoined verb phrase. Thus the original MRS realized as *a is large and b is large* can be mapped via these rules into an MRS which is realized as follows:

*a and b are large.*

*Both a and b are large.*

Similarly, the rule set includes mappings that can reverse the order of the conjuncts, to produce MRSs that will be realized as follows:

*b is large and a is large.*

*b is large, and a is large.*

---

<sup>3</sup> The full MRS includes, in addition to the outermost label *LTOP* and event variable *INDEX*, a set of scope constraints in *HCONS* which are not relevant for this discussion, but which will be essential as this work is extended to accommodate FOL expressions with quantifiers.

*b and a are large.*

*Both b and a are large.*

Examples of the paraphrase rules themselves are presented in the next section.

### 3 Paraphrase rules

Inspection of the exercise example sentences in the LPL textbook reveals a number of sources of linguistic variation in the English expression of propositions within the blocks language for Tarski's World, including the following:

- coordination/aggregation;
  - nominal phrases (*b is in front of a cube and a tetrahedron*);
  - predicates (*b is a cube and is large*);
- negation (*It is not the case that a and b are large*);
- pronouns (*If b is a cube, it is large*);
- partitives (*a and b are large, and both of them are cubes*);
- VP ellipsis (*If b is large, then c is*);
- sentence connectives (*if and only if, just in case, unless*);
- adjectives as pre-modifiers (*b is a large cube*);
- adverb addition (*If a is large, b is also large*);
- reordering (*a and b are cubes – b and a are cubes*).

Implementation of the MRS-to-MRS mapping to enable these variations resulted in a set of 143 paraphrase rules,<sup>4</sup> defined as a hierarchy of types within the LOGON-based formalism for semantic transfer, which is supported by the ACE engine.

An example of one of these paraphrase rule types accommodates the generation of sentences with verb-phrase ellipsis, as in the following alternation:

*b is large and c is large*

*b is large and c is*

The following rule applies to an input MRS where two event EPs have the same predicate name (the value of the PRED attribute in an EP), identifying in the CONTEXT attribute the first EP which will survive the rule's application unchanged, and in the INPUT attribute the second EP (and possibly additional EPs for that second verb phrase) which will be replaced by the ellipsis EP in the OUTPUT attribute.

---

<sup>4</sup> These rule definitions are included in the most recent version of the open-source ERG, within the subdirectory 'openproof', available at [www.delph-in.net/erg](http://www.delph-in.net/erg).

*Dan Flickinger*

```
basic_vp_ellipsis_gpr := monotonic_mtr &
[ CONTEXT [ RELS < [ PRED #pred, ARG0 event ] > ],
  INPUT   [ RELS < [ PRED #pred, LBL #h1,
                    ARG0 #e2 & event,
                    ARG1 #x3 & ref-ind ], ... > ],
  OUTPUT  [ RELS < [ PRED ellipsis_ref, LBL #h1,
                    ARG0 #e2,
                    ARG1 #x3 ] > ] ].
```

This output ellipsis EP preserves the inherent argument (ARG0) and the external (subject) argument (ARG1) of the deleted EP, to ensure the correct tense and agreement for the form of the verb *be* which will realize this ellipsis EP in the sentences generated from the resulting MRS.

The set of paraphrase rules is partially ordered to ensure that certain desired feeding relationships among the rules are enabled and that unwanted ones are prevented. If no restrictions on rule applicability are imposed when the paraphrase generator is invoked for an MRS, all of the rules are applied exhaustively and iteratively until no further application of any rule is possible, resulting in an often large set of derived MRSs. Each of these can be presented as input to the English generator, which can realize one or all of the sentences licensed by the ERG for that MRS. Invocation of the paraphrase engine can include positive or negative requirements on which paraphrase rules to apply, for example requiring only coordination variants that preserve the original order of the conjuncts, or excluding variants with pronouns or ellipsis.

## 4 Evaluation and discussion

The initial benchmark for this implementation was to provide sufficient MRS-to-MRS paraphrase rules in order to successfully generate the full range of sentence variants observed in the LPL textbook, exhibiting rich combinations of the linguistic phenomena described above. The present rule set accomplishes this task for a set of 77 FOL expressions drawn from the book, including the following example along with a few of the English paraphrases generated by the system:

Larger(**a**, **c**)&Larger(**e**, **c**)&¬Large(**a**)&¬Large(**e**)

*a and e are both larger than c, and neither of them is large.*

*a is larger than c and e is larger than c, but neither of them is large.*

*e is larger than c and a is larger than c; moreover, e isn't large, and it's not the case that a is large.*

But missing from the current 4,448 distinct sentences generated by the system for this one FOL is the following, which should be expected from seeing the first two outputs above.

*a and e are both larger than c, but neither of them is large.*

This illustrates the need for further adjustments either to the partial order imposed on the existing rule set, or tuning of the input conditions for one or more of the rules, to enable the substitution of *but* for *and* where the first conjunct is a sentence with a conjoined subject, and the second conjunct contains a partitive subject noun phrase with a pronoun anchored to that conjoined subject.

A more interesting evaluation of this paraphrase generator would involve a live study with students in the LPL course, to see if either on-the-fly generation of new exercises or rephrasing in English of incorrect answers could be shown to correlate with improved learning outcomes. A carefully controlled study of this type would be non-trivial to design and to carry out, but the generator in its current state should provide the functionality required for the FOL-to-English component of such a study.

## 5 Related work and next steps

The study of generation of natural language from logic dates back at least to the rule-based method of Wang (1980), which focused on problems involving quantifiers, unlike the present study, but developed a similarly decomposed approach to the translation task. Another rule-based approach is the semantic-head-driven method of Shieber et al. (1990), which is concerned primarily with the generation algorithm itself, and with the treatment of quantification, defining grammar-specific rules both for semantics-to-semantics mappings, and for semantics to surface forms; again the issue of paraphrasing is not addressed beyond variations due to lexical choice. More recent work such as that of Lu & Ng (1990) explores the use of statistical methods to generate English sentences from expressions in typed lambda calculus, including both log-linear and generative models. Both of these latter approaches also included application to multiple languages, which should also be possible with the present method, given an MRS-based grammar of another language, but no work has as yet been done in this direction. Another more recent approach by Kutlak & van Deemter (2015) enables improved generation in English output by defining simplification rules at the logic-to-logic level, rather than at the grammar-specific semantics level, but again no attention is given to paraphrasing.

An approach more closely aligned to the present one is the relatively early work of de Roeck & Lowden (1986), which identifies the important and vexing issue of how to minimize ambiguity in the natural language output of an automated generation system. This remains a challenge for the present approach, as can be seen by the following paraphrase currently and unfortunately generated by the system from an FOL with negation:

$\neg \text{Large}(\mathbf{a}) \& \neg \text{Large}(\mathbf{b})$

*It is not the case that **a** is large and **b** isn't large.*

This sentence makes a locally reasonable substitution for the negation predicate in the first conjunct clause, but creates an unwanted ambiguity where the second conjunct could be interpreted as within the scope of *It is not the case that...* Additional

*Dan Flickinger*

unwanted ambiguity emerges at many other points in the mapping space encompassed by the rules defined for the present system, including confusable antecedents of pronouns and of elided verb phrases. Clearly such ambiguity needs to be minimized, and has been reduced to a large degree by careful manual adjustment to input conditions for the rules, but the development of a more systematic method of detecting and avoiding such unwanted ambiguity remains for future research.

## Acknowledgments

I am grateful to David Barker-Plummer and his student Aaron Kalb of the Openproof project at Stanford's Center for the Study of Language and Information for presenting me with this challenge, and for their constructive critique during the development of the paraphrase generator. I am also grateful to the project for funding which supported the work. On a grander scale, I am indebted to John Nerbonne for illuminating the rich complexity of the mapping between natural language and logic during the years that we worked together in the Natural Language Project at Hewlett-Packard Laboratories.

## References

- Barker-Plummer, David, Jon Barwise & John Etchemendy. 2011. *Language, proof and logic*. 2nd edn. Stanford, CA: CSLI Publications.
- Carroll, John, Ann Copestake, Daniel Flickinger & Victor Poznanski. 1999. An efficient chart generator for (semi-)lexicalist grammars. In *Proceedings of the 7th European Workshop on Natural Language Generation*, 86–95. Toulouse, France.
- Copestake, Ann. 2002. *Implementing typed feature structure grammars*. Stanford, CA: CSLI Publications.
- Copestake, Ann, Dan Flickinger, Carl J. Pollard & Ivan A. Sag. 2005. Minimal recursion semantics: an introduction. *Research on Language and Computation* 3(4).
- de Roeck, A. N. & B. G. T. Lowden. 1986. Generating english paraphrases from formal relational calculus expressions. *Proceedings of the 11th Conference on Computational Linguistics, COLING 1986*. 581–583.
- Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6 (1) (Special Issue on Efficient Processing with HPSG). Dan Flickinger, Stephan Oepen, J. Tsujii & Hans Uszkoreit (eds.). 15–28.
- Flickinger, Dan. 2011. Accuracy vs. robustness in grammar engineering. In Emily M. Bender & Jennifer E. Arnold (eds.), *Language from a cognitive perspective: grammar, usage, and processing*, 31–50. Stanford: CSLI Publications.
- Kutlak, Roman & Kees van Deemter. 2015. Generating succinct English text from FOL formulae. *Proceedings of the First Scottish Workshop on Data-to-Text Generation*.

- Lønning, Jan Tore, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, Victoria Rosén & Erik Velldal. 2004. LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*. Uppsala, Sweden.
- Lu, Wei & Hwee Tou Ng. 1990. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 16 (1). 305–42.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar* (Studies in Contemporary Linguistics). Chicago, IL & Stanford, CA: The University of Chicago Press & CSLI Publications.
- Shieber, Stuart, Gertjan van Noord, Fernando C. N. Pereira & Robert C. Moore. 1990. Semantic-head-driven generation. *Computational Linguistics* 16 (1). 305–42.
- Wang, Juen-tin. 1980. On computational sentence generation from logical form. *Proceedings of the 8th Conference on Computational Linguistics, COLING 1980*. 405–411.



## Chapter 12

# Use and possible improvement of UNESCO's *Atlas of the World's Languages in Danger*

Tjeerd de Graaf

Fryske Akademy

### 1 Introduction

In countries such as the Netherlands, Germany and New Zealand, it is obvious which minority languages have to be regarded as being endangered, i.e. Frisian, Sorbian and Maori, respectively. In other countries, the situation is far more complicated and insight into various factors is required in order to reach an understanding of the overall sociolinguistic situation of a language with respect to its degree of endangerment. The following paper summarizes UNESCO activities focusing on gathering information on the degree of language endangerment and its visualisation in UNESCO's *Atlas of the World's Languages in Danger*. This Atlas considers various stages of endangerment and provides a survey of the available data on these languages in separate areas of the world. It is intended to raise awareness about language endangerment and the need to safeguard the world's linguistic diversity among policy makers, speaker communities and the general public, and to be a tool to monitor the status of endangered languages and the trends in linguistic diversity at a local level. A computer version of the Atlas makes it possible to correct and add information on particular items based on new available data. This keeps the contents on minority languages up-to-date, improves the Atlas and provides a source for comparative research and the preparation of materials for teaching and other purposes.

## 2 Assessing language vitality and endangerment

In 1995 UNESCO launched a Clearing House for the Documentation of Endangered Languages in Tokyo. Since then many international meetings have taken place, either addressing the problem of language endangerment in general or discussing a geographic approach (Africa, South America, the Russian Federation, etc.). Within the framework of these activities an International Expert Meeting was organized by the UNESCO headquarters in Paris in March 2003. There, an UNESCO ad hoc Expert Group on Endangered Languages presented a draft report entitled *Language Vitality and Endangerment* (2003) for discussion among a wide audience of linguists, language planners, representatives of NGO's, as well as members of endangered language communities. At the meeting, a final document was produced and among the main outcomes the following nine core factors were identified with the help of which the language situation can be assessed:

### Degree of endangerment

1. Intergenerational language transmission
2. Absolute numbers of speakers
3. Proportion of speakers within the total population
4. Loss of existing language domains
5. Response to new domains and media
6. Materials for language education and literacy

### Language attitudes and policies

7. Governmental and institutional language attitudes and policies, including official language status and use
8. Community members' attitudes towards their own language

### Urgency of documentation

9. Amount and quality of documentation

Factors from 1 to 6 are applied to assess a language's vitality and its state of endangerment. The most crucial single factor among them is factor 1, which determines the extent of language acquisition among the children within a community. It is obvious that a language without any young speakers is seriously threatened by extinction.

The dynamics of the processes of a given language shift situation is captured by factors 1 to 5. The proportion of speakers within a community (factor 3) reveals an important aspect of language vitality: is the minority language still an essential

indicator for being regarded a member of the community or not? Can a person be a member of the community without speaking the heritage language?

The introduction of formal education or new job opportunities for the members of a minority group may result in the loss of domains in which the heritage language has been used up to then (factor 4). A shift in religious affiliation of a community might result in the shift to another mother tongue, a language that is associated with the new religion (factor 5).

Factor 6 relates to the stage of development of a given language. Is there a community's orthography? Have the community members agreed on a common standard form of writing? Are teaching and learning materials for the language available? Is there literature, such as newsletters, stories, religious texts, etc. published in that language? Factor 7 deals with the government's policies towards a language and factor 8 assesses the speakers' attitudes towards their ethnic language. Finally, factor 9 attempts to evaluate the urgency for documentation by focusing on the quantity and quality of already existing and analysed language data.

Speech communities are complex and patterns of language use within these communities are difficult to explore. The evaluation of the state of vitality of any language is therefore a challenging task. Members of an ethnolinguistic minority or external evaluators can use the factors introduced above in order to describe a language shift situation and to analyse the kind and state of endangerment of a language. The UNESCO ad hoc Expert Group has introduced for each factor a grading system from 0 to 5. With factor 1, for instance, grade 5 stands for the use of the language by all members of the community, whereas grade 0 states that there are no longer any speakers of this language left. In applying all the factors to the language situation, a table of numbers is obtained, which characterizes the kind and state of endangerment for a language. The information in such tables can serve as a useful instrument not only for the assessment of the situation of a community's language, but also for the formulation of appropriate support measures for language documentation, maintenance, or revitalization.

### **3 Versions of UNESCO's *Atlas of the World's Languages in Danger***

The first edition of the Atlas was edited by Stephen Wurm and published in 1996. It comprised 53 pages including 12 pages of maps showing some 600 languages. As a first publication of its kind, the Atlas met with vivid scholarly and journalistic interest and soon became a valuable reference book for the wider public.

A second, thoroughly updated edition of the Atlas was published by UNESCO in 2001, and expanded to 90 pages including 14 pages of maps showing some 800 languages. The update reflected the fact that since the first edition of the Atlas, research on endangered languages and scientific interest and work in the field has proliferated.

The latest print version of the Atlas was published in 2010. The Atlas lists some 2,500 endangered languages approaching the generally-accepted estimate of about

half of the more than 6,000 languages of the world. It provides analytic reports on each region and attracts much academic, media and public attention. Hundreds of press articles in different parts of the world refer to the Atlas and underline its impact as awareness-raising instrument regarding language endangerment.

Since 2009, the interactive version of the Atlas is available, which provides the following data for more than 2,500 languages: name, degree of endangerment (indicated by a coloured dot as marker), location on the map and geographic coordinates, country, number of speakers, relevant projects, sources, ISO language codes.

The UNESCO online Atlas<sup>1</sup> is an interactive digital resource that can be continually enriched with updated and more detailed information, accessible globally, free of charge, to anyone with a computer and an internet connection. The online version shows, at the click of the mouse on the marker, the exact latitude and longitude coordinates of a central point in the area where a language is spoken. It provides a wealth of other information and permits interactive contributions from the world's linguists, census takers and, most importantly, language communities.

## 4 Degrees of endangerment for the Atlas

On the basis of the assessment of language endangerment, the *UNESCO atlas of the world's endangered languages* distinguishes the following six degrees with regard to intergenerational transmission (with in brackets the colour of the dots on the Atlas):

**Safe:** The language is spoken by *all generations*. There is no sign of linguistic threat from any other language, and the intergenerational transmission of the language seems uninterrupted.

**Vulnerable (white):** Most but not all children or families of a particular community speak the language as their first language, but it may be restricted to specific social domains (such as at home where children interact with their parents and grandparents).

**Definitively endangered (yellow):** The language is no longer being learned as the mother tongue by children in the home. The youngest speakers are thus of the *parental generation*. At this stage, parents may still speak their language to their children, but their children do not typically respond in the language.

**Severely endangered (orange):** The language is *spoken* only by *grandparents and older generations*; while the parent generation may still *understand* the language, they typically do not speak it to their children.

**Critically endangered (red):** The youngest speakers are in the *great-grandparental generation*, and the language is not used for everyday interactions. These older people often *remember* only part of the language but *do not use* it, since there may not be anyone to speak with.

---

<sup>1</sup> Web site: [www.unesco.org/languages-atlas/](http://www.unesco.org/languages-atlas/).

**Extinct (black):** There is no one who can speak or remember the language. In the Atlas those languages are indicated which became extinct since 1950.

According to the present Atlas data, nearly half of the languages spoken in the world are endangered.

## 5 Number of speakers and census data

Important factors determining the vitality of a language are the *absolute number of speakers* (factor 2) and the *proportion of speakers within the total population* (factor 3). It is impossible to provide a valid interpretation of absolute numbers, but a small speech community is always at risk. A small population is much more vulnerable to decimation (e.g. by disease, warfare, or natural disaster) than a larger one. A small language group may also merge with a neighboring group, losing its own language and culture.

The number of speakers in relation to the total population of a group is a significant indicator of language vitality, where *group* may refer to the ethnic, religious, regional, or national group with which the speaker community identifies. The report of the UNESCO ad hoc expert group on Endangered Languages uses a scale from 0 (extinct: no speakers of the language) to 5 (all speakers of the ethnic group speak the language) to refer to degrees of endangerment.

This situation can be illustrated by the census data of a number of Siberian languages, which also shows the change in time and the ongoing loss of the language and culture of these cases. As an example we provide the data for the Nivkh language, spoken by an ethnic minority in the Far East of the Russian Federation.

The Russian census in 2010 contained questions about personal data, citizenship (for the Nivkh *Russian*), nationality (*Nivkh*), education and language use (*Nivkh* or *Russian*). Here one has to distinguish two meanings of *Russian* (*Rossiyskiy* 'citizen of the Russian Federation' or *Russkiy* 'belonging to the Russian nationality'). The data for 2010 show a total number of representatives of the ethnic group of 4652, mother tongue speakers of the Nivkh language 8,5% and of Russian 91,5%, whereas nearly 100% used Russian in daily life. In 1959 the population size was 3717 with 76,3% mother tongue Nivkh and 23,7% Russian. Similar results are found for the case of other Siberian languages such as Yukagir and Koryak. In his book on the *Languages of the Northern Peoples in the XXth Century* Vakhtin (2001) provides similar data for the period between 1926 and 1989 and from this the following conclusion can be drawn: in 1926 most representatives of these Siberian peoples were monolingual in their own language, whereas more and more Russian took over and at present most of them have become monolingual in Russian.

This situation is illustrative for many endangered minority languages and it will be important to show this clearly by adding the related diachronic data to the Atlas in order to improve its quality further.

## 6 Materials for language education and literacy

Education *in* the language is essential for language vitality. There are language communities that maintain strong oral traditions, and some do not wish their language to be written. In other communities, literacy in their language is a source of pride. In general, however, literacy is directly linked with social and economic development. There is an urgent need for books and materials on all topics for various ages and language abilities (factors 6 and 9).

The UNESCO ad hoc Expert Group distinguishes several grades for this factor of language vitality. The highest grade (5) is given when the language has an established orthography, literacy tradition with grammars, dictionaries, recorded texts, literature and everyday media. Writing in the language is used in administration and education. Lower grades are given for languages where part of these properties are lacking and at the lowest level grade 0 is provided for languages where no orthography is available in the community.

Several ethnic communities in the world get support from organisations which assist them in the development of their language and culture by providing materials for language learning and teaching. For example, the Foundation for Siberian Cultures<sup>2</sup> – founded in 2010 – has the aim to preserve the indigenous languages of the Russian Federation and the ecological knowledge expressed in them (Kasten & de Graaf 2013). During our fieldwork expeditions to Sakhalin, Kamchatka, Northern Yakutia and Central Siberia we studied processes of language shift and language death for some minority peoples of Russia, in particular for the Nivkh of Sakhalin, the Itelmen and Koryak of Kamchatka, and the Yukagir of Sakha/Yakutia.

A digital library and ethnographic collections on the World Wide Web provide above all indigenous communities with open access to relevant scholarly resources and research materials. Recent or current projects are presented at regular shows on the internet in the form of alternating photo-video shows. This provides a forum through which indigenous communities can participate and be informed about how their traditions are presented and received abroad.

In the past we received research grants which made it possible to re-record material from collections of historical sound carriers according to up-to-date technology and to store them in safe places together with the related metadata. The results of present day fieldwork and the reconstructed data from sound archives provide important information for the preparation of language descriptions, grammars, dictionaries and edited collections of oral and written literature. These can also be used to develop teaching methods, in particular for the younger members of certain ethnic groups who do not have sufficient knowledge of their native language (de Graaf 2012).

Information about this kind of projects can be added to the data of the Atlas by texts or links to the related websites. This feedback will further improve the quality of the Atlas.

---

<sup>2</sup> Web site: [www.kulturstiftung-sibirien.de/](http://www.kulturstiftung-sibirien.de/).

## 7 Interactive Atlas user feedback

UNESCO has commissioned the Foundation for Endangered Languages<sup>3</sup> to monitor and process the feedback from users of the online edition of the Atlas. In this way the Interactive Atlas is constantly improved and updated. The feedback is evaluated by the editorial board, and validated for updates, or addition of new content. Users are invited to submit comments through different channels, in particular directly on the internet.

Each language entry in the online Atlas contains a tab for comments on any of the following elements:

- correct or complete this record (names, vitality degree, location, ISO code, etc.) and provide online or bibliographic data;
- share information on media or online resources (such as dictionaries, websites) for this language;
- describe a recent or current safeguarding or revitalisation project for this language.

It is also possible for users to suggest a new language for inclusion in the Atlas. They can do this by filling out an online form. The suggestions from users can be broadly categorised as covering the following areas:

- location of the markers;
- status on the endangerment scale;
- population figures and speaker numbers
- classification as a language, is it a language or a dialect?
- Additional bibliographic sources, especially new learning materials;
- personal anecdotes about contact with the speakers;
- ethno-political policy statements from representatives of minorities;
- general questions about UNESCO criteria.

The Foundation for Endangered Languages has appointed a set of regional consultants who are familiar with the language situation in the region which is considered in a specific comment. They form the editorial board which on the basis of these comments updates and improves the Atlas content. The problems and controversies related to this procedure are described by Moseley (2012).

In March 2011, 116 language entries had been updated in the Interactive Atlas thanks to users' feedback. At present (August 2015) the editorial board is considering many new suggestions for improvement of the Atlas and in recent publications

---

<sup>3</sup> Web site: [www.ogmios.org/index.php](http://www.ogmios.org/index.php).

*Tjeerd de Graaf*

several possibilities for this have been mentioned, such as by Kornai (2015) and Soria (2015), stressing the importance of “Digital Language Diversity”, the response of the language to new media in the digital domain (factor 5).

## 8 Conclusions

During the UNESCO International Expert Meeting on Improving Access to Multilingual Cyberspace, which was held in Paris, 28–29 October 2014, discussions took place about the future of the Atlas, for which further improvements were suggested. One of the recommendations was to upscale UNESCO’s *Atlas of the World’s Languages in Danger* to a more general *UNESCO World Atlas of Languages*, which is not limited to endangered languages, but provides an overall view of linguistic diversity, multilingualism and language change in the world.

For the time being the work continues on the existing version of the *UNESCO World Atlas of Languages in Danger*, for which quite modest financial support is required. A new UNESCO Advisory Group should develop a comprehensible and sustainable set of new indicators, based on the existing proposals and advantages of ICT. Depending on possible future financial resources, these should be implemented and added to the Atlas in combination with the results of initiatives elsewhere. This could lead to a *UNESCO World Atlas of Languages* as a new monitoring tool, which not only determines the vitality and possible state of endangerment of the world’s languages, but also measures language diversity and multilingualism in specific regions of the world.

## Acknowledgement

The author of this paper thanks Christopher Moseley for making some related texts available and for giving further support.

## References

- de Graaf, Tjeerd. 2012. How oral archives benefit endangered languages. In *NETLANG. Towards the multilingual cyberspace*, 269. Caen: MAAYA Network, C&F Éditions.
- Kasten, E. & Tjeerd de Graaf (eds.). 2013. *Sustaining indigenous knowledge: learning tools and community initiatives for preserving endangered languages and local cultural heritage*. Fürstenberg/Havel: Kulturstiftung Sibirien. [http://www.siberian-studies.org/publications/sustainingik\\_E.html](http://www.siberian-studies.org/publications/sustainingik_E.html).
- Kornai, A. 2015. A new method of language vitality assessment. In *Proceedings of the 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, 132. Moscow: Interregional Library Cooperation Centre.
- Language Vitality and Endangerment*. 2003. Document adopted by the International Expert Meeting on UNESCO’s Programme Safeguarding of Endangered Languages. Paris: UNESCO. <http://unesdoc.unesco.org/images/0018/001836/183699E.pdf>.

- Moseley, Christopher (ed.). 2010. *Atlas of the world's languages in danger*. 3rd edn. Paris: UNESCO Publishing. <http://www.unesco.org/languages-atlas/en/atlasmap.html>.
- Moseley, Christopher. 2012. *The UNESCO atlas of the world's languages in danger; context and process*. Cambridge: World Oral Literature Project.
- Soria, C. 2015. Towards a notion of "digital language diversity". In *Proceedings of the 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, 111. Moscow: Interregional Library Cooperation Centre.
- UNESCO atlas of the world's languages in danger*. 2011. Brochure published by the United Nations Educational, Scientific and Cultural Organisation, with the support of the Government of Norway. Paris: UNESCO Publications Office.
- UNESCO atlas of the world's languages in danger*. First edition, (1996), Wurm, S. (ed.); Second edition, 2001, Wurm, S. (ed.); Third edition, 2010, Moseley, C. (ed.). Paris: UNESCO. [http://publishing.unesco.org/details.aspx?Code\\_Livre=4728](http://publishing.unesco.org/details.aspx?Code_Livre=4728).
- Vakhtin, N. 2001. *Yazyki narodov severa v XX veke [Languages of the peoples of the north in the XXth century]*. St. Petersburg: Evropeyskiy Universitet.



## Chapter 13

# Assessing smoothing parameters in dialectometry

Jack Grieve

Aston University

This paper considers an approach that was suggested by John Nerbonne for assessing how to best set parameters for the smoothing of dialect maps using statistical methods. This approach involves correlating the smoothed maps for a regional linguistic variable generated under various different settings of a parameter to the underlying raw map and then graphing the results in order to estimate a reasonable value for that parameter. In order to test this method, the relative frequencies of numerous words were mapped across the counties of the contiguous United States based on an 8.9 billion word corpus of geocoded Tweets. These relative frequency maps were then smoothed using a Getis-Ord  $G_i^*$  local spatial autocorrelation analysis based on a nearest neighbor spatial weights matrix, where the number of nearest neighbors was varied from between 1 and 200 locations. The analysis suggests that setting the value of this parameter at between 25 and 50 nearest neighbors, or alternatively at approximately 10% of the total locations over which the variable was measured, generally yields acceptable results.

## 1 Introduction

One of the longest standing methodological problems in dialectology is how to make sense out of the complex patterns of regional variation that are generally exhibited by linguistic variables when mapped. The traditional solution to this problem has been to draw isoglosses by hand that divide the map into regions where the different values of the variable are more or less common. An alternative solution is to use statistical methods to automatically smooth dialect maps, including a Getis-Ord  $G_i^*$  local spatial autocorrelation analysis (Getis & Ord 1992; Grieve, Speelman & Geeraerts 2011; Grieve 2016).

Basically, a Getis-Ord  $G_i^*$  local spatial autocorrelation analysis is a statistical method that identifies underlying patterns of spatial clustering in the values of a quantitative variable that has been measured across a set of locations. A Getis-Ord  $G_i^*$  analysis functions by comparing the values of a variable around each location

over which it is measured. If the values of the variable tend to be relatively high, then that central location is assigned a positive  $z$ -score, whereas if those values tend to be relatively low, then that location is assigned a negative  $z$ -score. These  $z$ -scores are then mapped to identify underlying regional clusters of high and low value locations, much like drawing an isogloss.

This type of statistical approach to the analysis of dialect maps has several advantages over drawing isoglosses by hand. Most important, it allows for consistent, efficient, and replicable analyses. This is especially useful when analyzing and comparing maps for many different linguistic variables, for example when the results of dialect surveys are used to identify common patterns of regional linguistic variation. In such studies, there is a very real possibility that dialectologist bias will substantially affect the results of the analysis. On any given map, most dialectologists will usually broadly agree on the placement of isoglosses, but when this procedure is repeated for many different variables, for example as a first step toward the identification of common patterns of regional linguistic variation, small variations in how isoglosses are drawn can become amplified, possibly leading to major differences when isoglosses are aggregated, which can reflect the preconceptions of the dialectologist about where important dialect boundaries lie.

The use of statistical methods for identifying underlying regional signals in dialect maps therefore greatly limits the influence of dialectologist bias, but it does not eliminate this bias entirely. As is generally the case with all but the simplest statistical methods, including most of the methods that are commonly applied in dialectometry, there are numerous parameters that must be set by the dialectologist. Most notably, an important step in conducting a Getis-Ord  $G_i^*$  analysis is to define a spatial weights matrix, which specifies the relationship between every pair of locations over which the variable is measured. Essentially, the spatial weights matrix defines what constitutes a nearby location. For example, in the most basic type of spatial weights matrix, two locations are assigned a weight of 1 if they are considered to be nearby to each other and a weight of 0 if they are not considered to be nearby to each other. Proximity can be defined in various ways, including by the number of nearest neighbors, where for each location the  $n$  nearest neighboring locations are assigned a weight of 1 and all other locations are assigned a weight of 0. How one chooses to set the number of nearest neighbors is an important decision that affects the smoothness of the resultant maps. Specifically, the smaller the number of nearest neighbors taken into consideration, the more similar the resultant map will be to the original map. Of course, the goal of applying a local spatial autocorrelation analysis in the first place is to smooth the map and so it is always necessary to set this parameter to a value larger than 1, but otherwise setting this value is an important and often challenging decision (Getis 2009), where dialectologist bias can enter into the analysis. Perhaps most important, if these parameters are set too liberally, over-smoothing can result, where the smoothed map no longer accurately reflects the underlying regional pattern visible in the map for the variable under analysis.

Fortunately, it is possible to record and scrutinize the effects of these decisions, which is impossible when isoglosses are drawn by hand. Furthermore, given that a

local spatial autocorrelation analysis produces quantitative results, it is also possible to compare the smoothed maps to each other and to the original raw map in order to assess the degree of smoothing, and in particular to consider whether the maps have been over-smoothed. This paper considers one such approach to assessing smoothing parameters, which was suggested to the author by John Nerbonne at the 2014 Methods in Dialectology conference that he hosted in Groningen. Specifically, the suggestion was to assess the degree of smoothing of dialect maps by measuring the correlation between the raw map and the smoothed maps generated using different parameter settings.

## 2 Analysis

The corpus analyzed in this study consists of 8.9 billion words of geo-coded American mobile Twitter data, totaling 980 million tweets written by 7 million users from across the contiguous United States, downloaded between October 11th, 2013 and November 22nd, 2014 using the Twitter API (see Huang et al. 2016; Grieve, Nini & Guo 2016). To analyze patterns of regional linguistic variation in this variety of language, the corpus was geographically stratified by county using the longitude and latitude provided with each Tweet. In total the corpus contains 3,075 county equivalents out of a total of 3,108 county equivalents in the contiguous United States. On average, the corpus contains 2 million words per county, but the number of words per county ranges from 300 to 300 million words. Overall, 98% of the counties are represented by at least 10,000 words and 79% of the counties are represented by at least 100,000 words. Twitter provides a uniquely large and accessible source of geo-coded natural language data, which is also a highly informal variety of language that is participated in by millions of people from across the United States, making it a valuable source of data for dialectologists.

To test the effect of varying the number of nearest neighbors used to define the spatial weights matrix for a Getis-Ord  $G_i^*$  analysis of dialect maps, the relative frequencies of a series of words were measured across the counties in the corpus. This is not the type of linguistic variable commonly analyzed in dialectology, where lexical variation tends to be measured as alternations between equivalent forms (e.g. *pail* vs. *bucket*, *pop* vs. *soda* vs. *coke*); nevertheless, word frequencies still do generally show regional patterns and are therefore as suitable as any other type of linguistic variable for testing the effect of varying smoothing parameter settings. For example, the relative frequency map for the word *love*—the most common content word in the corpus—is presented in the first cell of Figure 1, showing that the usage of this word is relatively more common in the Upper South. Similarly, the relative frequency map for the word *know*—the second most common content word in the corpus—is presented in the first cell of Figure 2, showing that the usage of this word is relatively more common in the Deep South.

Next, smoothed maps for each of these words were generated using a Getis-Ord  $G_i^*$  analysis based on a series of 200 different nearest neighbors spatial weights matrices defined for between 1 and 200 nearest neighbors. As discussed above, each of these

Jack Grieve

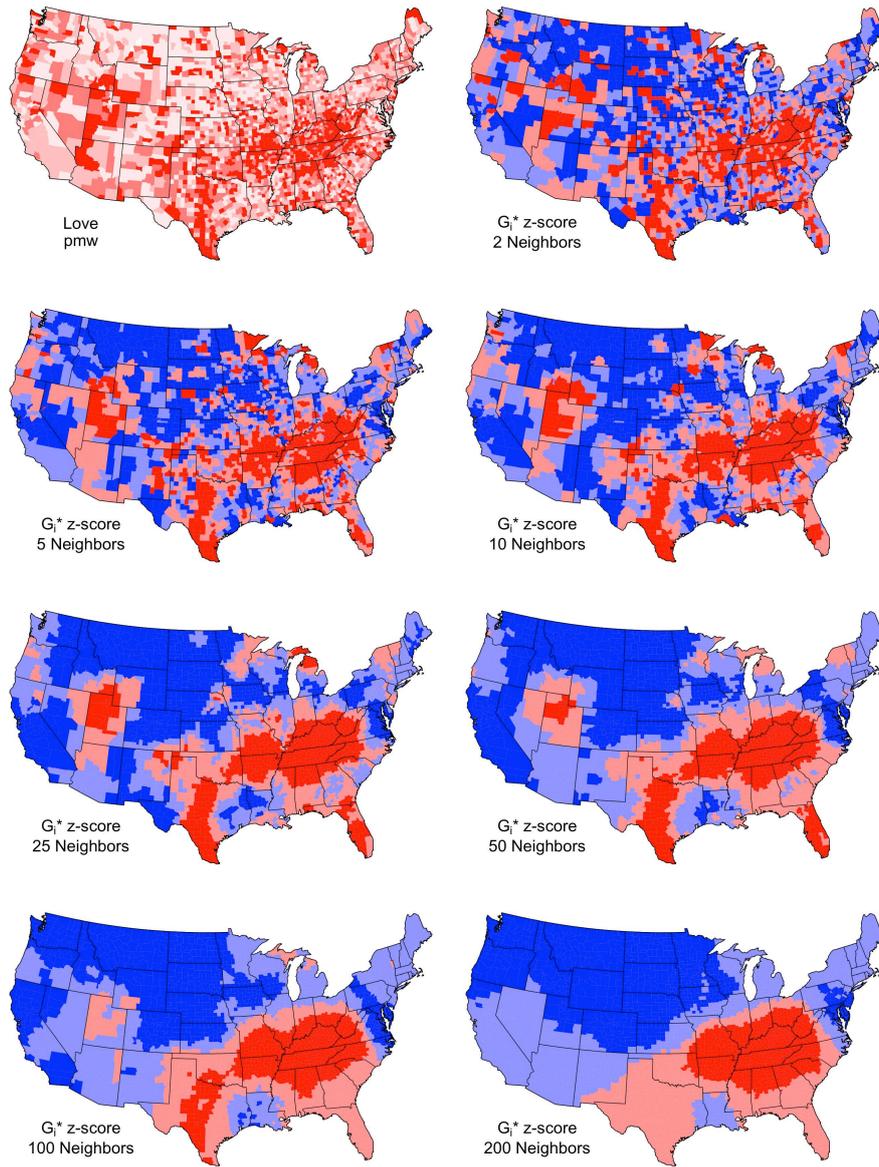


Figure 1: Relative frequency and local spatial autocorrelation maps for *love*.

13 Assessing smoothing parameters in dialectometry

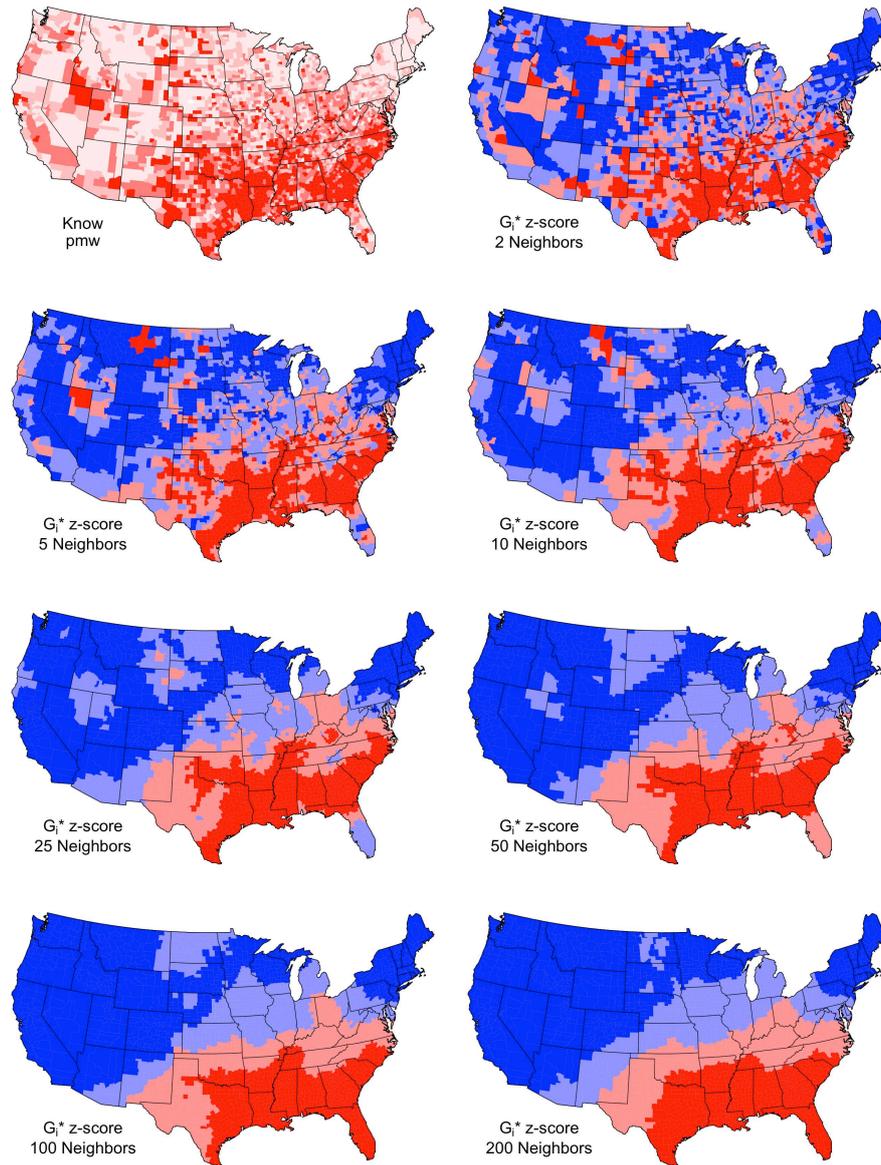


Figure 2: Relative frequency and local spatial autocorrelation maps for *know*.

*Jack Grieve*

analyses generates a  $z$ -score for each location over which the variable is measured, which can then be mapped to visualize the patterns of spatial clustering identified by the analysis. For example, the remaining cells in Figure 1 show the smoothed maps for *love* generated based on 2, 5, 10, 25, 50, 100, and 200 nearest neighbors spatial weights matrices, while the remaining cells in Figure 2 show the smoothed maps for *know* for these same parameter settings. When the number of nearest neighbors is set very low, the smoothed maps basically reproduce the raw data, whereas when the number of nearest neighbors is set very high, there is clearly over-smoothing present. For example, looking at the final smoothed map in the series for *love* (i.e. for 200 nearest neighbors), all of Utah is incorrectly identified as being a relatively low value region. Similarly, looking at the final smoothed map in the series for *love*, all of Florida is incorrectly identified as being a relatively high value region. It is therefore necessary to set the number of nearest neighbors somewhere between these two extremes in order to produce usefully smoothed maps that still accurately reflect the underlying patterns present in the raw data. In particular, looking at both sets of local spatial autocorrelation maps reproduced in Figures 1 and 2, it would appear that local spatial autocorrelation analysis based on between 25 and 50 nearest neighbors is ideal as values in this range strike a balance between over- and under-smoothing.

To assess how similar the local spatial autocorrelation maps are to the raw relative frequency maps upon which they are based, each local spatial autocorrelation map (i.e. the Getis-Ord  $G_i^*$   $z$ -scores measured over the 3,075 counties) were correlated to the corresponding raw maps (i.e. the relative frequencies measured over the 3,075 counties), following the suggestion made by John Nerbonne. The resulting Pearson correlation coefficients were then plotted against the number of nearest neighbors. The resulting graph for *love* is presented in the first cell of Figure 3 and the resulting graph for *know* is presented in the second cell of Figure 3. As one would expect, both graphs show that as the number of nearest neighbors increases the correlation between the raw map and the smoothed maps decreases, although overall the strength of the correlation remains substantial. The decrease, however, is not linear. Rather, the decrease starts off very steep and then gradually flattens. Furthermore, there notably appears to be an inflection point in both graphs between 25 and 50 nearest neighbors, which corresponds to the impressionistic analysis of the smoothed maps described above, where it was argued that conducting the local spatial autocorrelation analysis on spatial weights matrix based on between 25 and 50 nearest neighbors was best.

The other cells in Figure 3 present the results of the same analysis repeated for several other words, which were selected to represent a range of word frequencies and degrees of spatial clustering. Remarkably, all these graphs show very similar patterns, with inflections points falling in the same range, i.e. between 25 and 50 nearest neighbors, suggesting that this value represents a consistently applicable parameter setting for the smoothing of maps based on this dataset. This conclusion is supported by an analysis of the maps for these variables (not shown), which exhibit similar results to the smoothed maps presented for *love* and *know* in Figures 1 and 2—clearly exhibiting over-smoothing at higher parameter settings. Given that there are 3,075

### 13 Assessing smoothing parameters in dialectometry

locations in this dataset, it would therefore appear that using a number equal to approximately 10% of the total locations to set the spatial weights matrix is in general a reasonable value for this parameter.

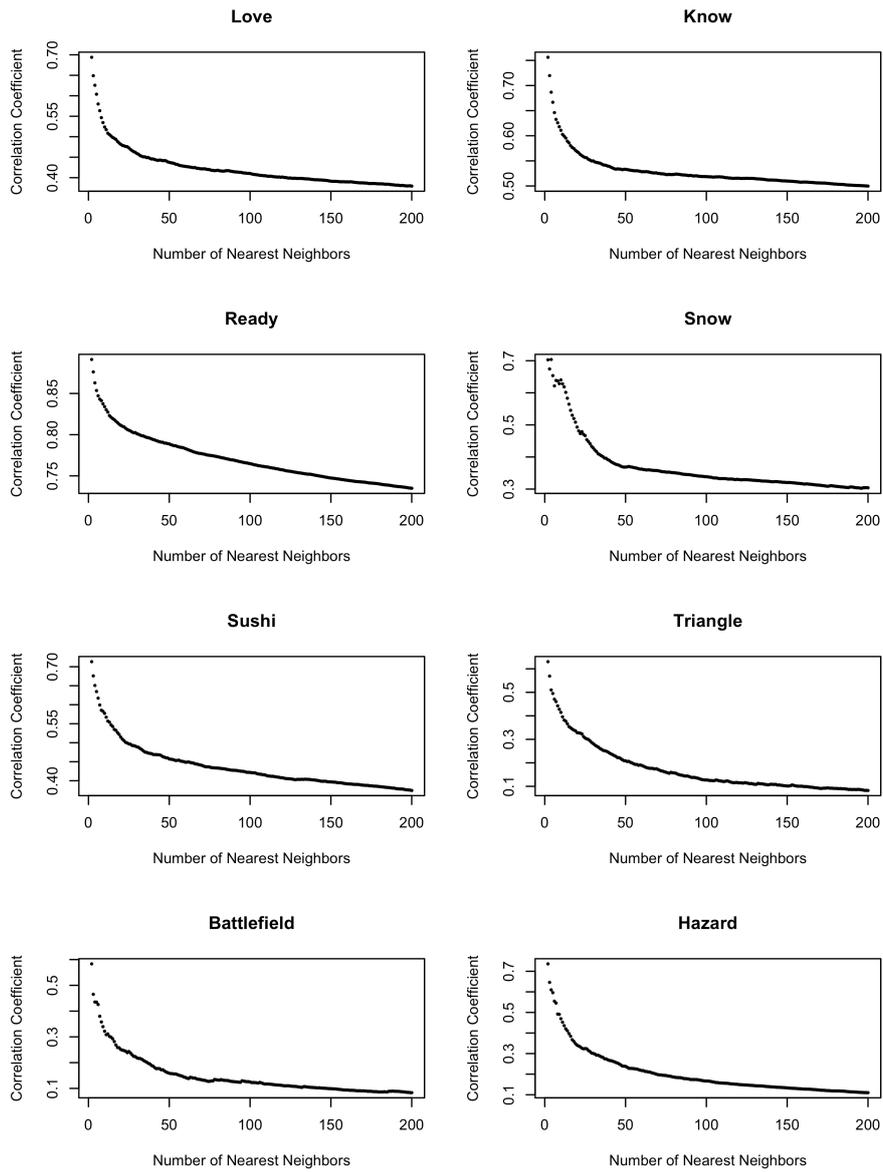


Figure 3: Nearest neighbor vs. correlation coefficient graphs.

### 3 Conclusion

This paper has briefly explored a general method for setting smoothing parameters for the analysis of individual patterns of regional linguistic variation in dialect maps, which was conceived by John Nerbonne. This method involves correlating maps smoothed using different parameter settings to the underlying raw maps and graphing these results. By inspecting the resultant graph, an approximate point of inflection is estimated and the smoothing parameter under consideration is then set to this value. In particular, this method was used in this paper to assess the number of nearest neighbors used to generate a nearest neighbor spatial weights matrix, an important step in conducting a Getis-Ord  $G_i^*$  local spatial autocorrelation analysis, which is an increasingly common method for smoothing dialect maps in dialectometry. Based on this approach, this study found evidence suggesting that using a number of nearest neighbors equal to approximately 10% of the total number of locations under analysis is a reasonable way to set this parameter for a Getis-Ord  $G_i^*$  analysis—generating usefully smoothed maps, while guarding against over-smoothing. Considerably more analysis both within this dataset and across other dialect datasets, however, is necessary to fully support this claim. In addition, it is important to test the applicability of this approach for setting the parameters associated with other types of spatial weights matrices as well as to test the applicability of this approach more generally for setting the parameters associated with other methods for smoothing used in dialectometry. Nevertheless, this relatively simple method appears to be a promising approach for helping to resolve an important modern methodological problem in dialectometry.

### References

- Getis, Arthur. 2009. Spatial weights matrices. *Geographical Analysis* 41. 404–410.
- Getis, Arthur & J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24. 189–206.
- Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge University Press.
- Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in American English online. *English Language and Linguistics*. To appear.
- Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23. 193–221.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*. To appear.

## Chapter 14

# Finding dialect areas by means of bootstrap clustering

Wilbert Heeringa

Fryske Akademy

In dialectometry cluster analysis is a means to find groups given a set of local dialects and their mutual linguistic distances. The weakness of cluster analysis is its instability; small differences in the distance matrix may strongly change the results.

Kleiweg, Nerbonne & Bosveld (2004) introduced composite cluster maps, which are obtained by collecting chances that pairs of neighboring elements are part of different clusters as indicated by the darkness of the border that is drawn between those two locations. Noise is added to the clustering process, which enables the authors to estimate about how fixed a border is. Nerbonne et al. (2008) use clustering with noise and bootstrap clustering to overcome instability. Both the work of Kleiweg, Nerbonne & Bosveld (2004) and Nerbonne et al. (2008) focus on boundaries which may be weaker or stronger.

We introduce a new flavor of bootstrap clustering which generates areas, similar to classical dialect maps. We perform a procedure consisting of four steps. First, we randomly select 1,000 times  $n$  items from  $n$  items with replacement. For each resampled set of items we calculate the aggregated distances. Second, on the basis of the distances we perform agglomerative hierarchical cluster analysis. We choose nearest neighbor clustering since this method reflects the idea of dialect areas as continua. On the basis of the tree we determine the number of natural groups by means of the elbow method. Third, for each pair of dialects we count the number of times that both dialects are found in the same natural group. Fourth, when two dialects belong to the same group in more than 95% of the cases, we mark them as 'connected'. In this way we will obtain networks which are the groups.

We apply the procedure to distances in the sound components measured with Levenshtein distance between a set of 86 Dutch dialects. We use material which was collected in the period 2008–2011.

## 1 Introduction

Jain & Dubes (1988: 55) define cluster analysis as “the process of classifying objects into subsets that have meaning in the context of a particular problem”. The goal of

clustering is to identify the main groups in complex data. In dialectometry cluster analysis is a means to find groups given a set of local dialects and their mutual linguistic distances. Goebel (1982) introduced cluster analysis in the field of dialectometry (see also Goebel 1984 and Goebel 1993).

The weakness of cluster analysis is its instability; small differences in the distance matrix may strongly change the results (Jain, Murty & Flynn 1999; Nerbonne et al. 2008). Kleiweg, Nerbonne & Bosveld (2004) introduced *composite cluster maps*, which are obtained by collecting chances that pairs of neighboring elements are part of different clusters as indicated by the darkness of the border that is drawn between those two locations. Noise is added to the clustering process, which enables the authors to estimate how fixed a border is. Given a distance matrix, a random value between 0 and a maximum is added to each distance, and subsequently the dialects are clustered. The maximum may be one or two standard deviations. This is repeated, e.g. 1,000 times, giving 1,000 clusterings, and the number of times that pairs of neighboring elements are part of different clusters in those 1,000 clusterings is counted. The results can be visualized in a map, where the darkness of the border between two locations represents the chance that the locations belong to different clusters.

In addition to *noise clustering* Nerbonne et al. (2008) also introduced *bootstrap clustering* to overcome instability. Given, e.g. a data set with transcriptions of 100 words for each local dialect, 100 words are randomly selected using replacement, Levenshtein distances are calculated between the dialects, and the dialects are clustered on the basis of the Levenshtein distances. When this is repeated, e.g. 1,000 times, the number of times that pairs of neighboring elements are part of different clusters is counted.

Nerbonne et al. (2008) show that noise clustering and bootstrap clustering produce similar results, but bootstrap clustering has the advantage that no noise ceiling needs to be specified.

Both the work of Kleiweg, Nerbonne & Bosveld (2004) and Nerbonne et al. (2008) focus on boundaries which may be weaker or stronger, i.e., they are gradual. This makes it harder to compare the maps with traditional dialect maps where the color distinctions give a visual representation of the borders between different dialect areas, for example, the map of Winkel (1901) and the map of Daan & Blok (1969).

We introduce a new flavor of bootstrap clustering which generates areas, similar to classical dialect maps. In our approach 1) we consider dialect groups as continua, i.e. each local dialect is not necessarily strongly related to any other local dialect in the same group; the local dialects in a group rather constitute a 'network' and 2) we take into account that not every local dialect can be classified with statistical confidence.

We apply the procedure to distances in the sound components measured with Levenshtein distance between a set of 86 Dutch dialects. Recorded transcriptions of older and younger speakers are used. Thus we are able to show the change of Dutch dialect areas in apparent time.

## 2 Data

In this paper we use a corpus database of recordings of 86 local Dutch dialects. This database was compiled in the period 2008–2011 by Heeringa & Hinskens (2014). The dialects are evenly spread over the Dutch and Frisian language areas and represent the major dialect regions.

In order to be able to measure dialect change in apparent time, at least two male speakers aged 60 or older, and two or more female speakers aged between 20 and 40 were recorded in each of the 86 locations. The males represent the older phase of a particular variety and the females the newer phase.

An scene of the Charlie Chaplin movie *The Kid* served as the basis of the recordings that were made. The scene can be regarded as a cross-section of plain, simple daily spoken language, and consists of 23 sentences, each containing an average of 7.6 words. We used a selection of 13 sentences for this study, which include a maximum of 125 words in the written standard Dutch version of the text.

Both the older male and the younger female speakers operated in small groups of at least two people. When a small group was being recorded, the individuals were first asked to write down a translation of the text in their own dialect, independently of each other. Then, together they compiled and wrote a consensus text upon which both of them agreed. Finally, they both read the consensus text aloud.

Phonetic transcriptions of the recordings were made. Usually, two recordings of the consensus dialect version of the story were produced by both the older males and the younger females. Since phonetic transcription is time-consuming, only one recording per group was transcribed, where the recording of the speaker who was the most autochthonous, had the clearest voice, and read the text most fluently was preferred. The transcriptions were made in IPA and digitized in X-SAMPA.

The recordings in our data set were transcribed by one transcriber. To ensure optimal consistency per item, transcriptions are made per sentence instead of per text. The same sentence was played (2 times 86 is) 172 times and transcribed. Subsequently, the next sentence was played 172 times and transcribed, etc.

For more details see Heeringa & Hinskens (2014).

## 3 Methodology

We perform a procedure consisting of four steps. First, we randomly select  $n$  items from  $n$  items with replacement 1,000 times. For each resampled set of items we calculate the aggregated distances. Second, on the basis of the distances we perform agglomerative hierarchical cluster analysis. Third, for each pair of dialects we count the number of times that both dialects are found in the same natural group. Fourth, when two dialects belong to the same group in more than 950 of the cases (95%), we mark them as ‘connected’. In this way we will obtain networks which are the groups. Below we will discuss each step in more detail.

### 3.1 Calculating distances and resampling

Distances in the sound components between dialects are measured with the aid of the Levenshtein distance metric (Levenshtein 1966). This algorithm was introduced into dialectology by Kessler (1995). The Levenshtein distance between two strings is calculated as the “cost” of the total set of insertions, deletions and substitutions needed to transform one string into another (Kruskal & Liberman 1999).

The aggregated distance between the two dialects is based on 125 word pairs (fewer if words were missing). We use normalized distance measures, calculating the aggregated distance between two dialects as the sum of a maximum of 125 word pair distances divided by the sum of the alignment lengths that correspond to the word pairs. For more details about the measurements see Heeringa & Hinskens (2014).

Now we calculate 1,000 times aggregated distances between the 86 local dialects on the basis of 125 words randomly chosen from 125 words, using either the transcriptions of the older or younger speakers.

### 3.2 Nearest neighbor clustering

Once we have obtained 1,000 distance matrices, for each distance matrix we apply agglomerative hierarchical cluster analysis. Each observation starts in its own cluster. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. We use *single-linkage*, which is also known as *nearest neighbor clustering*. In this kind of clustering at each step, the two clusters separated by the shortest distance are combined. The result is a binary hierarchical tree structure in which the dialect varieties are the leaves, and the branches represent the distances among (clusters or groups of) dialect varieties (Jain & Dubes 1988).

We choose nearest neighbor clustering since this method reflects the idea of dialect areas being continua, where the distance between geographically neighboring dialects is small, and the difference between geographically distant points may be large. This method also agrees with the little arrow method which the map of Daan & Blok (1969) is based on. Using the little arrow method, locations which have similar dialects according to the speakers are connected by arrows in the map. It then may happen that local dialects A and B are judged as similar, and dialects B and C are judged as similar, but A and C are judged as different by the speakers.

### 3.3 Count number of times that two dialects share the same group

On the basis of the tree we determine the number of natural groups. Dendrograms are binarily branching trees. Within a dendrogram different levels of detail can be distinguished. Starting at the root, a division into two groups is found. Then, if we delve a little deeper we find that one of the two groups is divided into two further groups. At the bottom of the tree are the leaves, and here we find a classification into the maximum number of groups, in our case 86, with each grouping containing a single variety. We thus have 85 levels, the first suggesting a division into two groups and the 85th suggesting a division into 86. For each division in  $i$  groups ( $2 \leq i \leq 85$ ),

we compute the variance in the original distances, as explained by the cophenetic distances of the part of the tree that gives a division in  $i$  groups. Cophenetic distances are distances between the dialects as reflected by the dendrogram. The cophenetic distance between two local dialects is the height of the dendrogram where the two branches that include the two objects merge into a single branch<sup>1</sup> (Sokal & Rohlf 1962).

In a graph, the variances are plotted against the number of groups as found in each of the 85 divisions. The initial clusters usually explain a great deal of the variance. However, at a certain point the marginal gain will drop, yielding an angle in the graph. This angle provides the number of natural clusters and is known as the *elbow* (Aldenderfer & Blashfield 1984). After the angle, the amount of explained variance in the distances increases much more slowly than before. An example of an elbow plot is shown in Figure 1.

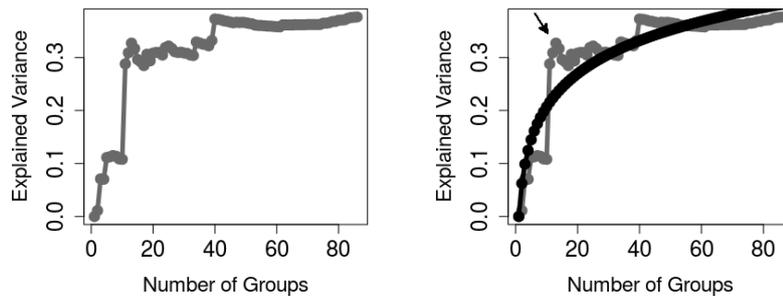


Figure 1: Left: elbow plot in which the variances are plotted against the number of groups. Right: a logarithmic regression curve is added. The elbow is found where the difference between predicted variance and ‘real’ variance is largest. The elbow is indicated by a little arrow. The number of natural groups is therefore 13.

We perform a linear regression analysis where the logarithmic number of clusters is the predictor and the explained variance the dependent variable. The curve which represents the explained variance predicted by the logarithmic number of clusters is added in the right graph in Figure 1. The elbow is found where the difference between the variance predicted by the logarithmic number of cluster and the ‘real’ variance is largest. The elbow is indicated by a little arrow and corresponds with 13 groups. The number of natural groups is therefore 13.

Now for each pair of dialects we count the number of times that both dialects are found in the same natural group. The number will vary between 0 (never) and 1,000

<sup>1</sup> Definition taken from *Wikipedia*, retrieved August 27, 2016, from <https://en.wikipedia.org/wiki/Cophenetic>.

*Wilbert Heeringa*

(always). The counts are graphically shown in Figure 2.

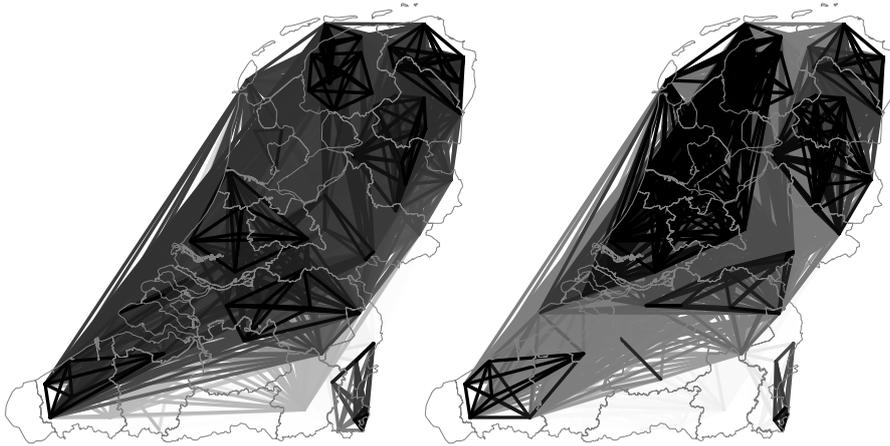


Figure 2: For each pair of dialects the number of times that both dialects are found in the same natural group is counted. The maps show counts obtained on the basis of transcriptions of older males (left) and younger females (right). Higher counts are represented by darker lines.

### 3.4 Create networks

When two dialects belong to the same group in more than 950 of the 1,000 bootstrap runs (95%), we mark them as being ‘connected’. Pairs of dialects which belong to the same group in more than 95% of the cases are connected by a line in the map (see Figure 3).

In this way we will obtain networks. For example, when dialects A and B are connected, and dialects C and D are connected, but dialects A and B are not connected with dialects C and D, we obtain two separated networks, each of which consists of two dialects. We consider each network as a group. Dialects which are a part of the same network, belong to the same group. Dialects which are not connected to any other dialect remain unclassified. Our procedure does not force a local dialect to belong to a group when it is not strongly related to any other local dialect.

## 4 Results and conclusions

The final classifications are shown in Figure 4. The number of groups has decreased from 12 to 10. The mixed Frisian group (orange on the ‘old’ map) has been absorbed by the group of Holland dialects (yellow), and the petrol group has expanded and absorbed the dialects of Kampen and IJsselmuiden (lighter green on the ‘old’ map). The

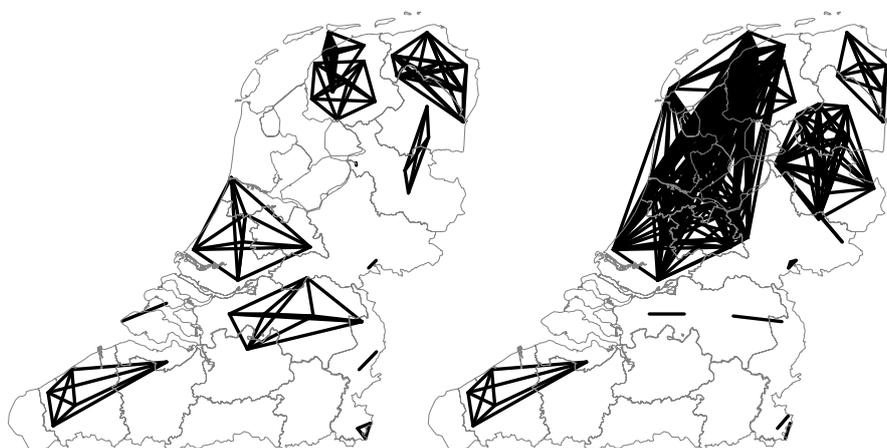


Figure 3: Local dialects that are connected by a line share in 95% of 1,000 bootstrap runs the same groups. Thus we obtain networks which are considered as dialect groups.

group of two Zeeland varieties (green/brown) has disappeared and the pink group has split into two smaller groups. The northern Limburg group (red/brown) has disappeared, and the group in the Southeast (red dots on the 'old' map) has split into two groups. On the 'new' map the red/brown dots represent transitional Limburg dialects, and the red dots Ripuarian dialects.

The number of unclassified dialects is larger for the older males than for the younger females: 31 versus 24, but the difference is not significant.

We conclude that the new cluster procedure produces clear area maps which are statistically justified and which can be easily compared to older dialect maps. Additionally changes of dialect areas can be clearly visualized by this approach.

## References

- Aldenderfer, M. & R. Blashfield. 1984. Cluster analysis. In *Sage university paper series on quantitative applications in the social sciences 07-044*. Newbury Park (CA): SAGE publications.
- Daan, J. & D. P. Blok. 1969. *Van Randstad tot Landrand; toelichting bij de kaart: dialecten en naamkunde*. Vol. XXXVII (Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam). Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.
- Goebel, H. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vol. 157 (Philosophisch-Historische

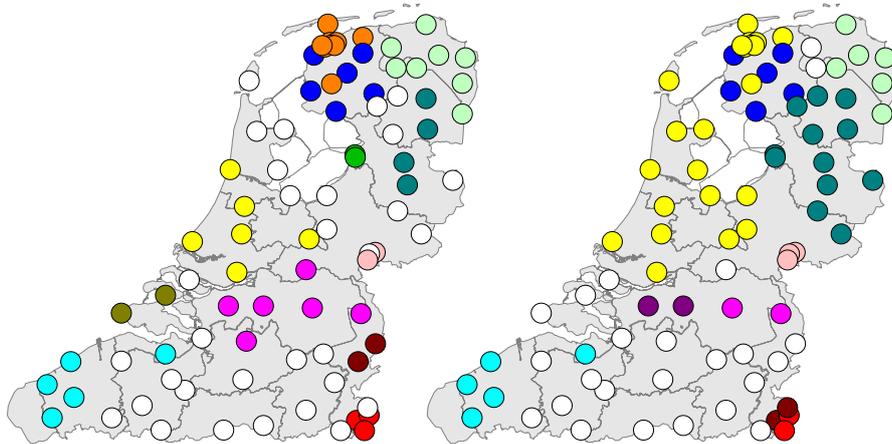


Figure 4: For the older males 12 groups are found (left) and for the younger females 10 groups are found (right).

- Klasse Denkschriften). With assistance of W.-D. Rase and H. Pudlatz. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.
- Goebl, H. 1984. *Dialektometrische Studien. Anhand italo-romanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Vol. 191, 192, 193 (Beihefte zur Zeitschrift für romanische Philologie). With assistance of S. Selberherr, W.-D. Rase and H. Pudlatz. Tübingen: Max Niemeyer Verlag.
- Goebl, H. 1993. Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In W. Viereck (ed.), *Proceedings of the International Congress of Dialectologists*, vol. 1, 37–81. Stuttgart: Franz Steiner Verlag.
- Heeringa, Wilbert & F. Hinskens. 2014. Convergence between dialect varieties and dialect groups in the Dutch language area. In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating dialectology, typology, and register analysis; linguistic variation in text and speech* (Linguae et Litterae: Publications of the School of Language and Literature), 26–52. Berlin & Boston: De Gruyter.
- Jain, A. K. & R. C. Dubes. 1988. *Algorithms for clustering data*. Englewood Cliffs, New Jersey: Prentice Hall.
- Jain, A. K., M. N. Murty & P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* 31(3). 264–323.
- Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60–67. Dublin: EACL.
- Kleiweg, P., J. Nerbonne & L. Bosveld. 2004. Geographic projection of cluster composites. In A. Blackwell, K. Marriott & A. Shimojima (eds.), *International Conference on Theory and Application of Diagrams*, 392–394. Springer.

14 *Finding dialect areas by means of bootstrap clustering*

- Kruskal, J. B. & M. Liberman. 1999. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff & J. Kruskal (eds.), *Time warps, string edits, and macromolecules. The theory and practice of sequence comparison*, 2nd edn., 125–161. 1st edition appeared in 1983. Stanford: CSLI.
- Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10(8). 707–710.
- Nerbonne, J., P. Kleiweg, Wilbert Heeringa & F. Manni. 2008. Projecting dialect distances to geography: bootstrap clustering vs. noisy clustering. In C. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data analysis, machine learning and applications*, 647–654. Springer.
- Sokal, R. R. & F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon* 11. 33–40.
- Winkel, J. te. 1901. *Geschiedenis der Nederlandsche taal*. Naar de tweede Hoogduitsche uitgave met toestemming van den schrijver vertaald door Dr. F. C. Wieder. Met eene Kaart. Culemborg: Blom & Olivierse.



## Chapter 15

# An acoustic analysis of English vowels produced by speakers of seven different native-language backgrounds<sup>1</sup>

Vincent J. van Heuven

University of Pannonia; University of Groningen; University of Leiden; Fryske Akademy

Charlotte S. Gooskens

University of Groningen; University of New England

We measured F1, F2 and duration of ten English monophthongs produced by American native speakers and by Danish, Norwegian, Swedish, Dutch, Hungarian and Chinese L2 speakers. We hypothesized that (i) L2 speakers would approximate the English vowels more closely as the phonological distance between the L2 and English is smaller, and (ii) English vowels of L2 speaker groups will be more similar as the L2s are closer to one another. Comparison of acoustic vowel diagrams and Linear Discriminant Analyses (LDA) confirm the hypotheses, with one exception: Dutch speakers deviate more from L1 English than the Scandinavian groups. The Interlanguage Speech Intelligibility Benefit was convincingly simulated by the LDA.

## 1 Introduction

In the past century English has evolved into the *Lingua Franca* of the world. It is now the language of commerce, international relationships and science *par excellence* (e.g. Rogerson-Revell 2007). The use of spoken English as a Lingua Franca (ELF) is not without problems, however. When a person learns to speak a foreign language, the pronunciation of the target language will differ from that of native speakers of

---

<sup>1</sup> This work was supported by The Netherlands Organisation for Research (NWO grant number 360-70-430) and by the European Union (TÁMOP-4.2.1.D-15/1/KONV-2015-0006).

that language and will be reminiscent of the sound patterns of the learner's mother tongue (e.g. Flege 1995). It is often easy to recognize the native language background of an ELF speaker by his non-native accent.

The present study compares the pronunciation of the (monophthongal) vowels of English produced by American L1 speakers with the same vowels pronounced by speakers of English from six different non-native backgrounds, i.e. Chinese, Dutch, Hungarian, Danish, Norwegian and Swedish. Chinese and Hungarian do not belong to the Indo-European language family. The other four languages are rather closely related to each other and to English. They are all members of the Indo-European Germanic language family group, be it in different branches. English is in the Anglo-Frisian branch, whereas Dutch is a West-Germanic language. The three Scandinavian languages are more closely related to each other (as members of the North-Germanic branch) than they are to either Dutch or English (Hendriksen & van der Auwera 1994). We will test the hypothesis that native languages that resemble one another yield foreign accents that are also similar to each other. Specifically, we expect the Danish, Norwegian and Swedish foreign accents in English to be so similar that one might speak of a Scandinavian accent.

Swarte (2016) measured linguistic distances between English, Dutch, German, Danish and Swedish. The two Scandinavian languages {Danish, Swedish} had the shortest distance between them, followed by the pair {Dutch, German}. English was more closely associated with the West-Germanic pair than with the Scandinavian pair. These linguistic distances were matched by experimentally established intelligibility results obtained in spoken-word translation and comprehension tasks. We predict that the Englishes spoken with a Scandinavian accent will be more alike than each of these accents is like the Dutch accent. Moreover Swarte's (2016) results lead us to expect that Dutch-accented English is closer to native English than the Scandinavian accents are. Finally, we predict that the Chinese and Hungarian accents differ more strongly from native English than the Germanic accents but we have no way of predicting which one will be closer to English.

We decided not to recruit human listeners but to simulate human listening through computer modeling. Linear Discriminant Analysis (LDA) has been used since the late 1990s to model and predict how human listeners with language background A identify the sounds of a foreign language B (e.g. Strange et al. 2005).

Intelligibility is greatest when interactants (speaker and listener) both use the same native language (e.g. Munro & Derwing 1995). When one interactant is a non-native, communication generally suffers (Cutler 2012). Communication in ELF may be as successful or even more successful when both interactants are non-native. This effect has been called the *interlanguage speech intelligibility benefit* (ISIB, Bent & Bradlow 2003). A realistic version of the ISIB hypothesis holds that the benefit will be found especially if speaker and listener have the same native language. If the LDA technique can be used to model and predict perceptual assimilation of non-native sounds to a native sound inventory (Strange et al. 2005), then the ISIB effect should also show up when we use the LDA technique.

## 2 Vowel systems of the seven languages

The languages targeted in this study have rich vowel inventories, as shown in Figure 1.

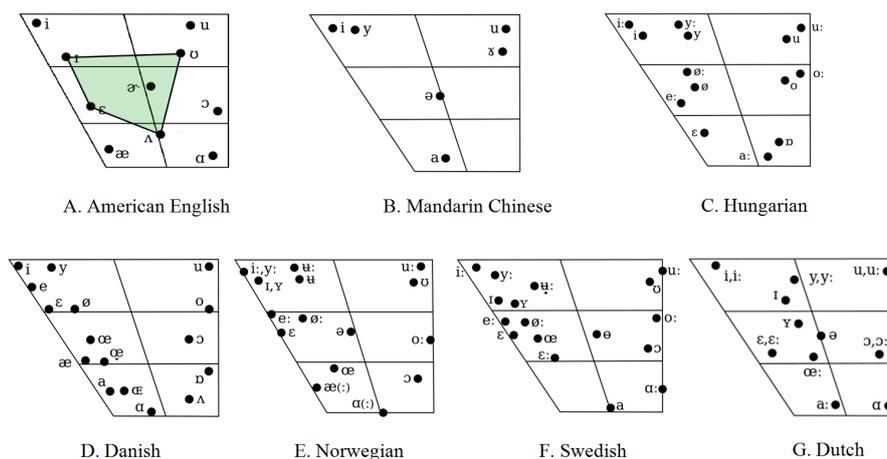


Figure 1: The monophthongs of (A) American English (Mannell, Cox & Harrington 2009), (B) Mandarin Chinese (Lee & Zee 2003), (C) Hungarian (Szende 1994), (D) Danish (Grønnum 1998), (E) Norwegian (Vanvik 1979), (F) Swedish (Engstrand 1999) and (G) Netherlandic Dutch (Gussenhoven 1992). The shaded polygon defines the English lax vowel subsystem.

The Scandinavian inventories offer a plethora of vowel types that would qualify as viable substitutes in English ('similar sounds' in Flege 1995). Every English monophthong can be mapped onto a distinct vowel in Norwegian at approximately the same position in the vowel space. The Danish inventory has no vowel pair corresponding to the *fool~full* contrast, but there are one-to-one mappings for all other English monophthongs. Swedish does have vowels matching this latter contrast but lacks the contrast between the *bed* and *bad* vowels. Dutch, lacks both the *fool~full* and the *bed~bad* contrast.

## 3 Method

The data for English spoken with American, Dutch and Mandarin accents were described by Wang & van Heuven (2006). For each language group ten male and ten female speakers were recorded. Non-native speakers were university students who had not specialized in English and had not spent time in an English-speaking environment. This type of speaker is representative of the typical ELF user in international settings. The  $3 \times 20$  speakers lived in the Netherlands at the time the recordings

were made, studying at Leiden University. The Hungarian ELF speakers (7 male, 10 female) were recorded in 2015 at the University of Pannonia (Hungary). The Scandinavian ELF speakers were recorded in 2016 at the University of Copenhagen (8 male and 12 female Danish speakers), the University of Oslo (10 male, 10 female Norwegian speakers) and the University of Stockholm (8 male, 12 female speakers). The inclusion criteria for the Hungarian and Scandinavian speakers were the same as for the Dutch and Chinese speakers.

Speakers produced all the 19 full vowels of English in a /hVd/ environment in a fixed carrier sentence *Now say h..d again*. Stimuli were presented to the speaker printed in normal English orthography on a sheet of paper. The pronunciation of the target vowels was exemplified by everyday key words rhyming with the /hVd/ targets (e.g. the unfamiliar target word *hoed* was cued by the more familiar words *road*, *showed*). Only the /hVd/ target words were used for acoustic analysis. Each speaker produced one token of each vowel; only the Hungarians recorded two tokens of each vowel type.

The onsets and offsets of the target vowels were determined by ear and by eye, using oscillograms and spectrograms. Formants were estimated by the Burg LPC algorithm implemented in Praat (Boersma & Weenink 1996). The optimal LPC model order and upper frequency cut-off were determined by trial and error, visually comparing formant tracks with the spectrogram. Vowel duration and the centre frequencies of maximally five formants were extracted; for each vowel token each formant frequency was averaged over the duration of the vowel. Formant frequencies were then psychophysically scaled in Barks (Traunmüller 1990). In all, measurements of 1,540 vowel tokens were available for statistical analysis.

## 4 Results

In American English as spoken in Southern California, ten vowels are recognized as monophthongs. The vowels in *caught*, *hot* and *father* have merged in that variety (and count as tense vowels in our analysis). The mid vowels in *pay* and *show* (not in Figure 1A) are commonly considered monophthongs (even though they are diphthongized to some extent in many varieties of English, including Californian). The central vowel in *bird* is not included since it only occurs immediately before /r/ – which makes it a positional allophone of the vowel in *but*. For the same reason we excluded all other /r/-coloured allophones.

Figure 2 displays the location of the ten English monophthongs in the acoustic vowel diagram with the first formant frequency (F1, representing vowel height) plotted from top to bottom, and the second formant frequency (F2, representing vowel backness and rounding) from right to left. This type of plot affords immediate visual comparison with the vowel diagrams in Figure 1.

The male and female vowel configurations are basically identical within each L1 group and yet differ systematically between groups. Using the American vowels as the reference set, we observe that there is a strict separation between the tense and lax subsystems. The six tense vowels (including /æ/ and /ɒ/) are on the outer perimeter of

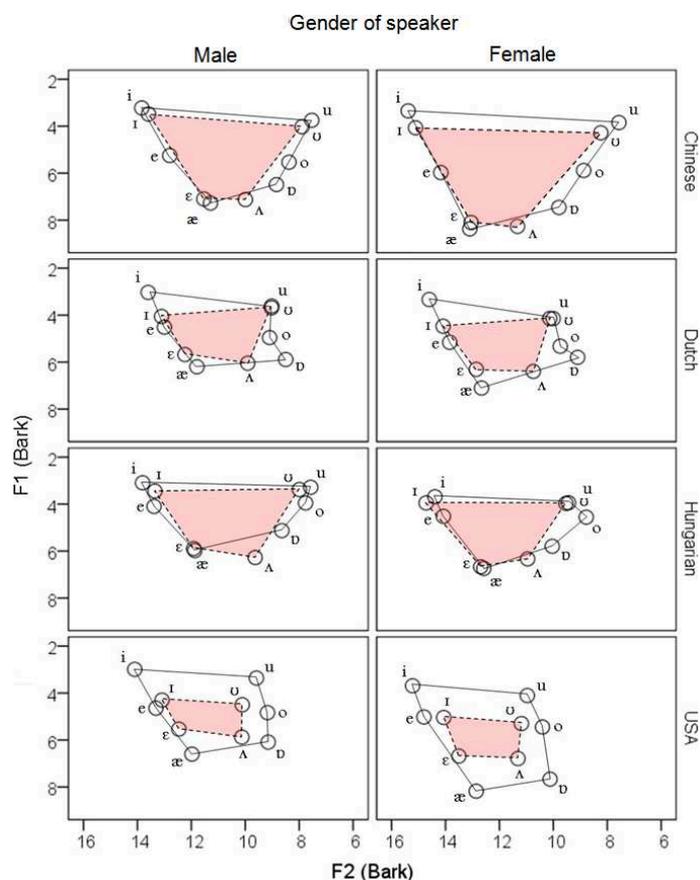


Figure 2: (Continued on page 142.)

the vowel space, while the four lax/short monophthongs form the corner points of an inner polygon. The members of the two pairs of mid vowels (/e, ɪ/ and /o, ʊ/) are rather close to one another in the spectral space but they are still distinct by a difference in duration – see below). This arrangement reproduces what has commonly been reported for (American) English and closely matches Figure 1A.

The three Scandinavian groups of ELF speakers display roughly the same organization of the English vowel system. They, too, can best be described in terms of an outer hexagon of six peripheral, tense vowels and an inner tetragon with four lax vowels. Details of the configurations differ between Danish, Norwegian and Swedish groups but the basic organization is a very good match with the native system. The Danes differ from the other Scandinavian groups in their location of /u/, which is as far back as the mid vowel /o/. In the pronunciation of the Norwegians and Swedes, however, /u/ is centralized, which mimics the way this vowel is pronounced in present day English (both in England and in the United States). It seems practically impossible to

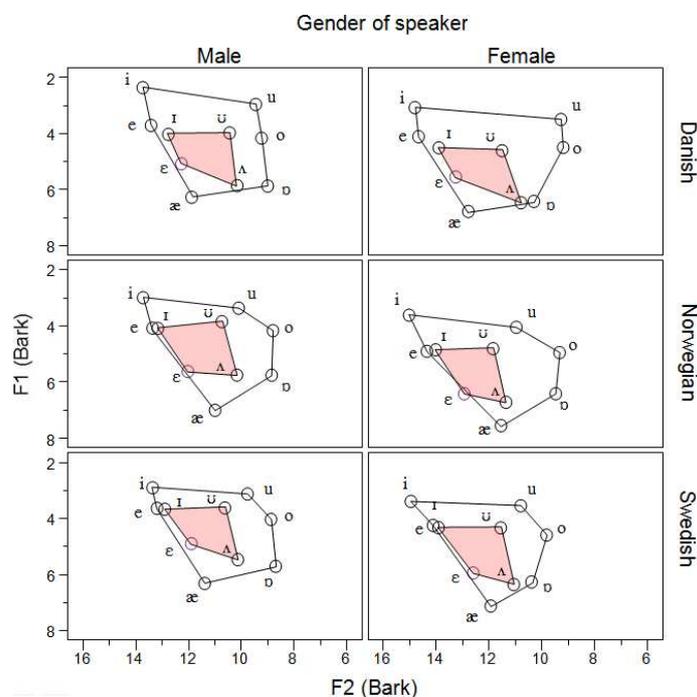


Figure 2: (Continued from page 141.) Mean F1 and F2 (Bark) of ten English monophthongs plotted for tense (open polygons) and lax (shaded polygons) vowels for 14 speaker groups.

tell the Norwegians and the Swedes apart on the basis of the spectral distribution of the English vowels. This is somewhat surprising, since we noted above that Swedes should be at a disadvantage when having to pronounce the contrast between / $\epsilon$ / and / $\text{æ}$ /, which does not exist in their L1 but does occur in Norwegian (and Danish). Either the Swedes have been so much exposed (through the media) to / $\text{æ}$ / in the neighboring languages Danish and Norwegian that this vowel is a familiar percept to them, or the pronunciation problem has been dealt with in the Swedish secondary school curriculum.

The Dutch ELF speakers deviate from English in several respects. Although the Dutch-accented vowels can also be divided into a tense and lax subsystem, the separation is poor in the bottom-left corner, where the contrast between / $\epsilon$ / and / $\text{æ}$ / is weak (though not completely absent) and no difference is made at all between / $u$ / and / $\text{ʊ}$ /.

The English vowel systems of the non-Indo-European groups do not seem to differentiate between the tense and lax subsystems – at least in terms of vowel quality. Although the Mandarin ELF speakers use a large vowel space – which confirms impressionistic claims (Zhao 1995) – they do not differentiate between / $i$ / and / $\text{ɪ}$ /, nor

between /u/ and /ʊ/. The vowel /æ/ is virtually the same as /ɛ/ while the distance between /ʌ/ and /æ, ɛ/ is so small that perceptual confusion can be expected.

The Hungarian speakers use a more contracted vowel space than the Mandarin speakers and show the same lack of contrast: poor separation between /i/ and /ɪ/, between /u/ and /ʊ/ and between /ɛ/ and /æ/. Moreover, the mid vowels /e/ and /o/ are spectrally quite close to /i, ɪ/ and /u, ʊ/, respectively – although they may still be distinct by duration.

Figure 3 plots the durations of the ten English vowels as produced by the seven speaker groups. Vowels are plotted from left to right in ascending order as determined for the native speaker group.

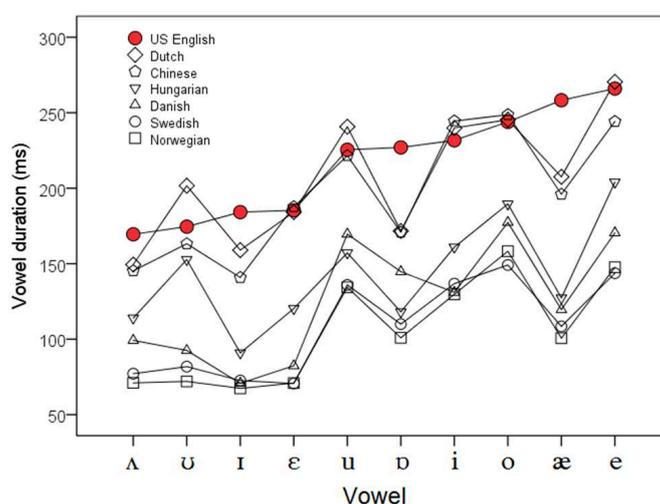


Figure 3: Duration (ms) of ten vowels of English produced by seven speaker groups. Vowels are in ascending order of length as observed for the US English native speaker group.

Although absolute durations differ from one language background to another (e.g. American, Dutch and Chinese speakers appear to produce longer vowels than Hungarian and Scandinavian speakers), there is a good deal of similarity in the relative vowel durations across the L1 backgrounds. The four lax vowels are indeed systematically shorter in the speech production of the native speakers, and these vowel durations do not overlap with any of the vowel durations of the six tense vowels. Crucially, the vowels /ɒ/ and /æ/ are clearly long. The same distribution of vowel duration is seen with the Scandinavian speakers: Danes, Swedes and Norwegians produce the four lax vowels with short durations, which are always shorter than any tense vowel duration. For the Scandinavian speakers, however, there is a tendency for the vowels /ɒ/ and /æ/, though longer than the lax vowels, to be shorter than the other tense vowels. The Hungarian speakers do not differentiate the duration of the tense versus lax members of the pairs /ɛ/~ /æ/ and /u/~ /ʊ/. The Dutch speakers have

slightly longer durations for the tense members of these two pairs (suggesting that these contrasts exist for at least some Dutch speakers of English).

As a last exercise we performed a series of 98 Linear Discriminant Analyses (LDAs). Each of the seven languages provided the training data in turn, and the resulting models were then tested on the same set of seven languages, yielding a  $7 \times 7$  matrix. Speaker normalization was performed prior to the LDA, by applying  $z$ -normalisation within individual speakers to vowel duration, F1 and F2. The results are shown in Figure 4, which plots the percentage of correctly identified vowels for each combination of training and test language. The left-hand panel presents the results when only F1 and F2 were entered as predictors; the right-hand panel shows the results of when the set of predictors was augmented with vowel duration. In both panels the data are plotted in ascending order of success of the training language with three predictors. Numerical values plotted in the right-hand panel are also given in Table 1.

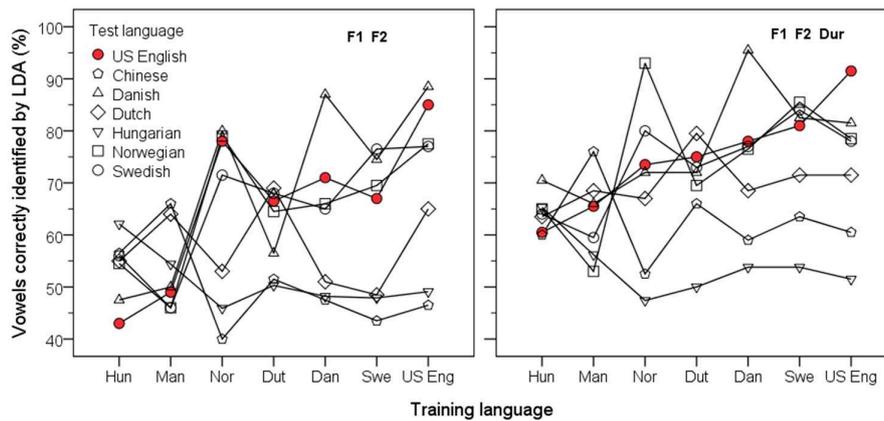


Figure 4: Correctly identified vowels by LDA (%) based on two (left-hand panel) or three (right-hand panel) acoustic predictors: F1, F2 and vowel duration as the optional third parameter. The models were trained in each of seven languages ( $x$ -axis), and tested in each of these languages (legend). There were 20 tokens (one token for each of 20 speakers) for each of ten English monophthongs (chance = 10%) per language, and 34 tokens (two tokens for each of 17 speakers with Hungarian L1). When training and test language were the same, the scores are based on cross-validation (leave-one-out method).

When including all three acoustic predictors, the vowels produced by the American native speakers are identified as intended most often when the test language is also US English (95.5% correct). The English vowels produced by Hungarian (51.5%) and Mandarin-Chinese (60.5%) ELF speakers are identified least successfully. The three Scandinavian groups are quite successful with scores around 80% correct. The Dutch ELF speakers are in the middle of the range (71.5%).

Table 1: Correct vowel identification (%) by LDA with F1, F2 and vowel duration as predictors (A). Also see Figure 4. Panel B lists the Relative ISIB values (see text).

		Training language						
		CHI	DUT	USA	HUN	DAN	NOR	SWE
Test language	CHI	<b>76.0</b>	66.0	60.5	60.0	59.0	52.5	63.5
	DUT	68.5	<b>79.5</b>	71.5	63.5	68.5	67.0	71.5
	USA	65.5	75.0	<b>91.5</b>	60.5	78.0	73.5	81.0
	HUN	56.2	50.0	51.5	<b>65.3</b>	53.8	47.4	53.8
	DAN	66.0	72.0	81.5	70.5	<b>95.5</b>	72.0	82.5
	NOR	53.0	69.5	78.5	65.0	76.5	<b>93.0</b>	85.5
	SWE	59.5	63.0	78.0	64.0	77.0	80.0	<b>84.0</b>

A. Raw scores

		Training language						
		CHI	DUT	USA	HUN	DAN	NOR	SWE
Test language	CHI	<b>13.0</b>	0.1	-7.4	-3.3	-8.6	-13.4	-5.0
	DUT	1.7	<b>9.9</b>	-0.1	-3.6	-2.8	-2.7	-0.8
	USA	-3.8	2.9	<b>17.4</b>	-9.1	4.2	1.3	6.2
	HUN	-2.6	-11.6	-12.1	<b>6.2</b>	-9.5	-14.3	-10.5
	DAN	-4.3	-1.2	6.3	-0.1	<b>20.6</b>	-1.2	6.7
	NOR	-16.0	-2.4	4.6	-4.3	3.0	<b>21.1</b>	11.0
	SWE	-9.1	1.5	4.5	-4.9	3.9	8.5	<b>9.9</b>

B. R-ISIB

The Interlanguage Speech Intelligibility Benefit (ISIB, see introduction) can readily be observed in panel B of table 1. The effect is seen most clearly if we convert the raw scores to relative ISIB scores (R-ISIB, Wang & van Heuven 2015) by subtracting the row and column means from each cell in the matrix, and then divide the cell contents by 2. This removes the main effects of training language and of test language from the results, so that only the interaction term remains – which is our R-ISIB measure. When the test data are presented after the LDA was trained with data from the same speaker group (which simulates the shared interlanguage), vowel identification scores are much better than when test and training data are from different speaker groups, with R-ISIB values of 14.1 and  $-2.3$ , respectively,  $t(47) = 6.2$  ( $p < .001$ ).

When test and training languages are different Scandinavian languages (six pairs), vowel identification is better than when Scandinavian languages are paired with other non-native Englishes (eighteen pairs) with mean R-ISIB values of 5.3 versus  $-6.0$ , respectively,  $t(22) = 4.8$  ( $p < .001$ ). This confirms our hypothesis that the

ELF vowels of Danes, Norwegians and Swedes resemble one another more than they are like the vowels produced by the other ELF speakers.

## 5 Conclusion

We studied the vowels of English produced by six groups of non-native (ELF) speakers and compared these tokens with the vowels produced by native speakers of American English. Non-native speakers were representative of the academically trained professional with no specialisation in English. The vowel configurations of the six ELF groups differed substantially from those of native English as well as from each other, in ways that could often – but not always – be predicted from traditional impressionistic vowel diagrams of the first language of the speakers. Scandinavian, and especially Danish ELF speakers, approximate the English vowels most closely, better than Dutch ELF speakers and much better than Hungarian and Mandarin-Chinese ELF speakers do. Counter to what Swarte's (2016) results would suggest, it is not the case that Dutch ELF is closer to native English than the Scandinavian ELF varieties are. It is beyond the scope of the present paper, however, to examine whether the (monophthongal) vowel systems yield different linguistic distances than other aspects of the phonology, vocabulary, morphology and syntax.

These conclusions follow from visual comparison of the acoustic vowel diagrams of the seven varieties of English and are quantitatively corroborated by Linear Discriminant Analyses in which American native listeners are simulated. Moreover, the ELF vowels produced by the three Scandinavian speaker groups resemble each other more than they share properties with the ELF vowels of Dutch, Chinese and Mandarin speakers. This would suggest that the phonologies, specifically the vowel systems, of the three Scandinavian languages are rather similar and produce the same type of transfer from native to foreign language.

Finally, our results are in line with the idea that similarity of non-native accents in English (or any other language) may serve as an experimental means to quantify phonological distance between languages even if these languages are genealogically unrelated.

## References

- Bent, Tessa & Anne R. Bradlow. 2003. The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America* 114. 1600–1610.
- Boersma, Paul & David Weenink. 1996. Praat, a system for doing phonetics by computer. *Report of the Institute of Phonetic Sciences Amsterdam* 132. [www.praat.org](http://www.praat.org).
- Cutler, Anne. 2012. *Native listening: language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.
- Engstrand, Olle. 1999. Swedish. In *Handbook of the International Phonetic Association: a guide to the usage of the International Phonetic Alphabet*, 140–142. Cambridge: Cambridge University Press.

- Flege, James E. 1995. Second language speech learning: theory, findings, and problems. In Winifred Strange (ed.), *Speech perception and linguistic experience: theoretical and methodological issues in cross-language speech research*, 233–277. Timonium, MD: York Press.
- Grønnum, Nina. 1998. Illustrations of the IPA: Danish. *Journal of the International Phonetic Association* 28. 99–105.
- Gussenhoven, Carlos. 1992. Dutch. *Journal of the International Phonetic Association* 22. 45–47.
- Hendriksen, Carol & Johan van der Auwera. 1994. The Germanic languages. In Ekkehard König & Johan van der Auwera (eds.), *The Germanic languages*, 1–18. New York: Routledge.
- Lee, Wai-Sum & Eric Zee. 2003. Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33. 109–112.
- Mannell, Robert, Felicity Cox & Jonathan Harrington. 2009. *An introduction to Phonetics and Phonology*, Macquarie University. <http://clas.mq.edu.au/speech/phonetics/>.
- Munro, Murrey J. & Tracey M. Derwing. 1995. Processing time, accent and comprehensibility in the perception of native and foreign accented speech. *Language and Speech* 38. 289–306.
- Rogerson-Revell, Pamela. 2007. Using English for international business: a European case study. *English for Specific Purposes* 26. 103–120.
- Strange, Winifred, Ocke-Sven Bohn, Kanae Nishi & Sonja A. Trent. 2005. Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America* 118. 1751–1762.
- Swarte, Femke. 2016. *Predicting the mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors*. Groningen: Groningen Dissertations in Linguistics 150.
- Szende, Tamás. 1994. Illustrations of the IPA: Hungarian. *Journal of the International Phonetic Alphabet* 24. 91–94.
- Trautmüller, Hartmut. 1990. Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America* 88. 97–100.
- Vanvik, Arne. 1979. *Norsk fonetikk*. Oslo: Universitetet i Oslo.
- Wang, Hongyan & Vincent J. van Heuven. 2006. Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. In Jeroen M. van de Weijer & Betteloe Los (eds.), *Linguistics in the Netherlands 2006*, 237–248. Amsterdam: John Benjamins. <http://hdl.handle.net/1887/14899>.
- Wang, Hongyan & Vincent J. van Heuven. 2015. The Interlanguage Speech Intelligibility Benefit as bias toward native-language phonology. *i-Perception* 6. 1–13. <http://ipe.sagepub.com/content/6/6/2041669515613661.full>.
- Zhao, Demei. 1995. *English phonetics and phonology: as compared with Chinese features*. Qingdao Shi: Qingdao hai yang da xue chu ban she.



## Chapter 16

# Impersonal passives in German: some corpus evidence

Erhard Hinrichs

Eberhard Karls University Tübingen

Nerbonne (1982b) offers cross-linguistic evidence that casts doubt on the unified Relational Grammar account of the passive construction as advancement of an object relation, as proposed by Perlmutter (1978) and Perlmutter & Postal (1977). The purpose of the present contribution is to present further corpus evidence in support of the semantic analysis of German impersonal passives proposed by Nerbonne (1982a; 1986). The corpus data are extracted from the TüPP-D/Z treebank of German, a linguistically annotated corpus, which uses as its data source the Scientific Edition of the taz German daily newspaper.

### 1 Introduction

Impersonal passive constructions have received considerable attention in syntactic theory in general and in the framework of Relational Grammar (RG) in particular, which considers grammatical relations as primitive elements of grammar. Perlmutter (1978) and Perlmutter & Postal (1977) propose a unified RG account of the passive construction as advancement of an object relation (in RG terminology: a 2-relation) to a subject relation (in RG terminology: a 1-relation). Personal passives as in (1) then differ from impersonal passives in that the former involve the promotion of an overt object NP, such as *John* in (1), and the latter promotion of a dummy element, such as German *es*, which can be overtly realized as in (2) or remain covert, as in (3).

- (1) John wurde geehrt.  
John AUX honored  
'John was honored.'
- (2) Es wurde gefeiert.  
it AUX celebrated  
'A celebration happened.'

*Erhard Hinrichs*

- (3) Hier wurde gefeiert.  
here AUX celebrated  
'Here a celebration happened.'

As Nerbonne (1982b) points out, only a subset of intransitive verbs can appear in the impersonal passive construction in German – typically those that imply intentionality on the part of the agent of the action described by the verb. If intentionality is absent, as in the case of *explodieren* 'explode' in (4), then the impersonal passive is unacceptable.

- (4) \* Es wurde explodiert.  
it AUX exploded  
(intended:) 'An explosion was performed'.

The contrast in grammaticality of impersonal passives with unaccusative verbs such as *explodieren* in (4) and with unergative verbs such as *feiern* in (2) and (3) is put forward by Perlmutter (1978) as supporting evidence for the RG analysis of impersonal passive and is explained in terms of the combined effect of the following three independently motivated principles: the Unaccusative Hypothesis, the Final 1 Law, and the 1-Advancement Exclusiveness Law. According to the Unaccusative Hypothesis, certain intransitive clauses such as *explodieren* have an initial 2-relation and no initial 1-relation. The Final 1 Law states that clauses with final unaccusative strata are not well-formed. Therefore, the initial 2-relation in an unaccusative stratum must be promoted to a 1-relation. The 1-Advancement Exclusiveness Law states that no clause can involve more than advancement to a 1-relation. Since impersonal passives involve the promotion of a dummy element to a 1-relation and since an unaccusative clause involves promotion of an initial 2-relation, impersonal passives of unaccusative clauses are then ruled out since they would have to involve two separate advancements to a 1-relation, which is ruled out by the 1-Advancement Exclusiveness Law. By contrast, impersonal passives with unergative verbs as in (2) are not precluded by the 1-Advancement Exclusiveness Law since they involve a single 1-advancement of a dummy.

Perlmutter (1978) and Perlmutter & Postal (1977) put forth the hypothesis that promotion of a dummy to subjecthood is a language universal and that a strong version of the Unaccusative Hypothesis considers the distinction between unaccusativity and unergativity uniform cross-linguistically. In a cross-linguistic study of impersonal passive constructions, Nerbonne (1982b) cites data from Estonian, German, Irish, and Lithuanian that call into question the strong version of the Unaccusative Hypothesis and the universal characterization of impersonal passives as promotion of a dummy element to subject.

## 2 Impersonal passives in German

Nerbonne (1982b) and Nerbonne (1986) present a wide range of empirical findings for German that dummy *es* in impersonal passives of German is not a subject, contrary

to the prediction of the RG analysis of Perlmutter and of Perlmutter and Postal. The evidence includes *inter alia* the fact that dummy *es* in impersonal passives can only appear in clause-initial position of German V2 clauses.

The lack of overt subjects in German impersonal passives is not limited to impersonal passives with intransitive verbs, but extends also to impersonal passives of verbs with dative objects. The German verb *helfen* ‘help’ takes a dative object and can be passivized, as shown in (5).

- (5) a. Dir wird geholfen.  
       you AUX helped(PART)  
       ‘You receive help.’
- b. Ihnen wird geholfen.  
       them AUX helped(PART)  
       ‘They receive help.’
- c. Es wird ihnen geholfen.  
       it AUX them helped(PART)  
       ‘They receive help.’

All examples in (5) are impersonal passives, with the sentence-initial position either occupied by the dative object of *helfen*, as in (5a) and (5b), or by the dummy element *es*. Please note the lack of number agreement between the sentence-initial second-person singular and third-person plural pronouns in (5a) and (5b) and the third-person singular finite form *wird* of the auxiliary as further evidence that the pronouns are not subjects.

## 2.1 The Unaccusative Hypothesis

Nerbonne (1982b) shows that the distinction between unaccusativity and unergativity is far from uniform cross-linguistically. He cites examples such as (6b), where an impersonal passive is formed with an un-accusative verb such as *sterben* ‘die’, which should not be possible according to the strong version of the Unaccusative Hypothesis of Perlmutter (1978).

- (6) a. Autofreiheit. Und dafür ist es [das Volk] auch gerne bereit zu zahlen.  
       Mit abgeholzten Wäldern, mit stinkender Luft und einem verbogenen  
       Rückgrat. Weil das natürlich auch Freiheit ist.  
       ‘Car freedom. And for that it is also gladly willing to pay. With roded  
       forests, stinking air, and a bent spine. Because that is, of course, also  
       freedom.’
- b. Ganz nebenbei: es wird auch gestorben für diese Freiheit.  
       entirely aside: it AUX also died(PART) für this freedom  
       ‘Incidentally: people die for this freedom.’

In the linguistic context immediately preceding (6b), shown in (6a), (6b) is mentioned as an instance of the willingness to sacrifice oneself in support of *Autofreiheit*.

Thus, dying is portrayed as an intentional act, albeit not necessarily as a conscious choice of any victim involved. Or as Nerbonne (1982b) puts it: they do what might be done willingly. More generally, Nerbonne assumes that the semantics of the impersonal passive construction in German carries volitionality as an implicature. Therefore, his semantic account of impersonal passives predicts that verbs such as *sterben* can occur in this construction precisely in contexts such as (6), where volition is implied.

## 2.2 Impersonal passives and reflexives

Nerbonne (1986) notes that the range of verbs that can appear in the impersonal passive construction in German excludes transitive verbs with accusative objects, which form personal passives with overt subjects. However, there seem to be exceptions to this empirical generalization, as the examples in (7), due to Bierwisch (2006), and in (8), due to Nerbonne (1982b), show.

- (7) a. Hier wird sich gründlich gereinigt!  
here AUX self carefully cleaned(PART)  
'Here one cleans oneself carefully.'
- b. Jetzt wird sich nicht unterhalten!  
now AUX self not conversed(PART)  
'No conversation!'
- (8) Jetzt wird sich versammelt.  
now AUX self gathered(PART)  
'People should now gather.'

Verbs such as *reinigen* 'clean', *unterhalten* 'converse, entertain', and *versammeln* 'gather' can appear in impersonal passives with reflexive *sich* as their accusative objects.

The grammaticality judgments for impersonal passives with reflexive *sich* tend to show quite a bit variability across speakers, however. Nerbonne (1982b: 90) comments on the acceptability of example (8) as follows: "These examples sound abominable to many speakers of German, but are perfectly acceptable, if a bit pushy, to many others, particularly in the South." While Bierwisch (2006) rates examples (7) as acceptable, he considers the examples in (9) as unacceptable and highly marginal.

- (9) a. \* Es ist sich häufig rasiert worden.  
it is self often shaved(PART) AUX  
(intended) 'People shaved themselves frequently.'
- b. ?? Heute wird sich geärgert.  
today AUX self be irritated (PART)  
'Today people are irritated.'

One difference between examples (9) and examples (7)-(8) concerns their illocutionary force. The sentences in (7) and (8) are commands, while the sentences in (9)

are not. Since reflexives typically require an antecedent for co-reference, the acceptability of impersonal passives, when used as commands, may be due to the addressees of the command serving as implied discourse antecedents for the reflexive.

### 3 Corpus data

With the exception of (6b), the above discussion has been based on grammaticality judgement of native speakers reported in the linguistics literature. With the availability of large linguistically annotated text corpora of German, it seems appropriate to dig further into the data available for impersonal passives in German and to obtain additional data points for the range of verbs that can appear in this construction. The corpus study below is based on the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z).<sup>1</sup>

#### 3.1 The TüPP-D/Z corpus

The TüPP-D/Z treebank uses as its data source the Scientific Edition of the taz German daily newspaper,<sup>2</sup> which includes articles from September 2, 1986 up to May 7, 1999. The corpus consists of 11,512,293 sentences with a total of 204,425,497 tokens. The texts are processed automatically, starting from paragraph, sentence, word form, and token segmentation. All sentences have been automatically annotated with clause structure, topological fields, and chunks, as well as parts of speech and morphological ambiguity classes. The topological field model (Herling 1821; Erdmann 1886; Drach 1937; Höhle 1986) is used to account for regularities in sentence structure and word order across different clause types of German. The part-of-speech annotation uses the STTS labels of the Stuttgart-Tübingen tagset (Schiller et al. 1999), the de-facto standard for the part-of-speech labelling of German text corpora.

Figure 1 illustrates the annotation layers of the TüPP-D/Z, using the passive sentence *Posthum wird der Künstler in diesem Jahr mit zahlreichen Veranstaltungen geehrt*. (Engl.: ‘After his death, the artist is honored this year by numerous events.’). The sentence is a verb-second main clause and hence classified as a V2 clause. This V2 node is further annotated by the topological fields VF, short for: Vorfeld (‘initial field’), the labels VCL and VCR for the left and right bracket of the clause, respectively, and the label MF, short for: Mittelfeld (‘middle field’). The left and right brackets of the clause are realized by the finite auxiliary (FA) *wird* and by the past participle (VP) *gehrt*, respectively. Hence these two labels are appended to the labels for the topological labels VCL and VCR. The Mittelfeld of the sentence in Figure 1 contains three constituents labelled as NC (short for: noun phrase chunk) and PC (short for: prepositional phrase chunk). The Vorfeld of a German V2 clause typically consists of a single constituent; this is also the case for the sentence in Figure 1, where this constituent is erroneously labelled as a noun phrase chunk (NC). This error is due to the mistagging of the word *posthum* by the STTS label NE (short for: named entity), instead of the

<sup>1</sup> [www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tuepp-dz.html](http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tuepp-dz.html)

<sup>2</sup> [www.taz.de](http://www.taz.de)

correct label of an adverb. This mistake then percolates up to the incorrect labelling at the chunk level. The annotation mistake for the Vorfeld constituent underscores the fully automatic, and thus error-prone annotation of the TüPP-D-Z, a fact that needs to be kept in mind when querying this treebank. A more in-depth description of the linguistic annotation can be found in the TüPP-D/Z stylebook (Müller 2004), and information about the actual XML encoding of linguistic annotation can be found in the TüPP-D/Z markup guide (Ule 2004).

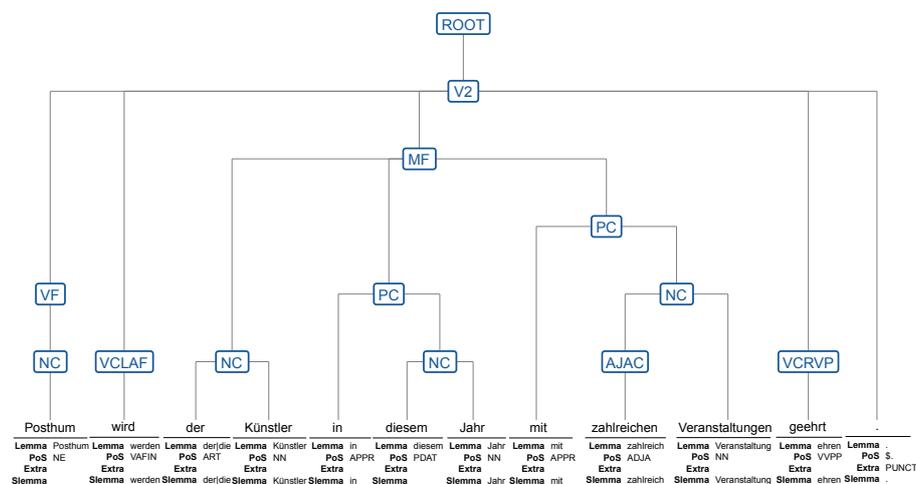


Figure 1: Example sentence from the TüPP-D/Z treebank of German.

### 3.2 TüPP-D/Z query results

This section summarizes the query results for the impersonal passives found in the TüPP-D/Z treebank. The results were obtained by the TüNDRA web application (Martens 2013), which uses the TIGERSearch (Lezius 2002) query language for treebank search. The corpus findings will focus on the empirical issues discussed in sections 2.1 and 2.2 above.

#### 3.2.1 Impersonal passives and unaccusative verbs

As discussed in section 2.1, Nerbonne’s semantic analysis of the impersonal passive construction in German assumes that there is an implicature of intendability associated with this construction. In Nerbonne’s account, this implicature is crucial for a semantic characterization of the range of verbs that can appear in German impersonal passives. In section 2.1, the verb *sterben* ‘die’, as used in example (6b), was mentioned as a case in point. It turns out that such examples, while rare, also occur

in the TüPP-D/Z treebank. The TüPP-D/Z contains twelve occurrences of impersonal passives with the verb *sterben* that include example (10).

- (10) Dort toben Kämpfe, dort wird gestorben.  
 there rage fights there AUX died(PART)  
 ‘There fights are raging, there dying is happening.’

Even more numerous than examples with the verb *sterben* are impersonal passives in the TüPP-D/Z with verbs such as *leiden* ‘suffer’ and *schwitzen* ‘perspire’. Such verbs denote involuntary acts, but allow readings of willingly submitting to such situations, as is implied in examples (11) and (13).

- (11) Da wird mit Größe gelitten und gestorben.  
 There AUX with greatness suffered(PART) and died(PART)  
 ‘There people suffer and die with greatness.’
- (12) Hier wird nicht geweint, hier wird auch nicht gelacht, hier wird  
 here AUX not wept(PART) here AUX also not laughed(PART) here AUX  
 gelitten.  
 suffered(PART)  
 ‘Here people do not weep, here there is no laughter, here they suffer.’
- (13) An den Tischen wird heftig gemurmelt und geschwitzt.  
 at the tables AUX intensively mumbled and perspired.  
 ‘At the tables there is intensive mumbling and sweating going on.’

### 3.2.2 Impersonal passives and reflexives

As discussed in section 2.2, the grammaticality judgments for impersonal passives with reflexive *sich* tend to show quite a bit of variability across speakers. Even in a corpus as large as the TüPP-D/Z with appr. 11.5 million sentences, the combination of impersonal passives and reflexives is rather rare. The main finding common to all TüPP-D/Z examples below is that the reflexives that occur in this construction tend to be so-called inherent reflexives, i.e., reflexives that do not occupy an ordinary argument position, but rather appear as improper arguments in the terminology of Bierwisch (2006). Such inherent reflexive verbs include *sich vergnügen* ‘to enjoy oneself’, *sich amüsieren* ‘to amuse oneself’, *sich genieren* ‘to feel embarrassed’, *sich freuen* ‘to rejoice’, and *sich entblöden* ‘to have the effrontery to do something’ in examples (14)-(18).

- (14) Auf in den Kampf, jetzt wird sich vergnügt!  
 up in to combat now AUX self enjoyed  
 ‘Up into combat, now let’s have fun.’
- (15) Was soll’s, jetzt wird sich amüsiert.  
 what shall it now AUX self amused  
 ‘Who cares, now let’s have a good time.’

- (16) Der Ruf der Stadt ist ruiniert – wo man auch hinschaut, es wird sich  
the image of the city is ruined – where one also looks it AUX self  
geniert.  
be embarrassed  
'The reputation of the city is ruined – wherever one looks, people are em-  
barassed.'
- (17) ... egal welches Wahlergebnis: es wird sich immer gefreut  
... no matter what election results it AUX self always happy  
'No matter what election result: people are always happy.'
- (18) ... und selbstbestimmtem Leben – es wird sich nicht mal entblödet, das für  
... and self-determined life – it AUX self foolish enough, that for  
Inhaftierte in den Knästen zu fordern!  
imprisoned in the jails to demand  
'... and autonomous life – people have the effrontery to demand that for the  
prisoners in jails!'

In addition to mono-lexemic inherent reflexives, the TüPP-D/Z also contains im-  
personal passives with inherent reflexives of multi-word expressions such as *sich die  
Mühe machen* 'to make the effort to' and *sich lustig machen* 'make fun of' in (19) and  
(20).

- (19) Und nun wird sich nicht einmal die Mühe gemacht, in der PDS genauer zu  
and now AUX self not even the effort made in the PDS sharper to  
rechnen.  
calculate  
'And now they do not even make an effort in the PDS to calculate more accu-  
rately.'
- (20) Es wird sich über Verlagshinweise zu den stetig steigenden Preisen  
it AUX self about publishers' comments on the always rising prices  
und veränderten Busfahrplänen lustig gemacht.  
and changed bus schedules fun made of  
'Fun is made of publishers' comments about steadily rising prices and  
changed bus schedules.'

While inherent reflexives make up the majority of data points for impersonal passives  
with reflexive *sich*, the TüPP-D/Z examples in (21) and (22) arguably involve transi-  
tive verbs with reflexive *sich* in a proper argument position of direct object. The verbs  
in question are *vereinigen* 'to unite' in (21) and *lieben* 'to make love' in (22), whereas  
the verb *sich sehnen* 'to long for' in the same sentence is an inherent reflexive verb.

- (21) ... es wird sich vereinigt und Otto Grotewohls und Wilhelm Piecks  
... it AUX self united and Otto Grotewohls und Wilhelm Piecks  
Händedruck auf Großfoto umkringelt.  
handshake on poster circled  
'People unite and circle Otto Grotewohls' und Wilhelm Pieck's handshake on

poster-sized photo.’

- (22) a. Ein Ehepaar (Nela Barsch und Jürgen Wink) bekommt Besuch (Eva  
a couple (Nela Barsch and Jürgen Wink) have visitors (Eva  
Manschott und Viktor Schefe),  
Manschott und Viktor Schefe)  
‘A married couple (Nela Barsch and Jürgen Wink) have visitors (Eva  
Manschott and Viktor Schefe).’
- b. es wird sich überkreuz gesehnt und geliebt.  
it AUX self crosswise longed and loved  
‘They long for each other and make love across couples.’

It is worth noting that both examples have the pragmatic force of directives. (21) describes politically correct behavior for citizens of the former German Democratic Republic, and (22) the plot of a theatre play. Hence, the addressees of the directives may serve as discourse-implied antecedents for the reflexives and make these, albeit extremely rare, examples acceptable in the same way as was hypothesized for examples (7) and (8) in section 2.2 above.

## 4 Conclusion

This paper has reported on a corpus study of German impersonal passives that uses the TüPP-D/Z treebank of contemporary German as its data source. The corpus findings lend further empirical support to the semantic analysis of German impersonal passives proposed by Nerbonne (1982a; 1986). More specifically, the corpus data confirm two claims inherent in Nerbonne’s analysis: (i) that there is an implicature of intendability associated with German impersonal passives, which can account for the acceptability of certain unaccusative verbs in this construction, and (ii) that the use of reflexive *sich* in German impersonal passives is restricted to the class of inherent reflexives.

## References

- Bierwisch, Manfred. 2006. German reflexives as proper and improper arguments. In Patrick Brandt & Erich Fuss (eds.), *Form, structure, and grammar: a festschrift presented to Günther Grewendorf on occasion of his 60th birthday*, 15–36. Berlin: Akademie Verlag.
- Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*. Frankfurt am Main.
- Erdmann, Oskar. 1886. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt. Erste Abteilung*. Stuttgart: Cotta.
- Herling, Simon. 1821. *Über die Topik der deutschen Sprache. Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache*, 394, *Drittes Stück*, 296–362. Frankfurt am Main.

- Höhle, Tilman. 1986. Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne (ed.), *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen*, 329–340. Tübingen: Niemeyer.
- Lezius, Wolfgang. 2002. TIGERSearch ein Suchwerkzeug für Baumbanken. In *Tagungsband zur Konvens 2002*.
- Martens, Scott. 2013. Tundra: a web application for treebank search and visualization. In Kiril Simov & Petya Osenova (eds.), *Proceedings of the twelfth Workshop on Treebanks and Linguistic Theories*, 133–144. Sofia University.
- Müller, Frank Henrik. 2004. *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Technical report. Tübingen, Germany: University of Tübingen, Seminar für Sprachwissenschaft.
- Nerbonne, John. 1982a. German impersonal passives: a non-structure-preserving lexical rule. In Dan Flickinger, Marcy Macken & Nancy Wiegand (eds.), *Proceedings of the 1st West Coast Conference on Formal Linguistics*.
- Nerbonne, John. 1982b. Some passives not characterized by universal rules: subjectless impersonals. In Brian D. Joseph (ed.), *Grammatical relations and relational grammar*, vol. 26 (OSU Working Papers in Linguistics), 59–92. Department of Linguistics, The Ohio State University, Columbus, OH.
- Nerbonne, John. 1986. A phrase-structure grammar for German passives. *Linguistics* 24(5). 907–938.
- Perlmutter, David M. 1978. Impersonal Passives and the Unaccusative Hypothesis. In Jeri J. Jaeger, Anthony C. Woodbury, Farrell Ackerman, Christine Chiarello, Orin D. Gensler, John Kingston, Eve E. Sweetser, Henry Thompson & Kenneth W. Whistler (eds.), *Proceedings of the fourth Annual Meeting of the Berkeley Linguistics Society*, 157–189.
- Perlmutter, David M. & Paul M. Postal. 1977. Fixme: universal character of passivization. In Kenneth Whistler, Robert D. Van Valin Jr., Chris Chiarello, Jeri J. Jaeger, Miriam Petruck, Henry Thompson, Ronya Javkin & Anthony Woodbury (eds.), *Proceedings of the third Annual Meeting of the Berkeley Linguistics Society*, 394–417.
- Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Technical report. Universität Stuttgart, Universität Tübingen.
- Ule, Tylman. 2004. *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Technical report. Tübingen, Germany: University of Tübingen, Seminar für Sprachwissenschaft.

## Chapter 17

# *In Hülle und Fülle* – quantification at a distance in German, Dutch and English

Jack Hoeksema  
University of Groningen

### 1 Introduction

Quantifiers are often equated with determiners and pronouns, such as *every* or *everybody*. Occasionally, adverbs of quantification are considered (Lewis 1975; de Swart 1991), or floating quantifiers (Dowty & Brody 1984; Sportiche 1988; Bach et al. 1995; Hoeksema 1996; Doetjes 1997; McCloskey 2000; Bošković 2004) The former are illustrated in (1-2) below, the latter in (3-4):

- (1) Polar bears are often dangerous.
- (2) Christmas parties are sometimes fun.
- (3) The polar bears were all hungry.
- (4) The Christmas parties were neither of them any fun.

Both types of expressions are well-known and have been well-studied. As noted by Lewis (1975) and subsequent literature, the sentences in (1-2) are equivalent to those in (5-6) on one of their readings:

- (5) Many polar bears are dangerous.
- (6) Some Christmas parties are fun.

The sentences in (3-4), in turn, are equivalent to the ones in (7-8):

- (7) All the polar bears were hungry.

- (8) Neither Christmas party was any fun.

The reader will have noted that the first two examples have indefinite subjects, whereas the second pair has definite subjects. This is typical for the two types of adverbial quantifiers. Adverbs of quantification have an effect on indefinites which they lack for definites. Sentences such as (9) below only have a temporal interpretation, except when they are given a kind-reading (“this kind of polar bear”):

- (9) This polar bear is often dangerous.

Recently, Morzycki (2011) called attention to a third type of adverbial quantifier, represented by the English expression *galore*. Like the adverb of quantification *often*, it may express the same quantificational force as *many*, but unlike *often*, it does not have an additional temporal reading. Morzycki (2011) argues that *galore* is a quantifier operating on kinds, in order to account for the fact that it is used mainly in combination with bare noun phrases:

- (10) We had polar bears galore on Nova Zembla.

Note that the following variants of (10) are ill-formed:

- (11) We saw {\*my/\*the/\*some/\*all/\*many} polar bears galore.

Morzycki (2011) proposes that *galore* is a stranded determiner, left behind after the NP is moved out of the DP, as follows (where *t* indicates the original position of the NP polar bears).

- (12) We saw [galore polar bears] → We saw [polar bears [galore *t*]]

Given a ban on double determiners, the ungrammaticality of (11) is accounted for. Note that determiners in general seem to be ruled out, not just definite determiners, or indefinite ones. In this respect, *galore* differs from both floating quantifiers and adverbs of quantification. It appears to form a third class of quantificational adverb. This raises the question whether it is one of a kind, or has counterparts, either in English or in other languages. In Hoeksema (2012), I argued that Dutch *bij de vleet* ‘in droves’ is in fact such a counterpart, and in the present paper, I will venture to do the same for German *in Hülle und Fülle*. If these claims can be substantiated, they are somewhat problematic for the analysis proposed by Morzycki, since the Dutch and German cases are syntactically prepositional phrases, and not lexical heads.

Before outlining the structure of the paper, let me say a few things about *in Hülle und Fülle* and *bij de vleet*. The rhyming idiom *Hülle und Fülle* consists of a word meaning ‘cover, clothing’ and another meaning ‘filling, food’. Together, they are all one needs to survive. At some point, this developed into the meaning ‘plenty’. Dutch *bij de vleet* contains the otherwise obsolete noun *vleet*, meaning fishing net. One might translate it as ‘by the net’. Fishermen caught herring *bij de vleet* in large numbers, and so this expression came to mean ‘in large numbers, galore’.

In the next section, I compare the distribution of *galore* with that of its proposed Dutch and German counterparts. The distributional data are from electronic corpora of the three languages. Section 3 contains the conclusions.

## 2 Distributional facts

### 2.1 The bare noun condition

We saw in the previous section that *galore* requires bare nouns as its associates. I will now show that this finding can be replicated for *in Hülle und Fülle* and *bij de vleet*. First, let's consider a mass noun:

- (13) Wir hatten Zeit in Hülle und Fülle  
we had.PL time a-plenty  
'We had lots of time.'

Adding a determiner to the noun in (13) leads to a degraded result in much the same way as it did in (11) above.

- (14) Wir hatten die/diese/einige Zeit in Hülle und Fülle  
we had.PL the/this/some time a-plenty  
'We had lots of the/this/some time.'

For Dutch, we can show a similar array of judgments:

- (15) We hadden geld bij de vleet  
we had.PL money a-plenty  
'We had money galore.'
- (16) We hadden het/zijn geld bij de vleet  
we had.PL the/his money a-plenty  
'We had the/his money galore.'

Nonetheless, there are several fairly robust and interesting exceptions to this generalization. For starters, the indefinite determiner *zulk* 'such' is acceptable with the associates of *bij de vleet*:

- (17) Zulke problemen waren er bij de vleet  
such problems were there galore  
'Such problems, there were galore.'

This observation is not restricted to Dutch. The Internet provides similar examples for English, e.g.

- (18) Humanity has had such problems galore.

as well as for German:

- (19) Solche Probleme gab es in Hülle und Fülle  
such problems there were galore  
'Such problems, there were galore.'

The case of *such* is interesting and not entirely straightforward. Among determiners, it enjoys a rather special status. In some ways it behaves as an adjective, in other ways it looks more like a determiner. To see this, just consider the following paradigm:

- (20) a. \* the such book  
b. \* a such book  
c. such a book  
d. no such book  
e. \* this such book  
f. such nice books  
g. every such book

While examples d and g are most compatible with adjectival status, the other data point more into the direction of determiner status. van der Auwera & Kalyanamalini (2015) treat it as a unique element, belonging to a category of its own. Consequently, we need not treat it as a direct counterexample to Morzycki's bare noun generalization.

More problematic are other cases, not discussed by Morzycki, where the associate of *galore* is not a bare NP. Pronouns, for instance, may serve as associates when they refer to a kind. The following is one of many examples to be found on the Internet:

- (21) I'm not sure what planet you live on, coz honour killings are very much part of the 'culture' in Pakistan. I did not say religion, but the culture. Open any Pakistani newspaper and you'll find them galore, and the men feel quite proud in admitting to it.

The use of *them* in the above example to refer back to the category of *honour killings* is acceptable, it would seem, precisely because it is a case of common noun anaphora. When *them* is referring to a set of individuals, it does not lend itself to be the associate of *galore*. Compare:

- (22) Q: Where are the boys?  
A: # O, I saw them galore.

Similar cases can be found in German and Dutch, respectively:

- (23) Probleme? Die haben wir in Hülle und Fülle  
problems those have we galore  
'Problems? We have them galore.'
- (24) Problemen? Die hebben we bij de vleet  
problems those have we galore  
'Problems? We have them galore.'

Finally, it would seem that full definite DPs, while clearly exceptional and rare, are not entirely ruled out either, compare the examples in (25, 26) from English, (27) from German and (28) from Dutch. The latter two are not from corpora but were adapted from corpus examples.

- (25) It's corroborated across Internet sites by the personal testimonies galore from women. [COCA]
- (26) Hallmark has those things galore. [Internet]
- (27) Das Geld steht in Hülle und Fülle zur Verfügung  
the money stands galore to disposition  
'The money is available in abundance.'
- (28) De boeken werden bij de vleet verkocht  
the books were galore sold  
'The books were sold in large numbers.'

We might take this as evidence against the bare noun generalization, but the fact of the matter is that the vast majority of cases involves bare nouns. I collected 100 examples of each of the three items studied here, *galore*, *in Hülle und Fülle* and *bij de vleet*. Examples were taken from the online corpora made available by Mark Davies (at corpus.byu.edu, cf. e.g. Davies 2011), the Institut für deutsche Sprache (IDS) in Mannheim (a.k.a. COSMAS II), and the Dutch newspaper corpora at delpher.nl and LexisNexis, as well as Google Books. All cases where the items and their associates did not form a full sentence were put aside. This is a very large group, since the three expressions are very popular in titles and section headers, photo captions, names of web shops and websites, etc. A typical example of this usage is seen in the following text, the introduction of a text announcing and advertising a cultural festival in Christchurch, New Zealand:

- (29) Culture Galore! All cultures are welcomed and celebrated at this free annual event which features music, dance, food and arts and crafts from more than 80 cultures from around the globe.

The data sets were divided into bare nouns, pronouns (personal, relative, demonstrative), DP (definite and universal noun phrases, including universal pronouns, such as *all*), *such N* (German *solche N*, Dutch *zulke N*), and the category *zero* when there was no overt nominal associate (more about this in section 2.3). The results are given in Table 1.

I assume that the strong preference for bare nouns has a semantic cause. It does not seem to stem from syntactic selection of noun phrases by determiners, a possibility that may seem acceptable for *galore*, but which lacks plausibility for the PP quantifiers *bij de vleet* and *in Hülle und Fülle*. Morzycki (2011) provides the following interpretation for *galore*, where  $k$  is a variable ranging over kinds, and  $\sqcup$  is an operator due to Chierchia (1998), sending kinds (in the sense of Carlson 1977) to properties. E.g. it will map mankind to the set of human beings. *Galore* acts as a quantifier comparable to *many/much*, differing from those quantifiers however in not making a

Table 1: Corpus data on type of associate, per language.

associate	English	German	Dutch
bare noun	94	89	88
pronoun	2	4	6
DP	3	6	4
such N	–	1	1
zero	1	–	1

distinction between plural count nouns and mass nouns. Applying the interpretation in (30) to a sentence such as *there were spiders galore* yields the interpretation: there were many instances of the kind spider.

$$(30) \llbracket \text{galore} \rrbracket = \lambda k \lambda g_{\langle e, t \rangle} . \exists x [\sqcup k(x) \wedge g(x) \wedge \text{amount}(x) \succ \text{standard}]$$

Assuming that pronouns may refer to kinds (Carlson 1977), and that the DPs and such NPs noted in Table 1 may also be viewed as kind-denoting, we could adopt Morzycki’s analysis for our purposes. While the bare plural *problems* denotes the kind of problems, the DP *such problems*, one might postulate, denotes the subkind of problems characterized by a given exemplar. To motivate this assumption, I want to argue that a DP such as *those cars* is acceptable as an associate of *galore* only in certain contexts. Compare:

- (31) \* Donald Trump owned those cars galore.  
(32) Donald Trump owned many of those cars.  
(33) Cadillacs are pretty rare in Europe, but in the USA we have those cars galore.

In other words, when *those cars* denotes a particular set of cars, such as the set (partially) owned by Trump, *galore* is no good, but when we are talking about a brand/type of car, we may use *galore* to quantify over exemplars of that brand. Note that some other types of generic DPs seem to be ruled out with *galore*, such as *an N* and *the N(s)*:

- (34) \* The tiger hunts alone galore.  
(35) \* For lunch, I prefer a light meal galore.  
(36) \* I am rooting for the Dutch galore.

In the case of generic indefinites such as *a light meal*, we may refer to Krifka et al. (1995), who argue that such indefinites do not refer to kinds. Kind-level predicates such as *extinct* do not apply to them, compare:

- (37) The dodo is extinct.  
(38) # A dodo is extinct.

Note that there is one reading on which (38) is acceptable, namely when the sentence means that a kind of dodo is extinct. It seems to me that on this particular reading, *galore* should be acceptable:

(39) Mostly, hominins are extinct, but we have one hominin *galore*: homo sapiens.

How about cases such as (34) and (36)? It seems very likely, that for these sentences, we need to look at the predicates as well. The interaction between predicate type and generic/non-generic DPs is complicated, but crucial (cf. Carlson 1977; Krifka et al. 1995; Oosterhof 2008; Barton, Kolb & Kupisch 2015). In the next section, we take a look at the predicates that show up with *galore*.

## 2.2 Predicates *galore*

In Carlson (1977), a three-way distinction is made between kind-level, individual-level and stage-level predicates, exemplified by (40), (41) and (42), respectively:

(40) Rats are not a threatened species.

(41) Rats are smart.

(42) Rats are available for vivisection.

Kind-level predicates are ruled out by the semantics of *galore*. E.g. *rats galore* quantifies over instances of the kind rats, according to Morzycki's analysis. Kind-level predicates apply to kinds, not their instances. The associates of *galore* come in three main groups: subjects, direct objects and objects of prepositions. The category of objects of prepositions is fairly marginal in my data set, except for English, where it accounts for 17 (out of 100) occurrences, hence I will ignore these cases in what follows. Only very rarely, a predicate nominal is accompanied by *galore* ("it will be bargains *galore* at the fire sale"). When the associate is a subject, the predicate is mostly *there + be*, a passive, or a locative construction. The following table gives an overview of the subject associates for the three languages studied. I treat the "subjects" of existential constructions all the same, although for German *es gibt* one might argue it involves an object (bearing accusative case). However, for the same of comparison, it seems best to treat the functionally equivalent DPs in *there is DP* and *es gibt DP* as the same.

While the table shows some crosslinguistic differences, a number of similarities stand out. Existential and locative sentences account for the majority of cases in each language. Agentive subjects are absent in my material, regardless of which of the three languages one looks at. I do not have an explanation yet for the fact that English has fewer subject associates than either Dutch or German, but we might venture a guess based on the distribution of direct object associates. For this category, I have also made a comparison of the predicates involved. An overview is presented in Table 3. Among the verbs involved, especially *have* (and its counterparts *haben/hebben*) stands out, but I also want to draw your attention to *find* (*finden/vinden*). Morzycki (2011) also mentions verbs of creation as an important group of verbs, but in my data

Table 2: Types of predicates with subject associate.

predicate	English	%	German	%	Dutch	%
existential	22	85	32	53	24	51
locative	1	4	12	20	3	6
passive	2	8	4	7	14	30
other	1	4	12	20	6	13
total	26	100	60	100	47	100

set verbs like *make* and *create*, while attested, do not stand out in any of the three languages researched.

Table 3: verbs with object associate.

verb	English	%	German	%	Dutch	%
have	26	48	10	28	7	14
find	1	2	6	17	36	12
other	27	50	20	56	38	75
total	54	100	36	100	51	100

There are deep and well-known connections between *have* and existential constructions involving *be* (see Benveniste 1966; Clark 1978; Partee 1999: *inter alii*). Many languages, such as French, build existential sentences with *have*, rather than *be*. If we assume that speakers of English, for whatever reason, pick an existential construction with *have* more often than do speakers of Dutch or German, we might have a possible explanation for the lower number of subject associates in English, and the higher number (both absolute and as a percentage) of occurrences with *have*, thereby linking these two observations. Note that the situation is in fact more complicated, since other verbs may also have an existential function (in the loosest sense of the word). Consider *find*. Instead of *There are many windmills in the Netherlands*, one might equally well write *The Netherlands have many windmills*, or *In the Netherlands, you will find many windmills*. Especially with generic subjects, such as *one* or *you*, *find* has developed an existential use. Of the 13 occurrences of *find*/*finden*/*vinden* in my material, 9 had generic subjects. Yet other ways to build an existential statement involve the use of *with*. *Visit the Netherlands with its many windmills!* *Visit the Netherlands where you will find many windmills.* *Visit the Netherlands where they have/there are many windmills.* Among the 17 English occurrences of associates of galore that are objects of prepositions, there were 9 introduced by *with*. To put this in perspective, the first 17 prepositions of this paper contain just one example of *with*.

The predicates we see most commonly in sentences with *galore* are not compatible with generic definites (or indeed definites in general). This might already explain why sentences such as (34) and (36) above are ungrammatical. They contain verbs of the wrong kind, and if they had verbs of the right kind, the definite generics would have been out of place. Of course, for a fuller account, we need to establish this for all verbs involved, including the ones labelled “other” in tables 2 and 3.

### 2.3 Zero associates

In a few cases, I could not find overt associates for *galore* or *bij de vleet*. Such cases are presumably also possible in German. First, take a look at a Dutch example:

- (43) Een buurt                    waar gekraakt wordt    bij de vleet  
      a neighbourhood were squatted becomes galore  
      ‘A neighbourhood where people squat galore.’

This sentence involves an impersonal passive. The relative clause has no subject or object, nor in fact any nominal constituent that could be the associate of *bij de vleet*. The verb *kraken* “to occupy by squatting” is used intransitively here, and has an implicit argument. My intuition is that *bij de vleet* does not quantify over the number of agents involved in the squatting, but over the number of houses or apartments involved. In fact, (43) would still be true if a single person did all the squatting, e.g. by occupying a different house every week. A similar effect can be noted in English with verbs of emission and ingestion, such as *eat, drink, sweat, pee*, etc., which may have implicit objects.

- (44) Some people were drinking galore, however my pint of PIMMS lasted me the entire night and did me proud! [Internet]

It may be that this kind of case leads to a new usage for *galore/bij de vleet*, as an adverbial intensifier, comparable to *a lot*. The Internet provides cases that can only be viewed in this light, such as the following (from an old journal, *The Music Trade Review*):

- (45) A toast to our Fletcher, who knows how to run  
      All things of departments; he’s bright as the sun!  
      Impartial and just to the boys on the floor,  
      We’re loyal to him, and we like him galore.

Neither *bij de vleet* nor *in Hülle und Fülle* can be used in this way to the best of my knowledge.

## 3 Conclusions

The main conclusions of this paper are as follows:

Jack Hoeksema

- *galore* is not alone: *bij de vleet* in Dutch and *in Hülle und Fülle* in German share its properties, as does English *aplenty*;
- we must distinguish three types of quantification at a distance: floating quantifiers, adverbs of quantification and *galore*-type quantifiers;
- *galore*-type quantifiers have bare nouns as their preferred associates, but pronouns, definite DPs and DPs introduced by *such* are possible as well, provided they have a kind-denoting reading;
- sentences with *galore* typically introduce new discourse referents and predicates are often existential or verbs of encountering;
- *galore*-type quantifiers are not stranded determiners, but PP modifiers (for Dutch and German the PP nature of the quantifiers is self-evident; *galore* itself comes from an Irish/Gaelic PP *go leor* meaning “in sufficiency, enough”).

## References

- Bach, E., E. Jelinek, A. Kratzer & B. Partee. 1995. *Quantification in natural language*. Dordrecht: D. Reidel.
- Barton, Dagmar, Nadine Kolb & Tanja Kupisch. 2015. Definite article use with generic reference in German: an empirical study. *Zeitschrift für Sprachwissenschaft* 34 (2). 147–173.
- Benveniste, Émile. 1966. “Être” et “avoir” dans leurs fonctions linguistiques. In Émile Benveniste (ed.), *Problèmes de linguistique générale, I*, 187–207. Paris: Gallimard.
- Bošković, Željko. 2004. Be careful where you float your quantifiers. *Natural Language & Linguistic Theory* 22 (4). 681–742.
- Carlson, Gregory N. 1977. *Reference to kinds in English*. University of Massachusetts, Amherst PhD thesis.
- Chierchia, Gennaro. 1998. Reference to kinds across languages. *Natural Language Semantics* 6. 339–405.
- Clark, Eve V. 1978. Locationals: existential, locative, and possessive constructions. In J. H. Greenberg (ed.), *Universals of human language*, vol. 4, 85–126. Stanford: Stanford University Press.
- Davies, Mark. 2011. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25. 447–65.
- de Swart, Henriëtte. 1991. *Adverbs of quantification: a generalized quantifier approach*. University of Groningen PhD thesis.
- Doetjes, Jenny. 1997. *Quantifiers and selection. On the distribution of quantifying expressions in French, Dutch and English*. University of Leiden PhD thesis.
- Dowty, David R. & Belinda Brody. 1984. The semantics of “floated” quantifiers in a transformationless grammar. In M. Cobler, S. MacKaye & M. T. Wescoat (eds.), *Proceedings of the 3rd West Coast Conference on Formal Linguistics*, 75–90. Stanford: Stanford Linguistics Association.

- Hoeksema, Jack. 1996. Floating quantifiers, partitives and distributivity. In Jack Hoeksema (ed.), *Partitives*, 57–106. Berlin: Mouton de Gruyter.
- Hoeksema, Jack. 2012. *Bij de vleet: een kwantor op afstand*. *TABU* 40 (3/4). 153–165.
- Krifka, Manfred, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Genaro Chierchia & Godehard Link. 1995. Genericity: an introduction. In Gregory N. Carlson & Francis Jeffrey Pelletier (eds.), *The generic book*, 1–124. Chicago: The University of Chicago Press.
- Lewis, David. 1975. Adverbs of quantification. In Edward L. Keenan (ed.), *Formal semantics of natural language*. Cambridge University Press.
- McCloskey, James. 2000. Quantifier float and wh-movement in an Irish English. *Linguistic Inquiry* 31 (1). 57–84.
- Morzycki, Marcin. 2011. Quantification galore. *Linguistic Inquiry* 42 (4). 671–682.
- Oosterhof, Albert. 2008. *The semantics of generics in Dutch and related languages*. Amsterdam: John Benjamins.
- Partee, Barbara. 1999. Weak NPs in have sentences. In J. Gerbrandy et al. (eds.), *JFAK [a Liber Amicorum for Johan van Benthem on the occasion of his 50th birthday]*. Accessible at: <http://www.illc.uva.nl/j50/>. Amsterdam: University of Amsterdam.
- Sportiche, Dominique. 1988. A theory of floating quantifiers and its corollaries for constituent structure. *Linguistic Inquiry* 19 (3). 425–449.
- van der Auwera, Johan & Sahoo Kalyanamalini. 2015. On comparative concepts and descriptive categories, such as they are. *Acta Linguistica Hafniensia* 47 (2). 136–173.



## Chapter 18

# The interpretation of Dutch direct speech reports by Frisian-Dutch bilinguals

Franziska Köder

University of Oslo

J. W. van der Meer

University of Groningen

Jennifer Spenader

University of Groningen

Frisian and Dutch both have a direct speech reporting construction and an indirect speech reporting construction with verb final word order. Frisian also has an additional indirect speech reporting construction, the embedded verb-second construction, which resembles direct speech in many respects. We investigated whether Frisian-Dutch bilinguals show negative transfer in their interpretation of direct speech in Dutch. We hypothesized that Frisian-Dutch bilinguals would rate an infelicitous embedded V2 construction in Dutch as higher than Dutch monolinguals. Further we hypothesized that when tested on their interpretation of direct speech reports in Dutch, Frisian-Dutch bilinguals would make more errors than their monolingual Dutch counterparts. Our results support both hypotheses.

## 1 Introduction

Consider two different reports about what Jan said in Dutch and Frisian.

- (1) a. Jan<sub>i</sub> zei: “Ik<sub>i</sub> ben ziek.” (Dutch)  
b. Jan<sub>i</sub> sei: “Ik<sub>i</sub> bin siik.” (Frisian)  
‘Jan said: “I am sick.”’
- (2) a. Jan<sub>i</sub> zei dat hij<sub>i</sub> ziek is.” (Dutch)

- b. Jan<sub>i</sub> sei dat hy<sub>i</sub> siik is. (Frisian)  
'Jan said that he is sick.'

Example (1) is a direct speech report, and Jan's words are repeated from his perspective. The pronoun *ik* 'I' therefore refers to the reported speaker, Jan, and not to the actual reporting speaker. Example (2) is an indirect speech report. The reporting speaker presents the content of what Jan said from her current perspective, referring to the absent Jan with the third-person pronoun *hij/hy* 'he'. In Dutch and Frisian, direct speech reports of assertions have verb-second word order (see 1a and 1b) and lack a complementizer. Indirect speech reports include a subordinate clause with verb-final word order and the obligatory complementizer *dat* 'that' (see 2a and 2b). Looking at these two examples alone it seems that there are clear grammatical and semantic markers of direct and indirect speech reports. However, Frisian has an additional indirect speech reporting construction, the embedded V2 construction shown in Example (3):

- (3) a. Jan<sub>i</sub> sei, hy<sub>i</sub> is siik. (Frisian)  
b. \*Jan<sub>i</sub> zei, hij<sub>i</sub> is ziek. (Dutch)  
'Jan said he is sick.'

This embedded V2 reporting construction has verb-second (V2) word order and no complementizer, like direct speech (de Haan 2010; Zwart 1997). However, the pronoun *hy* 'he' in (3) refers to Jan from the perspective of the reporting speaker. This means that with respect to the interpretation of pronouns and other indexicals, the embedded V2 construction in 3 is similar to indirect speech reports like (2).

Thus in Dutch, direct and indirect speech are clearly distinct (Köder 2016). But in Frisian – similar to other Germanic languages like German, and Danish – the embedded V2 construction combines direct speech features (V2 word order, absence of a complementizer) with indirect speech features (interpretation of pronoun), and can therefore be considered a mixed type of report in-between canonical direct and indirect speech (cf. Evans 2013). This means that features that in Dutch unambiguously distinguish direct speech from indirect speech appear in Frisian in an indirect speech reporting construction.

Because almost all Frisian speakers in the Netherlands are Frisian-Dutch bilinguals, we hypothesize that experience with Frisian will cause Frisian-Dutch bilinguals to display difficulties when understanding direct speech reports in Dutch compared to non-Frisian native speakers of Dutch. We experimentally tested this hypothesis and found that Frisian-Dutch speakers show significantly higher error rates in direct speech interpretation than their non-Frisian peers. Further, we found a significant difference in their evaluation of sentences of the three different speech reporting types.

## 2 Dutch and Frisian

Dutch and Frisian are two closely related languages that both have official status in the Netherlands. Most Frisian speakers live in the northern Dutch province of Friesland. Even though Dutch is the dominant language in education, administration and the media, 74% of the population of Friesland are Frisian-Dutch bilinguals, and the majority consider Frisian to be their first language (Hanssen et al. 2015; Province Fryslân, 2015). Due to intense language contact between the two languages, modern Dutch and Frisian exhibit many lexical, grammatical and phonetic similarities (Gooskens & Heeringa 2004; Heeringa & Nerbonne 1999). However, Frisian's larger inventory of constructions for reporting speech is one of the grammatical differences.

In this study, we investigate whether the difference in available constructions for reporting speech and thought in Frisian influence how Frisian-Dutch bilinguals interpret reported speech in Dutch. Previous research on language transfer in bilingual contexts indicates that such effects are possible (Nagy, McClure & Mir 1997; Müller 1998; Muysken 2000), but as far as we know there are no studies on the effect of language transfer on speech reporting.

## 3 Direct speech and indirect speech in Dutch

Previous studies have shown that speakers of Dutch find pronouns in direct speech more difficult to interpret than in indirect speech, due to the required perspective shift from reporting speaker to reported speaker (Köder, Maier & Hendriks 2015). Dutch children up until the age of twelve show difficulties interpreting direct speech reports, often interpreting pronouns in direct speech as if they were in indirect speech (Köder & Maier 2016). One interpretation of these findings is that Dutch children have a less clear-cut direct-indirect distinction than Dutch adults. Similarly, Frisian-Dutch adults might have a less rigid direct-indirect distinction in their Dutch grammar than non-Frisian Dutch adults due to a possible interference from Frisian. If this is correct, Frisian-Dutch bilinguals should make more mistakes than Dutch monolinguals when interpreting pronouns in Dutch direct speech, confusing it with the indirect embedded V2 construction. Furthermore, Frisian-Dutch bilinguals should rate the (ungrammatical) embedded V2 construction in Dutch as more acceptable than non-Frisian speakers of Dutch. To test these hypotheses, we tested Frisian-Dutch bilinguals and Dutch monolinguals with the Speech Report Experiment (Köder, Maier & Hendriks 2015) and asked a subset of the participants to rate the acceptability of speech reports on a questionnaire.

## 4 Method

### 4.1 Participants

36 Frisian-Dutch bilinguals ( $M_{age} = 28.8$ ,  $SD = 13.6$ ) and 115 Dutch monolinguals ( $M_{age} = 24.3$ ,  $SD = 8.6$ ) participated in the Speech Report Experiment. We classified

participants as Frisian-Dutch bilingual if they indicated proficiency in both Frisian and Dutch (among other languages). The group of Frisian-Dutch bilinguals therefore includes both early and late bilinguals with different levels of proficiency in speaking, comprehending and writing Frisian. We will use the term Dutch monolinguals to describe participants who listed only Dutch (among other languages). A subset of the participants (25 Frisian-Dutch bilinguals, 9 Dutch monolinguals) filled in a questionnaire including acceptability judgments.

## 4.2 Procedure

The Speech Report Experiment and the questionnaire were presented online. Instructions for the questionnaire were presented in writing, while instructions for the speech report test were presented auditorily. The task was completed individually and took about 10 minutes in total to complete.

## 4.3 Speech Report Experiment

The Speech Report Experiment is designed as an interactive game called *Who gets the ball?* The experiment consists of short animations that feature three animals (a dog, an elephant and a monkey) interacting with each other. For instance, the elephant walks over to the monkey and whispers into his ear who gets the football (Fig. 1a). Participants heard only an incomprehensible whispering sound. The monkey in turn walks to the dog and reports to him what the elephant has said using either a direct or indirect speech report (Fig. 1b). If the monkey uses for instance the direct speech report *Elephant said: "I get the football"*, the correct referent of the pronoun *I* is the speaker of the reported speech context, i.e., the elephant. In contrast, in an indirect speech report such as *Elephant said that I get the football*, the referent of *I* is the reporting speaker, i.e., the monkey. After each speech report, participants had to click on the animal that they thought got the object (Fig. 1c).

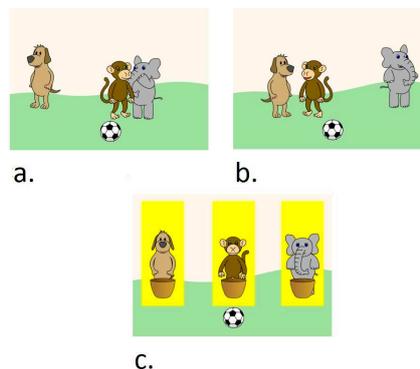


Figure 1: Example screenshots from a test item in the Speech Report Experiment.

The speech reports contain either a first-person (*ik* ‘I’), second-person (*jij* ‘you’) or third-person (*hij* ‘he’) pronoun (see Example (4a)). In total, we presented 30 test items in random order, five for each combination of report type (direct speech, indirect speech) and pronoun (*ik, jij, hij*).

- (4) a. Olifant zei: “Ik/Jij/Hij krijg(t) de voetbal”.  
‘Elephant said: “I/You/He get(s) the football”.’  
b. Olifant zei dat ik/jij/hij de voetbal krijg(t).  
‘Elephant said that I/you/he get(s) the football.’

Recall that direct and indirect speech reports in Dutch are clearly distinct: direct speech reports have verb-second word order in the report; indirect speech sentences have verb-final word order and include the complementizer *dat* ‘that’. In addition, our direct speech stimuli have an 800 ms break between reporting clause and quotation. For more detailed information on the entire procedure please consult Köder and Maier (2016).

#### 4.4 Acceptability judgments

After completing the speech report test, a subset of participants was asked to rate the acceptability of speech reports on a five-point Likert scale (totally agree, agree, neutral, disagree, totally disagree). Participants were presented with an original utterance (e.g. *Jan: I go to the store*) and were then asked to assess whether a particular sentence (e.g., *Jan said: “I go to the store”*) is a correct report of that utterance. We presented one Dutch direct speech report (5), one Dutch indirect speech report with verb-final word order (6) and one ungrammatical Dutch embedded V2 report (7).

- (5) Jan zei: “Ik ga naar de winkel.” (Dutch)  
‘Jan said: “I go to the store.”’  
(6) Bert<sub>i</sub> zei, hij<sub>i</sub> speelt goed voetbal. (Dutch)  
‘Bert<sub>i</sub> said he<sub>i</sub> plays soccer well.’  
(7) Anna zei dat ze niet van vis houdt. (Dutch)  
‘Anna said that she does not like fish.’

Frisian-Dutch participants were given three additional speech reports in Frisian:

- (8) Pyt sei: “Ik bin it paad bjuster.” (Frisian)  
‘Pyt said: “I have lost track.”’  
(9) Abe<sub>i</sub> sei, hy<sub>i</sub> hat in gles pakt. (Frisian)  
‘Abe<sub>i</sub> said he<sub>i</sub> has taken a glass.’  
(10) Froukje sei dat se op sinneskynwaar hoopt. (Frisian)  
‘Froukje said that she hopes for sunny weather.’

## 5 Results

### 5.1 Speech Report Experiment

Figures 2 and 3 show the accuracy of pronoun interpretation for Frisian-Dutch bilinguals and Dutch monolinguals. Figure 2 shows results for direct speech and Figure 3 shows results for indirect speech.

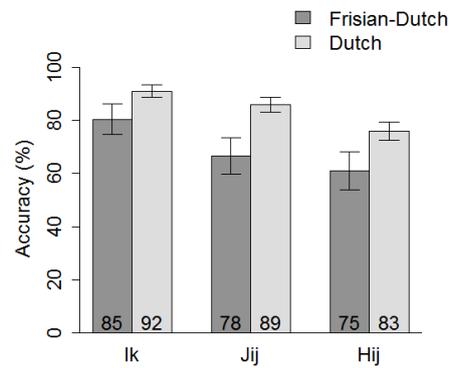


Figure 2: Percentage of correct pronoun interpretations in direct speech of Frisian-Dutch bilinguals and Dutch monolinguals. Error bars indicate 95% confidence intervals.

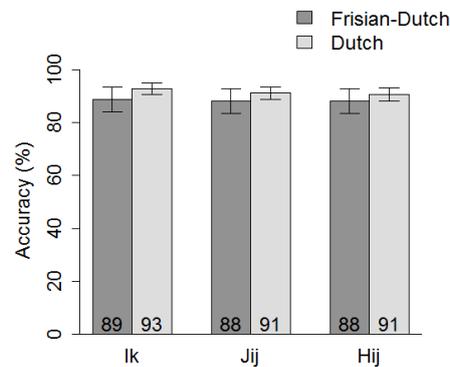


Figure 3: Percentage of correct pronoun interpretations in indirect speech of Frisian-Dutch bilinguals and Dutch monolinguals. Error bars indicate 95% confidence intervals.

We analyzed the accuracy data with mixed-effects logistic regression modeling with the software R (version 3.2.5). Step by step, we added fixed-effects factors and interactions and checked whether they improve the model fit significantly, as indicated

by an Akaike Information Criterion (AIC) decrease of more than 2. The best fitted model includes random intercepts for subjects and a three-way interaction between the fixed-effect factors report type (direct, indirect), pronoun (*ik*, *jij*, *hij*) and language (Frisian-Dutch, Dutch). The index of concordance of the model is 0.91, which indicates that it has predictive power. In direct speech, participants made more mistakes when interpreting the third-person pronoun *hij* than the first-person pronoun *ik* ( $\beta = 1.64$ ,  $z = 7.84$ ,  $p < .001$ ) and the second-person pronoun *jij* ( $\beta = 0.97$ ,  $z = 5.14$ ,  $p < .001$ ). The model already indicates that the language of the participants influences their performance on the Speech Report Experiment. We performed a multiple comparison analysis (Tukey contrasts) to find out in which respects Frisian-Dutch bilinguals differ from Dutch monolinguals in their interpretation of Dutch direct and indirect speech reports.

The results, reported in Table 1, show that both Frisian-Dutch and Dutch speakers made significantly more mistakes in direct than indirect speech. Comparing the performance of Frisian-Dutch bilinguals and Dutch monolinguals, we find that Frisian-Dutch bilinguals made more mistakes than Dutch monolinguals in direct speech, but not in indirect speech interpretation.

Table 1: Multiple comparisons of means (Tukey contrasts).

Linear Hypotheses	Estimate	SE	z-value	p-value
Dutch indirect – Dutch direct = 0	0.94	0.13	7.37	<0.001
Fris.-Dutch indirect – Fris.-Dutch direct = 0	1.65	0.19	8.54	<0.001
Fris.-Dutch indirect – Dutch indirect = 0	-0.56	0.46	-1.31	0.575
Fris.-Dutch direct – Dutch direct = 0	-1.27	0.44	-2.88	0.016

## 5.2 Acceptability judgments

The results of the questionnaire indicate that Frisian-Dutch bilinguals and Dutch monolinguals agreed that direct speech reports ( $M = 4.68$ ,  $SD = 0.77$ ) and indirect verb-final constructions ( $M = 4.74$ ,  $SD = 0.51$ ) are acceptable in Dutch, with no significant differences between the groups. As expected from the literature, all participants judged the Dutch embedded V2 construction mostly as unacceptable ( $M = 1.88$ ,  $SD = 1.14$ ). However, Frisian-Dutch bilinguals tended to rate the Dutch embedded V2 construction as more acceptable than Dutch monolinguals ( $t(31) = 1.92$ ,  $p = 0.06$ ). While none of the Dutch monolinguals found the Dutch embedded V2 construction acceptable ( $M = 1.44$ ,  $SD = 0.53$ ), four Frisian-Dutch bilinguals agreed or fully agreed that it is acceptable ( $M = 2.04$ ,  $SD = 1.27$ ).

The acceptability rating of the Frisian speech reports shows that most Frisian-Dutch bilinguals found the embedded V2 construction in Frisian acceptable ( $M = 2.88$ ,  $SD = 1.72$ ). However, the embedded V2 score is significantly lower than that of Frisian direct speech ( $t(28) = -5.51$ ,  $p < .001$ ) and Frisian verb-final indirect speech

( $t(45) = -3.45, p = .001$ ).

## 6 Discussion and conclusions

Our findings support our first hypothesis that Frisian-Dutch bilinguals make more mistakes in Dutch direct speech interpretation than non-Frisian Dutch speakers. We suggest that this is due to the available embedded V2 construction in Frisian. This construction resembles direct speech by having V2 word order and no complementizer, but requires pronouns to be interpreted from the reporting speaker's perspective as in indirect speech. The fact that Frisian-Dutch participants did not also make significantly more errors than Dutch participants in indirect speech indicates that they are not just generally more confused by speech reports in Dutch, but instead exhibit a more specific interference from the Frisian indirect embedded V2 construction.

Consistent with our second hypothesis, Frisian-Dutch bilinguals also rated the ungrammatical Dutch embedded V2 construction as more acceptable than Dutch monolinguals in the questionnaire. This suggests that Frisian-Dutch participants were unsure if the (ungrammatical) embedded V2 construction was possible in Dutch. We expect similar effects with other bilinguals e.g., German-Dutch, Danish-Dutch bilinguals because these languages have a similar embedded V2 construction.

## 7 Future research

Our questionnaire only used one item for each speech reporting type. A useful follow-up should include different version of multiple items in all three reporting types, and include control items of constructions that are unacceptable. Actually, colloquial Dutch seems to allow an embedded verb-second construction like that in (6) (see Zwart 1997), but our Dutch participants did give this item the lowest rating. We cannot judge whether this rating is evidence of marginal acceptability or actual error unless we also include such items for comparison.

We also only tested Frisian-Dutch speakers in Dutch, but it would also be interesting to investigate how Frisian speakers interpret all three constructions in Frisian. If the similarity between the embedded V2 construction and direct speech is the cause of the errors, then we would also expect to find that Frisians confuse direct and indirect embedded V2 speech reports in Frisian as well. This is a natural topic for further study.

## References

- de Haan, Germen J. 2010. *Studies in West Frisian grammar*. Vol. 161. John Benjamins Publishing.
- Evans, Nicholas et al. 2013. Some problems in the typology of quotation: a canonical approach. *Canonical morphology and syntax*.

- Gooskens, Charlotte & Wilbert Heeringa. 2004. The position of Frisian in the Germanic language area. *On the boundaries of phonology and phonetics*. 61–87.
- Hanssen, Esther, Arjen Versloot, Eric Hoekstra, Arina Banga, Anneke Neijt & Robert Schreuder. 2015. Morphological variation in the speech of Frisian-Dutch bilinguals: (dis)similarity of linking suffixes and plural endings. *Linguistic Approaches to Bilingualism* 5(3). 356–378.
- Heeringa, Wilbert & John Nerbonne. 1999. Change, convergence and divergence among Dutch and Frisian. *Philologia Frisica Anno*. 88–109.
- Köder, Franziska. 2016. *Between direct and indirect speech: the acquisition of pronouns in reported speech*. University of Groningen PhD thesis.
- Köder, Franziska & Emar Maier. 2016. Children mix direct and indirect speech: evidence from pronoun comprehension. *Journal of Child Language* 43. 843–866.
- Köder, Franziska, Emar Maier & Petra Hendriks. 2015. Perspective shift increases processing effort of pronouns: a comparison between direct and indirect speech. *Language, Cognition and Neuroscience* 30(8). 940–946.
- Müller, Natascha. 1998. Transfer in bilingual first language acquisition. *Bilingualism: Language and cognition* 1(03). 151–171.
- Muysken, Pieter. 2000. *Bilingual speech: a typology of code-mixing*. Vol. 11. Cambridge University Press.
- Nagy, William E., Erica F. McClure & Montserrat Mir. 1997. Linguistic transfer and the use of context by Spanish-English bilinguals. *Applied Psycholinguistics* 18(04). 431–452.
- Zwart, C. Jan-Wouter. 1997. *Morphosyntax of verb movement: a minimalist approach to the syntax of Dutch*. Vol. 39. Springer Science & Business Media.



## Chapter 19

# Mining for parsing failures

Daniël de Kok

University of Groningen

Gertjan van Noord

University of Groningen

Error mining is a technique to support the engineering of natural language parsers, by analysing parser output for a large set of inputs. It produces a list of properties (such as words or word sequences) of inputs which systematically lead to parsing failure. These properties typically point at omissions and mistakes in the grammar or the lexicon. Error mining can be applied to improve general purpose parsers, but is particularly suited to adapt parsers for novel text genres and topic domains.

In this article, a new error mining method is described, generalizing and extending earlier proposals by van Noord (2004), Sagot & de la Clergerie (2006) and de Kok, Ma & van Noord (2009). The new method improves the extraction of longer word sequences as features for the miner in comparison with the method of van Noord (2004), integrating the computation of suspicion as proposed by Sagot & de la Clergerie (2006). The extension allows the possibility to mine with character sequences as opposed to word sequences.

The new error miner is evaluated both quantitatively and qualitatively, and is shown to perform better than its predecessors.

## 1 Introduction

An important aspect of the engineering of natural language processing applications involves quality control. It is important to know for which types of input the result of the parser is not reliable. If we adapt the parser, it is important to check that the parser does not make new mistakes. And if we want to apply the parser for a novel text genre or topic domain, it is important to see what problems the parser faces in new contexts.

The importance of quality control has been recognized early on, and one of the first initiatives to do this in a systematic way has been pioneered by John Nerbonne and his colleagues (Nerbonne et al. 1993). They manually built a large catalogue of

example sentences displaying the relevant syntactic properties of German, with the explicit purpose of testing NLP systems. Treebanks constitute obvious resources for quality control but likewise involve manual labour.

A technique which does not rely on manual construction of example sentences or manual annotation of corpus sentences is *error mining*. The goal of error mining is to discover properties of input sentences which systematically cause parsing failure. The parser is applied to a large amount of sentences. For each sentence, we record (or estimate) whether the parse was successful. The error miner then lists properties which occur frequently in sentences that cannot be parsed, and which hardly if ever occur in sentences that were unproblematic.

Error mining thus requires that we know whether a parse for a given input is successful. Obviously, for previously unseen sentences we do not know the quality of the output of the parser. Instead, therefore, for hand written grammars and lexicons, we use a weaker notion of success. A parse is deemed successful if the parser is able to construct a parse tree dominating all words of the sentence. In other situations, there may be alternative possibilities to estimate parse success. If the parser produces a confidence score as part of the output, we may want to use that score as an indication of parse success. In the formalization below, we assume that parse success is not necessarily a binary distinction, but a score between 0 and 1.

The result of error mining consists of superficial properties of input sentences which ‘cause’ parsing failure. For instance, one such property could be the bigram “why .”. This indicates that sentences in which the word *why* is immediately followed by a full stop are typically not parsed correctly. The parser engineer will immediately realize (perhaps after further inspection of all sentences which contain this bigram) that verbs which introduce indirect questions should also be able to combine with bare WH-phrases, as in *They don’t know why*.

In this paper we review the earlier approaches to error mining by van Noord (2004) and Sagot & de la Clergerie (2006) (section 2). Extending work we described earlier (de Kok, Ma & van Noord 2009), we present a generalized error miner which combines the strength of these two former proposals (section 3).

In section 4 we discuss a simple evaluation method based on precision and recall. We apply this method to the error miners. The comparison shows that the generalized error miner out-performs the former proposals.

Error mining has been used for several parsing systems and several languages, including English, Dutch, French, German and Spanish. Error mining has also been applied to improve natural language generation (Gardent & Narayan 2012; Narayan & Gardent 2012).

## 2 Previous work

Two promising error mining techniques have been proposed. van Noord (2004) uses word n-grams of arbitrary length as its features and implements a simple, frequency-based suspicion scoring method. Sagot & de la Clergerie (2006) proposes a more sophisticated iterative suspicion scoring method. However, that method provides

only rudimentary feature extraction, since it only considers words and word pairs.

Our contribution in this paper is to combine an improved version of the more general feature extraction method of van Noord (2004) with the successful suspicion computation of Sagot & de la Clergerie (2006).

In the following description we use an error function  $\text{error}(s)$ , which is zero if the sentence  $s$  was parsable or one if it was not parsable. The suspicion of the feature  $f_i$ ,  $\text{susp}(f_i)$ , is the mean error score of the sentences in which the feature  $f_i$  occurs. Here,  $S$  is the bag of sentences, and  $f_i(s)$  is the number of occurrences of feature  $f_i$  in sentence  $s$ .

$$\text{susp}(f_i) = \frac{\sum_{s \in S} f_i(s) \cdot \text{error}(s)}{\sum_{s \in S} f_i(s)} \quad (1)$$

If word n-grams are used as features, a potential problem arises. If a particular word sequence  $w_i \dots w_j$  has a high suspicion, then all longer word sequences which contain  $w_i \dots w_j$  will necessarily have a high suspicion too. This is undesirable. Therefore, if n-grams are included as features in the error miner, a selection criterion is required. van Noord (2004) proposes to add a longer n-gram only if its suspicion is higher than all of the n-grams it contains:

$$\text{susp}(w_h \dots w_i \dots w_j \dots w_k) > \text{susp}(w_i \dots w_j) \quad (2)$$

As a result, there usually is only a small number of long n-grams which the error miner needs to take into account. This also implies that there is no need to set a value for  $n$  a priori; instead, the data determines which longer n-grams are required. As a further heuristic, if a longer n-gram satisfies the selection criterion, then the corresponding shorter n-grams are no longer used as features for the error miner.

The error mining method described by Sagot & de la Clergerie (2006) addresses an issue in the miner of van Noord (2004) where features that co-occur frequently with problematic features also get a high suspicion. It does so by taking the following characteristics of suspicious features into account during feature selection: if a feature occurs in parsable sentences, it becomes less likely that it is the cause of a parsing error; the suspicion of a feature depends on the suspicions of other features in the sentences where it occurs; a feature observed in a shorter sentence is initially more suspicious than a feature observed in a longer sentence.

The method introduces the notion of *observation suspicion*,  $\text{susp}(f_i(s))$  which is the suspicion of feature  $f_i$  in sentence  $s$ . The (global) suspicion of a feature is the average of all observation suspicions,

The observation suspicions are dependent on the feature suspicions, making the method iterative. The observation suspicion is defined as the feature suspicion, normalized by suspicions of the other features that are active within the same sentence:

$$\text{susp}^{n+1}(f_i(s)) = \frac{\text{error}(s) \cdot \text{susp}^{n+1}(f_i)}{\sum_{f_j \in F(s)} \text{susp}^{n+1}(f_j)} \quad (3)$$

Here,  $F(s)$  is the set of features that are active (have a non-zero frequency) in sentence  $s$ . Since the mining procedure is iterative, the suspicion of a feature is redefined to depend on the observation suspicions of the previous iteration:

$$\text{susp}^{n+1}(f_i) = \frac{\sum_{s \in S} \text{susp}^n(f_i(s))}{\sum_{s \in S} f_i(s)} \quad (4)$$

Given the recursive dependence between feature suspicions and observation suspicions, starting and stopping conditions are defined for iterative mining. The observation suspicions are initialized by uniformly distributing suspicion over the features that are observed in a sentence:

$$\text{susp}^0(f_i(s)) = \frac{\text{error}(s) \cdot f_i(s)}{\sum_{f_j \in F(s)} f_j(s)} \quad (5)$$

Mining is stopped when the process reaches a fixed point where the feature suspicions have stabilized.

### 3 n-gram expansion

While the iterative miner described by Sagot & de la Clergerie (2006) only mines on features that are word unigrams and bigrams, our experience with the miner described by van Noord (2004) has shown that including n-grams that are longer than bigrams as features during mining can capture many additional phenomena.

We propose a feature extraction method that adds and expands n-grams when it is deemed useful. This method iterates through a sentence unigram by unigram and expands unigrams to longer n-grams when there is sufficient evidence that the expansion will be useful. We then combine this feature extractor with the selection method of Sagot & de la Clergerie (2006). Within this extractor, we use the definition of suspicion given as Equation 1, except that we do not use the frequency  $f_i(s)$  directly, but use a binary value which indicates if the form  $f_i$  occurs in the sentence  $s$ .

The expansion method is based on the following observation: if we consider the word bigram  $w_1, w_2$ , either one of the unigrams or the bigram can be problematic. If one of the unigrams is problematic, the bigram will inherit suspicion of the problematic unigram. If the bigram is problematic, the bigram will have a higher suspicion than both of its unigrams. Consequently, we will want to expand the unigram  $w_1$  to the bigram  $w_1, w_2$  if the bigram is more suspicious than both of its unigrams. If the bigram is just as suspicious as one of its unigrams, such an expansion is not necessary, since we want to point to the cause of the parsing error as exactly as possible.

The same procedure is applied to longer n-grams. For instance, the expansion of the bigram  $w_1, w_2$  to the trigram  $w_1, w_2, w_3$  is only permitted if  $w_1, w_2, w_3$  is more suspicious than its bigrams. Given that the suspicion of  $w_3$  aggregates to  $w_2, w_3$ , we account for  $w_3$  and  $w_2, w_3$  simultaneously in this comparison.

The general criterion is that the expansion to an n-gram  $i \dots j$  is permitted when  $\text{susp}(i \dots j) > \text{susp}(i \dots j - 1)$  and  $\text{susp}(i \dots j) > \text{susp}(i + 1 \dots j)$ .

This method differs from that of van Noord (2004) in that the method of van Noord (2004) considers all n-grams in a sentence, while our method does not consider  $w_i \dots w_j \dots w_k$  if the expansion to  $w_i \dots w_j$  failed.

In Table 1, we illustrate our method to expand the n-gram feature *voor* to *voor uur van de* in the following sentence:

- (6) De Disney-topman staat voor uur van de waarheid.  
 the Disney top executive stands before hour of the truth.  
 ‘The moment of truth has come for the Disney top executive.’

The counts in this example are based on real data.

Table 1: Expansion of the feature *voor* to *voor uur van*.

Expansion	$\text{susp}(i \dots j)$	$\text{susp}(i \dots j - 1)$	$\text{susp}(i + 1 \dots j)$	Expand
<i>voor</i> → <i>voor uur</i>	$\frac{48}{50}$	$\frac{778949}{9590152}$	$\frac{116975}{1563498}$	yes
→ <i>voor uur van</i>	$\frac{40}{40}$	$\frac{48}{50}$	$\frac{856}{9779}$	yes
→ <i>voor uur van de</i>	$\frac{30}{30}$	$\frac{40}{40}$	$\frac{297}{3748}$	no

While this expansion method looked promising in our initial experiments, we found it to be too eager. This eagerness is caused by sparsity in the data. Since longer n-grams are less frequent, they also tend to be more suspicious if they occur in unparseable sentences. The expansion criterion does not take data sparseness into account.

We introduce an expansion factor to handle sparseness. This factor depends on the frequency of an n-gram in the set of unparseable sentences, and asymptotically approaches one for higher frequencies. As a result the burden of proof is inflicted on the expansion: the longer n-gram needs to be relatively frequent or much more suspicious than its (n-1)-grams. The expansion criteria are changed to  $\text{susp}(i \dots j) > \text{susp}(i \dots j - 1) \cdot \text{extFactor}(i \dots j)$  and  $\text{susp}(i \dots j) > \text{susp}(i + 1 \dots j) \cdot \text{extFactor}(i \dots j)$ , where

$$\text{extFactor}(i \dots j) = 1 + \exp(-\alpha \sum_{s \in S} \text{error}(s) \cdot f_{i \dots j}) \quad (7)$$

As we show in section 4.5,  $\alpha = 0.01$  proved to be a good setting.

After error mining, we can extract a list of forms, ordered by suspicion. However, normally we are interested in forms that are both suspicious and frequent.

## 4 Evaluation

### 4.1 Methodology

In the early papers on error mining, error mining methods have been evaluated primarily manually. Both van Noord (2004) and Sagot & de la Clergerie (2006) conduct a qualitative analysis of highly suspicious features. But once one starts experimenting with various extensions, such as n-gram expansion and expansion factor functions, it is difficult to gauge the contribution of modifications through a small-scale qualitative analysis.

To be able to evaluate changes to the error miner, we have supplemented qualitative analysis with a automatic quantitative evaluation method. Such a method should judge an error miner in line with the interests of a grammar engineer:

- an error miner should return features that point to problems that occur in a large number of sentences;
- the features that are returned by the error miner should be as exact as possible in pointing to the problem.

The first requirement is relatively easy to test – the error miner should return features that occur in a relatively large number of unparsable sentences. It is less clear how the second requirement should be tested, since it requires that a human checks the features to be relevant. However, if we assume that the quality of features correlates strongly to their discriminative power, then we would expect a miner to return features that only occur in unparsable sentences. These characteristics can be measured using the recall and precision metrics from information retrieval:

$$R = \frac{|\{S_{unparsable}\} \cap \{S_{retrieved}\}|}{|\{S_{unparsable}\}|} \quad (8)$$

$$P = \frac{|\{S_{unparsable}\} \cap \{S_{retrieved}\}|}{|\{S_{retrieved}\}|} \quad (9)$$

Consequently, we can also calculate the f-score (we use  $\beta = 0.5$ ).

### 4.2 Material

In our experiments, we use the Flemish Mediargus newspaper corpus. This corpus consists of 67 million sentences (1.1 billion words). For 9.2% of the sentences no full analysis could be found. Flemish is a variation of Dutch written and spoken in Belgium, with a grammar and lexicon that deviates slightly from standard Dutch. Previously, the Alpino grammar and lexicon were never systematically modified for parsing Flemish.

We now proceed to discuss the results of the quantitative and qualitative evaluation. We will first compare the miners described in van Noord (2004) and Sagot & de la Clergerie (2006). Then, we examine the performance of the expansion method

that we proposed and compare it to the competition. Finally, we will conclude this section with an qualitative evaluation of iterative error mining with our expansion method.

### 4.3 Scoring function

After error mining, we can extract a list of forms, ordered by suspicion. However, normally we are interested in forms that are both suspicious and frequent.

$$\text{delta}(f_i) = \sum_{s \in S, \text{error}(s) > 0} f_i(s) - \sum_{s \in S, \text{error}(s) = 0} f_i(s) \quad (10)$$

$$\text{score}(f_i) = \text{susp}(f_i) \cdot \text{delta}(f_i) \quad (11)$$

$$\text{score}(f_i) = \begin{cases} 0 & \text{if } \text{delta}(f_i) \leq 0 \\ \text{susp}(f_i) \cdot (1 + \ln(\text{delta}(f_i))) & \text{if } \text{delta}(f_i) > 0 \end{cases} \quad (12)$$

We use the scoring function that performed best for a specific error miner. In the case of the iterative miner of Sagot & de la Clergerie (2006) and the miner proposed in this work, the scoring function in equation 11 is used in the experiments below. For the miner of van Noord (2006), the scoring function in equation 12 is used below.

### 4.4 Iterative error mining

Our first interest is if, and how much iterative error mining outperforms error mining with suspicion as a ratio. To test this, we compared the method described by van Noord (2004) and the iterative error miner of Sagot & de la Clergerie (2006). For the iterative error miner we included all bigrams and unigrams without further selection. The left graph in Figure 1 shows the F-scores for these miners after  $N$  retrieved features.

The iterative miner of Sagot & de la Clergerie (2006) clearly outperforms the miner of van Noord (2004), despite the fact that the latter has a more sophisticated feature extraction method. That this happens is understandable – suppose that 60% of the occurrences of a frequent feature is in unparsable sentences. In such a case, the ratio-based miner would assign a suspicion of 0.6. But, since the feature is relatively frequent, it would still be ranked very high, even though there is plenty of evidence that it is not responsible for parsing errors. This also manifests itself in the performance of scoring functions – the ratio-based miner was the only miner to perform best with the scoring function in equation 12. This indicates that relying too much on frequencies is dangerous in ratio-based mining. However, relying purely on suspicion would return many forms with a low frequency.

Another interesting characteristic of these results is that the performance of the error miners seems to fit a logarithmic function. This is not surprising, since it shows that there are some very frequent patterns indicating errors and a long tail of less frequent patterns indicating errors. The fact that there is a long tail of infrequent patterns does not make the task of the parser engineer hopeless. In fact, a single error in the parser will often surface in multiple patterns. As an example, consider the Dutch determiner *zo'n* ('such'). In standard Dutch, this determiner combines with a singular noun. In Flemish, the determiner can combine with plural nouns as well. That usage of the determiner *zo'n* gave rise to parsing errors. This particular error shows up in many patterns which occur high in the list of relevant features: *zo'n momenten*; *zo'n mensen*; *Op zo'n momenten*; *zo'n omstandigheden*; *zo'n zaken...* Generalizing patterns to include part-of-speech tags would be useful here.

#### 4.5 n-gram expansion

In our second experiment, we compare the performance of iterative mining on uni- and bigrams with an iterative miner that uses the n-gram expansion algorithm that was described in section 3. We use  $\alpha = 0.01$ , which provides good performance across the board. In the second graph of Figure 1, we compare our miner that uses word n-gram expansion with the miner of Sagot & de la Clergerie (2006). We observe that our method for the inclusion of longer n-grams is beneficial to error mining.

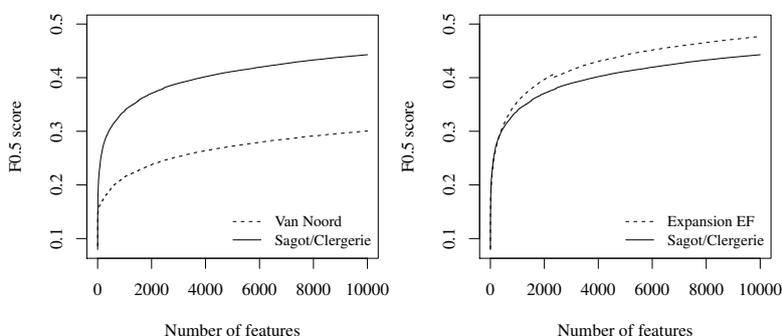


Figure 1: Left:  $F_{0.5}$ -scores after retrieving  $N$  features for ratio-based mining versus iterative mining with unigrams and bigrams. Right:  $F_{0.5}$ -scores after retrieving  $N$  features for the miner of Sagot and de la Clergerie (2006) and the miner proposed here

#### 4.6 Further analysis

We found many interesting longer n-grams in the results of the miner proposed in this article that could not have been captured by the miner of Sagot & de la Clergerie

(2006). If we inspect the most relevant 50 items, using the relevance score function given in equation 11, we do not find any unigrams, 16 bigrams, and 34 longer N-grams.

One very instructive example from this list of 50 most relevant items is the trigram *Het zijn zij* ('It are they'). This example illustrates a convincing case where a bigram would not suffice. Each of the three words of the pattern, *het*, *zijn* and *zij*, are amongst the most frequent words in Dutch. Also, the bigrams *Het zijn* and *zijn zij* are very frequent, and not suspicious at all: the first bigram occurs 56947 times, including 6937 parsing failures; the second bigram occurs 5588 times (696 parsing failures). However, the trigram is very suspicious: it occurs 220 times, leading to parsing failure in 212 cases. The trigram is found in Flemish cleft sentences such as *Het zijn zij die de fouten maken* ('It is them who make the mistakes').

The pattern surfaces because in standard Dutch, in contrast to Flemish, the sentence would be phrased as *Zij zijn het die de fouten maken*.

## 5 Conclusions

We combined iterative error mining with a new feature extractor that includes n-grams of an arbitrary length, taking care that n-grams are long enough to capture interesting phenomena, but not longer. We dealt with the problem of data sparseness by introducing an expansion factor that softens when the expanded feature is very frequent.

In addition to the generalization of iterative error mining, we have introduced a method for automatic evaluation that is based on the precision and recall scores commonly used in information retrieval. This allows us to test modifications to error minings quickly without going through the tedious task of ranking and judging the results manually.

Using this automatic evaluation method, we have shown that iterative error mining improves upon ratio-based error mining. We have also shown that the use of a smart feature extraction method improves error miners substantially. The inclusion of longer n-grams captures many interesting problems that could not be found if a miner restricted itself to words and word bigrams.

## References

- de Kok, Daniël, Jianqiang Ma & Gertjan van Noord. 2009. A generalized method for iterative error mining in parsing results. In *GEAF*.
- Gardent, Claire & Shashi Narayan. 2012. Error mining on dependency trees. In *ACL 2012*. <http://dl.acm.org/citation.cfm?id=2390524.2390607>.
- Narayan, Shashi & Claire Gardent. 2012. Error mining with suspicion trees: seeing the forest for the trees. In *COLING 2012*.
- Nerbonne, John, Klaus Netter, Abdel Kader Diagne, Judith Klein & Ludwig Dickmann. 1993. A diagnostic tool for German syntax. *Machine Translation* 8(1). 85–107.

*Daniël de Kok & Gertjan van Noord*

- Sagot, Benoît & Éric de la Clergerie. 2006. Error mining in parsing results. In *ACL-44: proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 329–336. Sydney, Australia: Association for Computational Linguistics.
- van Noord, Gertjan. 2004. Error mining for wide-coverage grammar engineering. In *ACL*, 446. Barcelona, Spain: Association for Computational Linguistics.
- van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In *TALN 2006 Verbum Ex Machina, actes de la 13<sup>e</sup> Conference sur le Traitement Automatique des Langues naturelles*, 20–42. Leuven.

## Chapter 20

# Looking for meaning in names

Stasinou Konstantopoulos

National Centre of Scientific Research “Demokritos”

This article discusses the concept that proper names are not semantically empty denotations, but characterize in many, often subliminal, ways their denotee. The discussion is driven by computational experiments on using just a name to guess the linguistic and cultural background of a person and the positive or negative polarity of a fictional character’s role.

### 1 Introduction

This article discusses the concept that proper names are not semantically empty denotations, but they characterize, sometimes subliminally, their denotee. To give a first, straightforward example, the reader will intuitively agree that one can guess (or at least considerably narrow down) a person’s linguistic and cultural background from that person’s name alone. But there might also be less intuitive correlations between names and their bearers’ properties.

To qualify such a broad and general claim into a more concrete object of investigation, we will narrow its scope to correlations between how a name ‘sounds’, its phonetic properties, and properties of the name’s bearer. In our language guessing example, we expect that guesses are based to large extent on knowledge of a language’s orthographic conventions, characteristic morphological markers, and characteristic lexical units. The question we discuss here is if *phonetic* characteristics are also relevant to our intuition about where somebody comes from, and also what other properties can be intuited besides origin.

However, in the case of actual people’s names it is hard to imagine that they would be associated with anything but pleasant and positive meanings, if anything at all. But if we turn our attention to *fictional* characters’ names, we expect their creators to have (consciously or not) named them to sound like they behave, so that there will be both positively and negatively sounding names. Ponder, for example, whether it would sound natural if *Hannibal Lecter* was a positive character and *Frodo* was a negative one. If it wouldn’t, and we had never heard of these names before, that would

imply a prior on what a positive or negative character should sound like, essentially a semantics that is linked to the sound of the name.

## 2 Names and computational methods

In fiction and art names serve a more complex functional role than denotation alone and should be treated differently than other proper names (Markey 1982; Nicolaisen 2008). Since they are chosen or invented to satisfy a different role, creators apply different criteria in selecting a name for their characters than the conventions or aesthetics used to select people's names. One of these criteria is the intuitions and preconceptions about the character that the name alone implies to the audience (Rudnyckij 1959; Algeo 2010). In fact, Ashley (2003) suggests that a literary name must be treated as a 'small poem' with all the wealth of information that such a statement implies.

This more general observations can sometimes be framed into specific intuitions and preconceptions in the context of each given work. Chen (2008), for instance, argues that the ethnic-marked names created by Carl Barks for the *Uncle Scrooge* comic books in the 1950s and 1960s contributed to the books' success by feeding into the isolationist feelings of post-war US. However, there is legitimate concern regarding the validity of generalizations made by studying individual creative works (Butler 2013).

This creates an opportunity for computational methods that can extract patterns from larger bodies of literary work than what is manually feasible. Naturally, more relevant to us are methods that concern the meaning of isolated words rather than the grammatical structure that combines them into text. Coming back to guessing linguistic and cultural background from a name, the relevant literature originates in *speech synthesis* where *language identification* is used to select different pronunciation models for foreign names. Starting from hand-crafted rules (Spiegel 1985), the speech synthesis community has since moved to machine-learned language identification models (Font Llitjós & Black 2001). The same general methodology has been applied to automatic transliteration of named entities for the purposes of *multi-lingual information extraction* (Virga & Khudanpur 2003) and *machine translation* (Huang 2005).

In these applications, the language identification pre-processing was reported to improve accuracy on the overall task, but the accuracy of the language identification step itself was not reported. The first results on how indicative a name is of its linguistic background comes from experiments on applying *n*-gram modelling to a corpus of European names and nationalities, compiled by harvesting from the Web information about football players and their national squad eligibility (Konstantopoulos 2007). This work and subsequent analyses of the same data (Konstantopoulos 2010; Florou & Konstantopoulos 2011) aimed at comparing language identification of people's surnames versus common words and identifying the features that make the former more characteristic of their linguistic background than the latter. But they have also yielded lateral results more directly related this discussion, and specifically the

analysis of the roots and of the derivational morphology used in surname formation.

### 3 Word form and meaning

In another strain of related work, form-meaning systematicity has been investigated in the context of exploring the idea that the mental lexicon would be maximally easy to organize if there were a transparent, structure-preserving relationship where words sound similar to the extent that they mean similar things (Shillcock et al. 2001). This idea has been tested in English (Shillcock et al. 2001; Monaghan, Shillcock & Christiansen 2014; Gutiérrez, Levy & Bergen 2016) and Spanish (Tamariz 2008), typically restricted to mono-morphemic words to avoid confusing simultaneous form/meaning similarity due to derivation for form-meaning systematicity.

This body of work provides significant evidence of form-meaning systematicity. It should be noted, however, that this work assumes a *distributional semantics* to interpret ‘meaning’ and is, accordingly, evaluated on the task of using form features to predict what other words the given word co-occurs with. As Gutiérrez, Levy & Bergen (2016: Section 6) also note, these experiments tell us very little about human intuition regarding the meaning of an unknown word heard outside of any context.

Naturally, such questions offer themselves to psycholinguistics research such as the experiments conducted by Ramachandran & Edward (2001) to identify cross-lingual and cross-cultural correlations between nonce words and shapes. But, and remaining on the more familiar ground of computational linguistics, they also offer themselves to computational experiments on predicting the meaning of words from their form, including non-distributional interpretations of ‘meaning’.

### 4 Names of fictional characters

This brings us back to names in fiction: the names of fictional characters reflect (possibly subconsciously) a perception shared between the creator and the audience of what a character’s name ought to sound like. The personal preferences or experiences of the creator might add noise, but given a large enough corpus fictional characters’ names can uncover analogies and familiarities within a given linguistic and cultural background.

Motivated by this idea, Papantoniou & Konstantopoulos (2016) created a corpus of names of fictional characters in motion pictures, annotated with the *polarity* of their role in the plot. Eight annotators identified 1102 positive and 434 negative characters in 202 movies. The annotation guidelines stressed that only clear-cut cases should be annotated and the overall setup made it easy for the annotators to avoid making a commitment and move on to the next character or movie, yielding a high degree of agreement (Table 1).

Obviously, the fact that the annotators made their decision by looking at the cast credits cannot be construed as anything deeper than their knowledge of the movie’s plot. That is to say, regardless of whether the character’s name was as common

as *Jane Smith* or as unique as *Darth Vader*, it was already familiar and the results of the annotation task are *not* meant to be used to extract any conclusions about predicting polarity from nonce word forms. They can, however, be used to create computational models that predict polarity where we can experiment by controlling what background knowledge we make available to the machine learning algorithm and observing the resulting prediction accuracy. By identifying the background features that are the best predictors, we can become informed about the inventory of phonological characteristics, semantic and pragmatic analogies, and other devices that creators use to share their perception of the character with their audience.

As noted earlier, experimenting with very difficult language identification tasks (Florou & Konstantopoulos 2011: Nordic surnames) has shown that even closely related, and otherwise difficult to separate backgrounds (e.g., Norwegian and Danish) exhibit different patterns regarding what derivational morphology is applied to roots from different semantic classes (such as in occupational, locative, or patronymic surnames). In the Nordic surname experiments,  $n$ -gram modelling was unable to identify such fine patterns which were only discovered by human observation and verified by encoding them as a DCG that was evaluated on the corpus.

This led Papantoniou & Konstantopoulos (2016) to define more sophisticated features that encode prior theoretical work as well as more informed features that incorporate lexical knowledge:

- the literary analysis of poems is a natural place to look for theoretical insights regarding how words sound. Swooshing past the staggering volume of relevant work through the centuries, we assume the framework developed by Kaplan & Blei (2007) for the computational analysis of the phonemic, orthographic, and syntactic features of English-language poems. Of these features, *alliteration* (as in *Peter Pan*), *consonance* (*Freddy Krueger*), and *assonance* (*Frodo*) can be directly applied to names outside any context. Such features encode dependencies longer than what can be discovered by  $n$ -gram modelling, so that the prior knowledge that alliteration, consonance, and assonance might be relevant needs to be encoded in pre-computed features.
- An increasing volume of work investigates *phonological iconicity*, the existence of non-arbitrary relations between phonetic representation and semantics. The

Table 1: Inter-annotator agreement on the role polarity task.

Measure	Value
Percentage Agreement	0.963
Hubert Kappa Agreement	0.980
Fleiss Kappa Agreement	0.973
Krippendorff Alpha Agreement	0.979

findings are often based on individual works and not generalizable, for example Miall (2001) notes that passages about Hell from Milton's *Paradise Lost* contain significantly more front vowels and hard consonants than passages about Eden, which contain more medium back vowels. When some generalizations have been made, these can be contradictory. Auracher et al. (2011), for instance, found that across different languages (including remote ones), nasals relate to sadness and plosives to happiness, parallels across remote languages, which might be consistent with the earlier finding that sonorants (including nasal /m/) is more common in tender poems (Fonagy 1961) but contradicts another previous finding that plosives correlate with unpleasant words (Whissell 1999). Clearly the relevant discussion in literature and psychology is far from mature, but there is growing evidence that phonological iconicity is a real phenomenon worth investigating in the context of names.

- Soundex phonetic distance and Levenshtein lexicographic distance to positive or negative terms in SentiWordNet (Esuli & Sebastiani 2006), a linguistic resource for *sentiment analysis* that annotates WordNet terms with an estimated degree of positive, negative or neutral hue. This makes character polarity prediction aware of the negative sentiment in *Darth Vader* through its similarity to the term *dark*.
- Socio-linguistic pragmatics, such as familiarity and gender. How familiar a name sounds is estimated via the frequency of its appearance in the *Social Security Death Index (SSDI)*, the publicly available database of all deceased US citizens since 1936. First names were matched against gender by scraping male and female first names from the multitude of Web sites that list baby names for prospective parents. This gives character polarity prediction access to the information that *Jane Smith* is a common female name and *Hannibal Lecter* is a rare male name.

Papantoniou & Konstantopoulos (2016) used these features to learn from their manually annotated corpus a decision tree that predicts character polarity (Table 2). Movie metadata such as genre and crediting order were expected to be very good discriminants, and were also included to the feature set for comparison only.

By comparing the performance of all features (F = 82%), only metadata (F = 71%), and all name-intrinsic features (excluding metadata, F = 80%), we can immediately understand that name-intrinsic features are better discriminants than metadata. This validates the core hypothesis that there is a correlation between what fictional character names look and sound like and the role they play in the plot of the fictional work they appear in. And among all intrinsic features, the phonetic ones appear to be the best discriminants. In fact, removing any other feature category *increases* performance, leading us to believe that all other features are actually adding noise (rather than discriminatory power) to the feature space.

Table 2: Performance of polarity prediction for different feature settings.

	Recall	Precision	$F_{\beta=1}$ score
Without IMDB metadata	80%	80%	80%
Only metadata	73%	70%	71%
Only phonetic features	79%	79%	79%
Without poetic features	84%	83%	83%
Without consonance feature	82%	82%	82%
Without SentiWordNet features	81%	81%	81%
Without phonetic features	80%	79%	79%
Without social features	81%	80%	80%
All features	82%	82%	82%

## 5 Concluding remarks

An interesting result was that the ‘unfamiliar sounding’ feature is not discriminative, refuting the hypothesis that the concept of the ‘other’ is stereotyped negatively. A more thorough investigation (and, in fact, one that is more inline with prior theories) will refine the ‘unfamiliar’ class into different ethnic backgrounds. Although not directly targeting any linguistic conclusions, from the wider humanities perspective such an investigation could give a tool for exploring whether ‘bad guy’ names in major US productions follow political developments to shift from German-sounding to Slavic-sounding to Arabic-sounding.

The result that was most relevant to language was the discriminative power of phonetic features. Although the current level of theoretical understanding of iconicity and its underlying mechanisms is far from complete, it helped formulate features and verify that they are discriminative. On the computational linguistics front, the findings presented here are also too focused on a particular language and domain to be a sound basis for grand generalizations, but they do point to various interesting directions. It would, for example, be interesting to extend the experiments to written literature to observe if there are differences between spoken names (as in films) and names that are only meant to be read (as in literature). In addition, using written literature will allow pushing earlier than the relatively young age of motion pictures.

## Acknowledgements

I would like to gratefully acknowledge having used in this paper the results and insights obtained from the excellent work on language identification by Eirini Florou and on polarity prediction by Katerina Papantoniou for their MSc theses.

## References

- Algeo, John. 2010. Is a theory of names possible? *Names* 58(2). 90–96. <http://dx.doi.org/10.1179/002777310X12682237915106>.
- Ashley, Leonard R. N. 2003. *Names in literature*. B: Authorhouse.
- Auracher, Jan, Sabine Albers, Yuhui Zhai, Gulnara Gareeva & Tetyana Stavniychuk. 2011. P is for happiness, N is for sadness: universals in sound iconicity to detect emotions in poetry. *Discourse Processes* 48(1). 1–25. <http://dx.doi.org/10.1080/01638531003674894>.
- Butler, James Odelle. 2013. *Name, place, and emotional space: themed semantics in literary onomastic research*. University of Glasgow PhD thesis.
- Chen, Lindsey N. 2008. Ethnic marked names as a reflection of United States isolationist attitudes in Uncle Scrooge comic books. *Names* 56(1). 19–22. <http://dx.doi.org/10.1179/175622708X282901>.
- Esuli, Andrea & Fabrizio Sebastiani. 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC '06)*, 417–22.
- Florou, Eirini & Stasinou Konstantopoulos. 2011. A quantitative and qualitative analysis of Nordic surnames. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), Riga, Latvia, 11–13 May 2011*, vol. 11 (NEALT Proceedings Series), 74–81. <http://hdl.handle.net/10062/17292>.
- Fonagy, Ivan. 1961. *Communication in poetry*. William Clowes.
- Font Litjós, Ariadna & Alan W. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proceedings of Eurospeech 2001, Aalborg, Denmark*.
- Gutiérrez, Elkin Darío, Roger Levy & Benjamin K. Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the ACL (ACL 2016), Berlin, Germany, 7–12 August 2016*, 2379–88.
- Huang, Fei. 2005. Cluster-specific named entity transliteration. In *Proceedings of Human Language Technology Conference and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP 2005), Vancouver, BC, Canada*, 435–442.
- Kaplan, David M. & David M. Blei. 2007. A computational approach to style in American poetry. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, 553–8. <http://dx.doi.org/10.1109/ICDM.2007.76>.
- Konstantopoulos, Stasinou. 2007. What's in a name? In *Proceedings of Computational Phonology Workshop, held at the International Conference on Recent Advances in NLP (RANLP 07), Borovets, Bulgaria, September 2007*.
- Konstantopoulos, Stasinou. 2010. Learning language identification models: a comparative analysis of the distinctive features of names and common words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010), Valletta, Malta, 19–21 May 2010*.
- Markey, T. L. 1982. Crisis and cognition in onomastics. *Names* 30(3). 129–142. <http://dx.doi.org/10.1179/nam.1982.30.3.129>.

- Miall, David. 2001. Sounds of contrast: an empirical approach to phonemic iconicity. *Poetics* 29(1). 55–70.
- Monaghan, Padraic, Richard Shillcock & Morten H. Christiansen. 2014. How arbitrary is language? *Phil. Trans. of the Royal Society of London B* 369(1651).
- Nicolaisen, William F. H. 2008. On names in literature. *Nomina* 31. 89–98.
- Papantoniou, Katerina & Stasinos Konstantopoulos. 2016. Unravelling names of fictional characters. In *Proceedings of the 54th Annual Meeting of the ACL (ACL 2016), Berlin, Germany, 7–12 August 2016*.
- Ramachandran, Vilayanur S. & Edward. 2001. Synaesthesia: a window into perception, thought and language. *Journal of Consciousness Studies* 8(12). 3–34.
- Rudnycky, Jaroslav B. 1959. Function of proper names in literary works. *Internationalen Vereinigung für moderne Sprachen und Literaturen* 61. 378–383.
- Shillcock, Richard, Simon Kirby, Scott McDonald & Cris Brew. 2001. Filled pauses and their status in the mental lexicon. In *Proceedings of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*.
- Spiegel, Murray F. 1985. Pronouncing surnames automatically. In *Proceedings of the 1985 Conference of the American Voice Input/Output Society*, 109–132.
- Tamariz, Monica. 2008. *Exploring the adaptive structure of the mental lexicon*. Edinburgh University PhD thesis.
- Virga, Paola & Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the Workshop on Multilingual and Mixed-language Named Entity Recognition, held at ACL 2003*, 57–64.
- Whissell, Cynthia. 1999. Phonosymbolism and the emotional nature of sounds: evidence of the preferential use of particular phonemes in texts of differing emotional tone. *Perceptual and Motor Skills* 89(1). 19–48. <http://dx.doi.org/10.2466/pms.1999.89.1.19>.

## Chapter 21

# Second thoughts about the Chomskyan revolution

Jan Koster

University of Groningen

### 1 Introduction<sup>1</sup>

According to a widespread belief, the field of linguistics was revolutionized since the middle of the previous century by the ideas of Noam Chomsky at the Massachusetts Institute of Technology in the USA. This revolution, as I will discuss in this article, is a myth. To be sure, to prevent misunderstanding at the outset, it has to be said that linguistics in the intended sense has been a great success. In contrast with the often dull field of the 1950s, Chomskyan linguistics has rejuvenated the field and even given it intellectual mass appeal occasionally. The field has not only exploded quantitatively but also qualitatively. We have seen growth that, no doubt, is unique in the history of the humanities. And yet, it has been my conviction for a long time that something is wrong with the field, not only in its technical development, but particularly in the way it is interpreted at a meta-theoretical level.

Before Chomsky, European structuralism had a broadly Saussurean orientation. I will argue that, technically speaking, the first 20 years of generative grammar, far from being revolutionary, showed a gradual reinvention of the structuralist wheel. Even more confused has been the persistent meta-theoretical reinterpretation of language (in some narrow sense) as a specialized biological faculty. The technical problem and the meta-theoretical problem are largely independent but, in practice, conspire to create the current theoretical stagnation.

Before going on, it is useful that I first give a short summary of the relevant aspects of the Saussurean heritage, contrasting it along the way with some central Chomskyan tenets. The reader should keep in mind that I am mostly focused on syntax

---

<sup>1</sup> I would like to thank Christina Gagné for discussion and helpful comments on an earlier draft of this article. All remaining errors are my own.

and semantics and that I leave the evaluation of developments in phonology happily to others. According to Saussure, language (in the narrowest possible sense) is not an individual-psychological or biological faculty (as for Chomsky) but a system of signs. Signs are invented artifacts and, as such, belong to the shared culture of some language community (see also Popper 1972).

Signs according to Saussure have three dimensions, the best known of which are the overt *signifiant* (audible visible, readable, whatever is the sensory mode) and the covert *signifié* (some conceptual substance). Trivially, the relation between *signifiant* and *signifié* is conventional, as has been a common idea at least since Aristotle.

The second covert dimension of the Saussurean linguistic sign is often overlooked and is about the relations between a simple sign and its environment. Thus, every competent speaker who knows the English word *book* also knows that it can be preceded by an article: *the book*. Part of a linguistic sign, then, can be “under water”. Thus, we have signs like *Art book*, *book PP*, *Art book PP*, etc. (where variable *Art* stands for article and variable *PP* for prepositional phrase). Each sign has a set of more or less fixed, invisible environments, which can be made visible by substituting the variables by constants: *the book*, *books from Geneva*, *the books from Geneva*, etc. The predictable environments of signs fall under what Saussure called “syntagmatic relations” and were later called the “valency” of a sign (mostly applied to verbs, see Tesnière 1959). Syntactic structures, then, are complex signs, which spell out the environmental properties of simpler signs. Both simple signs and complex signs conform to the conventions of the “trésor commun”, the non-individual “langue.”

Needless to say, our conventions are constrained by our individual biological properties, for the same trivial reasons as why all our forms of culture are constrained by our nature. Note also, that there are no specific derivational levels that can be referred to as “the interfaces”. Each linguistic sign, including the simplest morpheme, is a three-fold interface, connecting “signifiant”, “signifié” and conventional combinatorial potential (“syntagmatic relations”). There are also sign systems, like traffic signs, that miss a significant syntagmatic dimension and are therefore only interfacing two dimensions (*signifiant* and *signifié*).

## 2 Some theories of generative syntax since the 1950s

### 2.1 (Extended) Standard Theories

The following three are the most important versions of Chomskyan generative grammar since the 1950s:

- (1) a. (Extended) Standard Theories (1955–1973)
- b. Government–Binding Theories (1973–1995)
- c. Minimalist Theories (1995– )

This is a rough periodization which cannot be exact for the simple reason that key ideas from the various periods overlap. The idea of a Chomskyan revolution is particularly based on the much hyped first period (1a). It started in 1955 with the

voluminous *The Logical Structure of Linguistic Theory*, was popularized by *Syntactic Structures* (1957) and culminated in *Aspects of the Theory of Syntax* (1965), the so-called Standard Theory. This period was concluded by the Extended Standard Theory of Chomsky (1971) and the lexicalism of Chomsky (1970).

The next period (1b) sought to improve explanatory adequacy further by limiting the initial hypothesis space not only by X-bar theory (Chomsky 1970) but, most typically, also by “conditions on transformations”, as in the paradigmatic articles Chomsky (1973) and Chomsky (1977). Movement rules, the most important transformations in this framework, were gradually reduced to the schema “Move Alpha”. This type of theorizing culminated in another paradigmatic work, *Lectures on Government and Binding* (1981) (GB). In spite of the fact that GB-style theorizing was gradually hoped to be superseded, particularly after Chomsky (1995), by minimalist speculations (1c), “normal science” in generative linguistics is largely determined by GB-type analyses, up until the present day. Many linguists are only terminologically affected by Minimalism, replacing “Move Alpha” by “internal Merge” and re-baptizing good old bounding domains as “phases” (yes, I know, there are subtle differences).

Many generative linguists believe that (1a-c) is a continuous story of progress, which started with the glorious revolution of the late 1950s. In my opinion, this self-image of the field is false. Only the first period (1a) saw some potentially revolutionary ideas, but by the early 1970s it was, or should have been, clear that they were all ill-conceived. For concreteness’ sake, consider the following tenets of the early theories:

- (2) a. syntax-based theories instead of sign-based theories
- b. two-step sentence generation: PS-rules (kernel) and transformations
- c. multiple levels of representation: deep and surface structure
- d. use of formal methods (mathematics, logic)

About (2d) we can be brief: after some initial results, mathematical linguistics practically disappeared from Chomskyan practice since the 1970s. This is ironical because the initial interest in mathematics was a formidable tool of propaganda in establishing the field’s prestige and image of revolutionary paradigm shift. Since the 1970s, interest in mathematics and logic mostly lived on in formal semantics (Generative Semantics, Montague-inspired work, etc.).

From a Saussurean structuralist’s point of view, talking about syntax independent of the properties of signs is bound to be a failure. And so it happened. *Syntactic Structures* (1957) generated structures without looking at the internal properties of lexical items. However, as soon as a lexicon is added, it appears that the PS-rules mimic the valencies that are also spelled out as internal properties of the lexical items (Chomsky 1965). Chomsky (1970), and particularly Chomsky (1981) drew the obvious conclusion, namely that doing things twice can only be prevented by projecting syntactic structures directly from lexical items. This was no revolutionary new insight but the reinventions of the Saussurean wheel, according to which syntactic structures are defined as the syntagmatic properties (valencies) of signs. Exit revolutionary tenet (2a).

With (2a) and (2d) gone, the idea of a Chomskyan revolution had little to boast beyond the interrelated (2b) and (2c). These were perhaps the most characteristic features of the intended revolution. Especially the term “deep structure” was of great propagandistic value at the time, obscuring the fact that it stood for something close to and as unexciting as the “strings underlying kernel sentences” in Chomsky (1957).

The history of (2b) and (2c) is very interesting. The *prima facie* revolutionary innovation of generative grammar was in the transformations, not in the PS-rules. The transformations were tightly connected with the new idea of multi-level representation, as they formed the core of the mapping from deep structure to surface structure. Given the fact that these were key concepts of the alleged revolution, it is, in retrospect, astonishing how fast transformations disappeared from the theoretical scene around 1970. Major classes of transformations, like pronominalizations and equi-NP-deletion, appeared to be ill-conceived and were replaced by rules of construal, the essence of which is the local sharing of properties (see Koster 1987: 8ff). Before long, construal was demonstrated to be also the better alternative for movement transformations (with NP-movements and Wh-movements as the major families; see Koster 1978).

The end of transformationalism was considerably speeded up by insights about structure-preservingness. Emonds (1970, followed by some later elaborations) demonstrated that major movement transformations were structure-preserving in the sense that they created structures that could be independently created by the base rules (X-bar schemata). Others, including Chomsky, argued that movement rules left a “trace”, like the empty NP-object in (3):

(3) *What* did you see [<sub>NP</sub> ...]

Of course, this very terminology of structure-preservingness and “traces” is odd in retrospect, as what we see is just “base-generated” *What* (for instance, as the Spec of CP) and the “trace” [<sub>NP</sub> ...] as the spelled out valency of *see*. The then current terminology presupposed transformationalism, while what actually was demonstrated was that transformationalism is false, also for “movement” rules. Altogether, then, it is fair to say that transformationalism was as dead as a doornail by the mid 1970s.

This outcome should have had serious consequences for how we evaluate the so-called Chomskyan revolution. The least we can say is that, in retrospect, it was more hype than substance. However, I think we should go one step further: it created perhaps impressive enthusiasm and empirical momentum but conceptually, it was (at least until the mid-1970s) a complete failure and left the field more or less where it was before the mid-1950s. With transformations gone (2b), multiple-level representation (2c) was gone as well. What remained was X-bar structures projected from the lexicon as the foundation for property sharing (construal). This is just a variant of traditional *Wortgruppenlehre*, entirely compatible with Saussurean and other pre-1950s structuralist assumptions (see also Ries 1928, De Groot 1949 and, with a more English-oriented account of the parts of speech, Jespersen 1924).

## 2.2 Government–Binding and Minimalist Theories

With the generative enterprise diminished to a more or less traditional *Wortgruppenlehre*, the focus of the field shifted to “principles and parameters”, with the ambition to restrict the class of possible grammars to a very few or even one, with open parameters to account for the differences among languages. The theory of parameters never was developed beyond superficial descriptions close to the data, like the VO/OV-parameter supposed to account for the difference between VO- and OV-languages. Principles were often empirical generalizations for specific domains, like bounding theory for “movements” and binding theory for anaphoric constructions. The “principles and parameters” framework has had two varieties so far, the Government-Binding versions (1b) and the Minimalist versions (1c). For present purposes, it is important to keep in mind that the GB-versions are based on X-bar theory and therefore a continuation of the traditional *Wortgruppenlehre*. This was implicitly denied by Chomsky, as will be clarified in a moment. Minimalist theories are more radical in that their core is a step away from traditional word groups and constructions. This core is the Galilean grail known as “Merge”, which produces word groups only through interactions known as “mappings to the interfaces” (sensori-motoric and conceptual-intentional).

The most important question from the current perspective is whether the failed revolution of the first period (1a) got a new chance in the “principles and parameters” framework. The short answer is: no, the field gradually disintegrated into conceptual chaos. There were plenty of linguists who took the consequences of the developments of around 1970 and got rid of transformationalism altogether, thereby implicitly recognizing the failure of the revolution (Brame 1978; Bresnan 2001; Pollard & Sag 1994; Koster 1978; 1987). Chomsky, in contrast, preferred denial and sought to save transformationalism with the transformational residue “Move Alpha”. This enabled him to leave the revolutionary illusion intact, with a full-fledged defense of multiple levels of representation (from then on called D-structure, S-structure and Logical Form (LF); see Chomsky 1981 and Koster 1987 for a critique). Never mind that these levels later on disappeared through the minimalist backdoor, generously ignoring the fact that the relevant insight had been available for decades.

The most representative sub-theory of the GB-framework is the so-called bounding theory, with Subjacency as its core principle. So, if we want to know if generative grammar ever deserved the predicate “revolutionary” after all, it is useful to focus on Subjacency. We would minimally expect Subjacency to be non-trivial and innovative. It appears that Subjacency, and the problems it seeks to solve, is an artifact of the transformational framework that was already about to collapse by the time Subjacency was introduced in Chomsky (1973).

According to Subjacency, no transformational rule involves two categories X and Y across **two** bounding nodes. The bounding nodes were NP and S' (or the parametric variant S), which were later re-baptized as DP and CP (IP). Subjacency was supposed to be a more principled, deeper replacement of the famous, but somewhat disparate list of island constraints of Ross (1967), which in turn replaced the more appealing but empirically inadequate A-over-A Principle of Chomsky (1964).

Two questions immediately come to mind with respect to Subjacency in the current context: 1) is it empirically adequate, and 2) does it exceed the conceptual boundaries of traditional, pre-Chomskyan *Wortgruppenlehre*. I do not have the space here to do full justice to these questions, but in summary, answers amount to the following. As for empirical adequacy, Subjacency appeared to be wrong on two counts. First of all, the number two (highlighted in the description of Subjacency above) appeared to be unmotivated. Whenever Subjacency applies in the relevant contexts, one node is enough. Thus, the relations anachronistically described by movement transformations, in the unmarked case, never apply across single NP (DP)-boundaries, as correctly stipulated by the NP-constraint of Bach & Horn (1976). Never in such cases, a second boundary node (like S' (CP), S (IP) or a second NP (DP)) has to be specified. Second, an interesting, largely ignored discovery was made as soon as other languages were taken into account. Thus, largely on the basis of Dutch but with an eye for lots of other languages, van Riemsdijk (1978) concluded that PP had to be added to the inventory of bounding nodes. Koster (1978) generalized bounding to all lexical categories (maximal projections, extended XPs in the sense of Grimshaw 2000).

The significance of these discoveries has been completely misunderstood. It was implicitly shown that, far from being the effect of a fancy 2-node condition on fancy transformational rules, bounding phenomena showed an effect that could have been obvious from traditional *Wortgruppenlehre*, namely that the "size" of a word group is determined by the valency of its lexical head. For instance, the DP complement of a P is only interpreted as such within a PP, which is the exclusive domain of the P's syntagmatic relations. In the unmarked case, then, bounding is what is to be expected on lexicalist assumptions: it is limited to maximal projections (word groups). In some languages, like English, in a very narrow range of contexts, domains can be extended further by absorption of V-, A-, or P-projections into a more encompassing V-projection (see Koster 1987).

In short, the core principle of the next "revolutionary" stage, GB-theory, is reducible to what follows from the traditional definition of a word group.

What about Minimalism (1c, Chomsky 1995)? The core notion of Minimalism is recursive Merge (plus mappings to the interfaces). The potential revolutionary character of Merge is believed to rest on its radical "Galilean" nature, i.e., its status of a perfect object that only produces the "messy" data in interaction with external factors. In principle, this seems to be an interesting move, but in practice it comes down to the observation of properties that are hardly controversial and entirely compatible with most theories of syntactic structure, both traditional (like *Wortgruppenlehre*) and modern (like Construction Grammar).

First of all, it should be noticed that Merge is like pre-lexicalist, pre-1970 grammar in that it introduces hierarchical recursive structure independent of the lexicon. But since the merged objects also have hierarchical, recursive structures as their valency, Merge reintroduces the redundancy that X-bar theory was designed to get rid of. This does not look like progress. However, one can maintain that Merge at least introduces the right properties: binary branching and adjacency of the related cate-

gories. Reformulated as a constraint on representations, the substance of Merge is preserved and the redundancy problem is solved. As before, syntactic structures are projected from the lexicon (instead of being generated with Merge) and they are accepted as long as they conform to the constraints formerly derived from Merge. I have shown elsewhere that this view of Merge as a meta-constraint on representations has other interesting consequences as well, as stated in the theory of triads (see Koster 2007; 2015).

For the belligerent among us, this is bad news because it makes the properties of Merge compatible with almost any framework on the market. Theorists of Construction Grammar, for instance, can postulate constructions without fear, as long as their constructions conform to the locality constraints formerly contributed to Merge.

Most significantly, the locality constraints are also entirely compatible with pre-revolutionary *Wortgruppenlehre*. Again, there is no reason to complain about the descriptive-analytic productivity of the field since the mid 20th-century, but the idea of revolutionary conceptual innovation is in urgent need of revision. The revolution was mainly a matter of hype and rhetoric. As for substance, there is almost complete continuity with the tradition. Philosophically speaking, the pre-WWII European ideas about language were even significantly better, as will be discussed next.

### 3 Against biolinguistics

The main shift since the 1950s has been from an external, socio-cultural concept of language to language seen as a matter of an individual faculty, also referred to as a genetically determined universal grammar (UG) or as I-language (where I stands for individual, internal and intensional). Ultimately, it is hoped, I-language can be unified with biology, whence the term “biolinguistics” (see Lenneberg 1967; Jenkins 2000; Hauser et al. 2002).

Inspired by Saussurean ideas, the focus of pre-Chomskyan linguistics in Europe was on the collective “langue”, a public system of signs serving the needs of symbolization. Of course, individual aspects were also recognized, but those were relegated to “parole” and seen as outside the language system in the narrow sense. Seeing language as a phenomenon external to the individual was the common view. Major philosophers, like Karl Popper, for instance, saw language as a “World 3” phenomenon (see Popper 1972). All of this was in accordance with the insight that one of the most characteristic aspects of human minds is their living in symbiosis with shared external memories (see Donald 1991). Language was somehow seen as the pivotal phenomenon bringing this symbiosis about. I think this view is correct and that the view of language as an individual faculty is wrong.

Language is a socio-cultural phenomenon because it is based on morphemes and words. Words do not grow from trees or in wombs but are artifacts invented by someone and adopted by the community, usually of the inventor (with exceptions such as loanwords). Words thus invented and adopted are maintained as part of the community’s cultural record, in the form of oral or written traditions. Morphemes and words have, as we saw, a valency that can be lexicalized to form more complex

signs. These valency patterns are also part of the record of a language community, even if many community members are not aware of that. However, if somebody, on purpose or by accident, deviates from the accepted norm, most community members will notice that. Valency patterns are recursive because each realized pattern opens up new slots with new realizable contexts, etc., *ad infinitum*.

Needless to say, such complex cultural objects with recursion, can only be handled by brains that are able to do so. However, from the relation between the ability and complex objects, it does not follow that the ability in question is a language ability. The relation is not intrinsic but accidental, a relation also known as an application. Applications must have an agentive cause, a context that somehow give the object a function. The distinction that we commonly make between biological and cultural applications (or functions) depends on the nature of the agentive cause. If the cause is a quasi-agent, like natural selection, we call the application 'biological'. Examples of biological functionality are the natural functions of the organs of the body. Similarly, so-called bio-computation, like the kind involved in mammalian vision, is biological because the functionality is not caused by human intervention.

However, if the agentive cause is human invention, we commonly call the application 'cultural'. Interesting examples are those that involve both agentive- and non-agentive functionality. Consider organs like the lungs. The lungs have an obvious biological function brought about by non-agentive causation, for instance, by natural selection. However, if we wish we can also give the lungs a cultural application, mediated by human-made artifacts, like trumpets and other wind instruments. This example shows that the status of the application (biological or cultural) has nothing to do with the innateness of the structures involved. Obviously, the lungs are innate in the relevant sense. Nor does it matter how many applications there are. In this case, the lungs have exactly one cultural application.

The only thing that matters is the nature of the agentive cause of the application. By that criterion, Chomsky has it wrong when he compares linguistic functionality with the functionality of the organs or the computations involved in the mammalian visual system. The linguistic application of our capacity for recursion (innate or not) is agentive because mediated by human-made artifacts like the trumpet. In this case by linguistic signs, like morphemes, words or phrases.

It must be concluded therefore that biolinguistics is an untenable proposition. All things cultural, from sports to music and language, exploit the more or less innate capacities of our body and brain, crucially as a matter of free agentive application. This in contradistinction to the biological functions of the body, which do not enjoy such freedoms of application.

I will end with an argument against biolinguistics based on intentionality. Inspired by medieval examples, Franz Brentano (1874) developed the valuable insight that there is an essential difference between merely physical states and mental states. The difference is intentionality or "aboutness". Thus, an arbitrary object, like a stone is not about anything. However, the *word* "stone" involves mental states and is therefore about something, for instance about stones.

The relevant notion of aboutness deserves some further clarification. One might

say, for instance, that instruments are about something. Thus, a thermostat can be said to be about temperature and an alarm clock about time. Obviously, however, thermostats and alarm clocks do not have mental states. Insofar as these instrumental objects are about something, it is thanks to the user. Therefore, we say in these cases that the objects have “derived intentionality”. In general, intentionality involves mental states that relate a target (what the states are about) with a source. The most interesting aspect of intentionality, highly relevant for cognitive science, is the question what is the source of intentionality.

Crucial is the insight that brains and brain states are tools serving the user and therefore intentional objects with derived intentionality. The question, then, is what is the source of their intentionality. There is a misguided tendency to look for the source at the brain itself (“we are our brains”). Bennett & Hacker (2003) have called this “the mereological fallacy”: confusing the part (the brain) with the whole (the person using the brain). In short, people, not their instruments such as computers and brains, are the sources of intentionality.

The crux of the argument is that people (living agents) cannot be defined in purely physical or biological terms because they have, next to a physical identity a socio-cultural identity and a history. The latter parts of our identity are at least as important as the former. In general, therefore, the intentionality of the mental is a decisive barrier against its naturalization in our theories of cognition.

This is directly relevant as an argument against biolinguistics. According to Chomsky, “knowledge of language” (I-language) is a state of the mind/brain. Language learning, in his view, is a development from an initial state  $S_0$  to a relatively stable state  $S_s$ , where  $S_0 = UG$  (see, for instance, Chomsky 2007). However, it is a serious error to say that knowledge of language is a brain state. This would be of the same calibre as saying that knowledge of time is a state of your alarm clock. Like clock states, brain states have derived intentionality and only represent knowledge thanks to the necessary source and target of their intentionality. We have already seen that the source of human intentionality cannot be naturalized. This suffices to make “biolinguistics” an unattainable goal. However, things are much worse for biolinguistics. As a state of the mind/brain  $S_s$ , Chomsky’s “knowledge of language” is intentional, therefore not only has a source but also a target (what the mental state is about). What, then, is Chomsky’s state  $S_s$  actually about? It never has become clear.

One thing is certain, however: there is no known biological answer to the question what language-as-a-mental state is about. Luckily, as was proposed earlier on in this article, there is a perfectly satisfying, traditional structuralist answer: knowledge of language is about systems of complex signs (Saussure: ‘langue’) and their use (‘parole’). If this answer is correct, biolinguistics must also be rejected from the target side of the intentionality relation. This is so, I can only repeat, because signs are not growing from trees or in wombs but are inventions by human agents and preserved in our cultural records.

## 4 Conclusion

The history of Chomskyan generative grammar is in urgent need of revision. All what seemed revolutionary about it in the 1950s and 1960s turned out to be untenable, often as early as in the 1970s. Later revisions failed to reanimate the revolution and were more than once a step back in the direction of pre-Chomskyan models of grammar, leaving a GB-style analytic-descriptive kind of normal science, with theoretical notions mostly compatible with both traditional and more recent kinds of frameworks. The “biolinguistic” effort of recent years is doomed to failure, as it continues the fundamental error of seeing language (in some narrow sense) as an individual mental state rather than as a Sasseurian “trésor commun”.

## References

- Bach, Emmon & George Horn. 1976. Remarks on ‘Conditions on transformations’. *Linguistic Inquiry* 7. 265–299.
- Bennett, Maxwell & Peter Hacker. 2003. *Philosophical foundations of neuroscience*. Malden, MA: Blackwell.
- Brame, Michael. 1978. *Base generated syntax*. Seattle: Noit Amrofer.
- Brentano, Franz. 1874. *Psychologie vom empirischen Standpunkte*. Leipzig: Von Duncker & Humblot.
- Bresnan, Joan. 2001. *Lexical-functional syntax*. Malden, Mass.: Blackwell.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
- Chomsky, Noam. 1964. *Current issues in linguistic theory*. The Hague: Mouton.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Chomsky, Noam. 1970. Remarks on nominalization. In Roderick Jakobs & Peter Rosenbaum (eds.), *Readings in English transformational grammar*, 184–221. Waltham, MA: Ginn & Company.
- Chomsky, Noam. 1971. Deep structure, surface structure and semantic interpretation. In Danny Steinberg & Leon Jakobovits (eds.), *Semantics: an interdisciplinary reader in philosophy, linguistics and psychology*, 183–216. Cambridge: Cambridge University Press.
- Chomsky, Noam. 1973. Conditions on transformations. In Stephen Anderson & Paul Kiparsky (eds.), *A festschrift for Morris Halle*, 232–286. New York, etc.: Holt, Rinehart & Winston.
- Chomsky, Noam. 1975. *The logical structure of linguistic theory*. New York: Plenum Press.
- Chomsky, Noam. 1977. On Wh-movement. In Peter Culicover, Thomas Wasow & Adrian Akmajian (eds.), *Formal syntax*, 71–132. New York: Academic Press.
- Chomsky, Noam. 1981. *Lectures on government and binding: the Pisa lectures*. Dordrecht: Foris.
- Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Mass.: MIT Press.
- Chomsky, Noam. 2007. Approaching UG from below. In Uli Sauerland & Hans-Martin Gärtner (eds.), *Interfaces + recursion = language?*, 1–29. Berlin: Mouton de Gruyter.

- de Groot, Albert W. 1949. *Structurele syntaxis*. Den Haag: Servire.
- de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Charles Bally & Albert Sechehaye (eds.). Lausanne & Paris: Payot.
- Donald, Merlin. 1991. *Origins of the modern mind*. Cambridge, Mass.: Harvard University Press.
- Emonds, Joseph. 1970. *Root and structure preserving transformations*. Cambridge, Mass.: MIT PhD thesis.
- Grimshaw, Jane. 2000. Locality and extended projections. In Peter Coopmans, Martin Everaert & Jane Grimshaw (eds.), *Lexical specification and lexical insertion*, 115–133. Amsterdam: John Benjamins.
- Hauser, Marc, Noam Chomsky & Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science* 298. 1569–1579.
- Jenkins, Lyle. 2000. *Biolinguistics*. Cambridge, Mass.: MIT Press.
- Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen & Unwin; New York: Holt.
- Koster, Jan. 1978. *Locality principles in syntax*. Dordrecht: Foris.
- Koster, Jan. 1987. *Domains and dynasties*. Dordrecht: Foris.
- Koster, Jan. 2007. Structure-preservingness, internal merge, and the strict locality of triads. In Simin Karimi, Vida Samiiian & Wendy Wilkins (eds.), *Phrasal and clausal architecture: syntactic derivation and interpretation*, 188–205. Amsterdam-Philadelphia: John Benjamins.
- Koster, Jan. 2015. Aspects and beyond: the case for generalized construal. In Ángel J. Gallego & Dennis Ott (eds.), *50 years later: reflections on Chomsky's Aspects* (MIT Working Papers in Linguistics 77), 159–169. Cambridge, MA.
- Lenneberg, Eric. 1967. *Biological foundations of language*. New York: John Wiley & Sons, Inc.
- Pollard, Carl & Ivan Sag. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Popper, Karl. 1972. Epistemology without a knowing subject. In Karl Popper (ed.), *Objective knowledge: an evolutionary approach*, 106–152. Oxford: Oxford University Press.
- Ries, John. 1928. *Zur Wortgruppenlehre: mit Proben aus einer ausführlichen Wortgruppenlehre der deutschen Sprache der Gegenwart*. Prag: Taussig & Taussig.
- Ross, John Robert. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT PhD thesis.
- Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Editions Klincksieck.
- van Riemsdijk, Henk. 1978. *A case study in syntactic markedness: the binding nature of prepositional phrases*. Dordrecht: Foris.



## Chapter 22

### Good maps

William A. Kretzschmar, Jr.

University of Georgia; University of Glasgow

What makes a good map? The question is really quite simple: a good map is what allows you to find what you are looking for. However, since we look for different things—addresses, highways, cities, rivers, mountains, territories, languages, dialects, and many more—there is no such thing as the best map. A single map cannot show all the different things we want to find. Thus good maps are the best we can hope for, maps that let us find particular things. Bad maps, in the same way, are bad because they show us something other than what we want to find. A London Tube map is bad for finding a street address in the city. A street map of London may be bad because it only shows the major roads of the city and not the address we are trying to find. Some maps do a poor job of representing what we need to know, like the neighborhood map of London drawn on a napkin in a pub, not to scale, missing roads or landmarks; they show us what the person drawing the map remembers, which may not be enough for us to find the address we want. Good maps, therefore, have an information model that matches what we are trying to find, one that has the right information in three ways: the right kind of information, information at the right scale, and information that is accurate. Linguistic maps follow the same principles as London maps. They need to show the right kind of information about language; they need to show that information at the right scale; and they need to show their information accurately.

Early linguistic maps did not show the right kind of information about language. Maps with isoglosses assumed that a linguistic feature was used in one place but not in another place, so that it was reasonable to draw a line that separated the area of use from the area of non-use. Language in use, however, never distributed itself so neatly into separate areas. As I have shown in detail elsewhere (Kretzschmar 1992; 2003), drawing lines to represent the use of linguistic features always includes (usually unstated) generalizations about frequency, that some feature is used *more often* on one side of a line from the other side. Isoglosses can also be drawn differently at the option of the mapmaker, to delineate broader or narrower areas of usage because the data is sampled, never a complete record of the linguistic usage of an area.

*William A. Kretzschmar, Jr.*

These problems provoked different interpretations of the isogloss, as a limit of occurrence (the classical definition) or as a transitional area where the use of one feature changed over to the use of another feature. Definition as a transitional area did not solve the problem of the isogloss as a limit, however, because it is never the case that features are used uniformly and uniquely in any area: there is always some mixture of different realizations for the same feature, even in places that ought, we think, to be the center of some area of usage. The basic problem with isoglosses is that, while we may want to represent areas of uniform usage of a linguistic feature, areas of uniform usage simply do not occur, not in any language, not for pronunciations, words, or grammar. For evidence all we need to do is look at maps of features in American English as collected for the Linguistic Atlas Project ([www.lap.uga.edu](http://www.lap.uga.edu)). It is possible to make thousands of maps on the web site, and no map shows any area where a feature is uniformly used. We may want language to look that way, features divided into complementary regions, but that is not how it works as people use language, and thus a map based on the assumption of areas of uniform usage will always show the wrong kind of information about language.

When mapmakers bundle isoglosses in order to show dialect areas, areas where several features coexist while they do not coexist on the other side of the bundled lines, the distributional problem just gets worse. Unstated assumptions about frequency multiply, and subjective choices about where to draw the lines similarly amplify. As William Moulton has written (1968: 456):

“Ideally, an investigator might have plotted all possible isoglosses and let the dialect divisions fall where they may. In practice this was never done, since a plotting of all possible isoglosses seemed to reveal no clear geographical structure at all and even to refute the very notion of ‘dialect area’—which was what the investigator set out to demonstrate in the first place. Accordingly, what the investigator typically did was to develop some sort of intuitive idea of the areas he wanted to find; he was then able to pick and choose isoglosses—especially bundles of isoglosses—that could be patched together so as to reveal the desired areas.”

Moulton in 1968 was talking about the work of Hans Kurath and other linguists who had pursued the isoglossic model in the first surveys of European languages and American English. This practice continues today in Labov, Boberg & Ash (2006), as shown in Figure 1.

The lines on this map enclose dots (speakers) of different colors (degrees of fronting) so that they are not limits of occurrence or transition areas, and the Northern areas drawn can only be subjective regions. Of course, Labov and his colleagues know a great deal about American English, and we have reason to want to trust their judgment; we just need to realize that that is what the map shows us, judgment, and not objective fact about dialect regions in American English. If we want to find “Northern” speakers of American English the map tells us where to look, but the problem is that the idea of “Northern” speakers is not well supported by the map. If we find the map helpful we agree with Labov and his team, but language does not

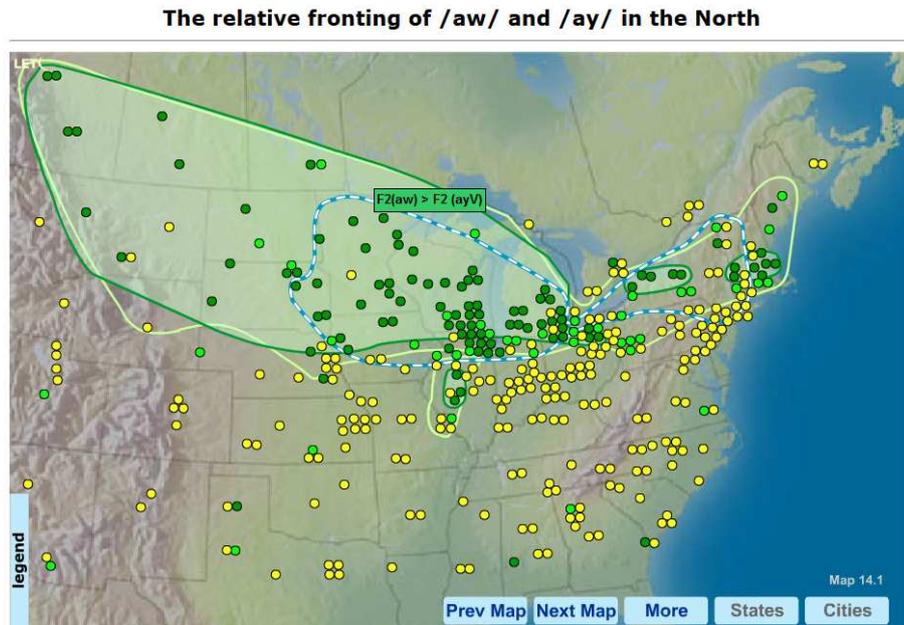


Figure 1: Northern Fronting of /aw/ and /ay/ (Labov, Boberg, and Ash 2006, viewed online at <http://www.atlas.mouton-content.com/secure/generalmodules/anae/unit0031/genunstart.html> on 9/8/2016).

work quite so neatly, and it is a mistake to think that all of the speakers inside of the lines talk the same.

Figure 1 also illustrates the problem of scale. Labov and his colleagues chose two people for their survey from each metropolitan statistical area in North America. They were able to make a national survey that way, but their data does not allow users to make good generalizations about smaller areas. So, for example, Figure 1 shows two yellow dots in Birmingham and two yellow dots in Montgomery, plus a green dot near Tuscaloosa, but on the basis of these five speakers we cannot make a generalization about all of Alabama, or even about urban Alabama. The Linguistic Atlas of the Gulf States interviewed 127 speakers from Alabama, which gives a much better indication of language use at the state scale in Alabama, and it is possible on the Atlas web site to plot responses of those Alabama speakers. In older-style North American sociolinguistics it was thought that sampling did not matter, that choosing just a few speakers from a place was enough to make good generalizations, but the emergence of community-of-practice studies has shown that every place has many different groups of speakers, not homogenous speech.

Modern linguistic maps typically use statistics to make generalizations about how language is used across space. My own work has focused on single features, such as a single variant answer (like *pail* or *bucket*) for the question about the container for wa-

William A. Kretschmar, Jr.

ter from a well. Figure 2 shows a view of the *pail* response after processing with the statistic known as density estimation (the data is available on the Atlas web site). The map of the Middle and South Atlantic States has been divided in approximately 3000 locations, and the locations have been shaded to represent the likelihood that *pail* might be elicited there, the darker the more likely, based on the nearest neighbors of each location. The sample for the Atlas did not select speakers from all 3000 locations (there were only 1162 informants), so the statistic generalizes from the known locations to make estimates for all of the locations. It is clear from the map that usage of *pail* does not occur in neat areas, and that within any area there may be higher and lower estimates of the probability of eliciting the form. The nearest neighbors method preserves this granularity of responses. It is of course possible to smooth the estimates in order to make neat areas, as shown in Figure 3, the same data processed with smoothing. Figure 3 is less accurate, to my way of thinking, because it creates smooth areas where there really are none; smoothing responds to the idea that language should be used in neat regions.

Recently Ilkka Juuso and I have experimented with a multivariate approach to density estimation, as shown in Figure 4. Different variants for a question, here the question about what to call the event often known as a thunderstorm, are shown in a different color. The least-frequently-occurring variants drop out of the picture, leaving more common variants (*thunderstorm* in blue, *thundershower* in forest green, *thundercloud* in olive green, *thundersquall* in purple), and for those variants the greater the intensity of the color the greater likelihood of elicitation at any location. Figure 4 shows how a multivariate density estimation map compares to a univariate density estimation map, by sketching the outlines of the regions where *thundercloud* appeared in a univariate analysis.

This map retains its accuracy because of the different colors and different intensities. It does not smooth the data into discrete regions. It also operates at its own level of scale, one where only common variants appear on the map; this map will not tell the user where people say any of the dozens of uncommon words for thunderstorm. Still, multivariate density estimation effectively addresses the question of who says what where with regard to thunderstorms.

Grieve, Speelman & Geeraerts (2013) have recently applied modern methods to the data gather by Labov for his North American Atlas. Figure 5 shows raw values for F2 in /ay/ for Labov's data, and just as we would expect, they fail to pattern themselves into neat areas.

Grieve and colleagues then apply spatial autocorrelation statistics to Labov's data (Kretschmar 1992 had introduced spatial autocorrelation to the field), as in Figure 6 for the same data shown in Figure 5. The prevalence of colors in some regions indicates that the F2 values of /ay/ do have similar neighbors. Grieve and colleagues then conduct a factor analysis on all 38 vowel variables available in the Labov data, and identify four factors that account for a great deal of the variance in the spatial autocorrelation scores (together, 86%). Multiple vowels are implicated in each factor. A hierarchical agglomerative cluster analysis was then conducted on the factor scores, as shown in Figure 7.

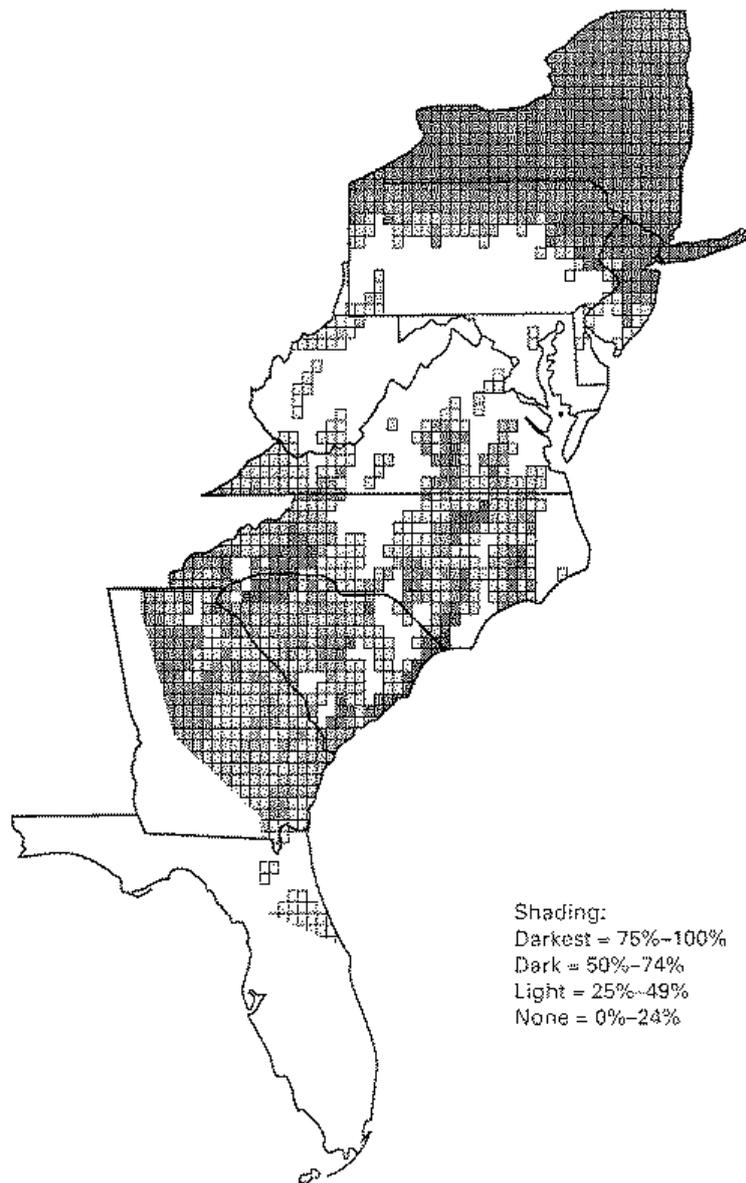


Figure 2: *Pail* responses in the Middle and South Atlantic States, processed with density estimation, nearest neighbors method.

William A. Kretschmar, Jr.

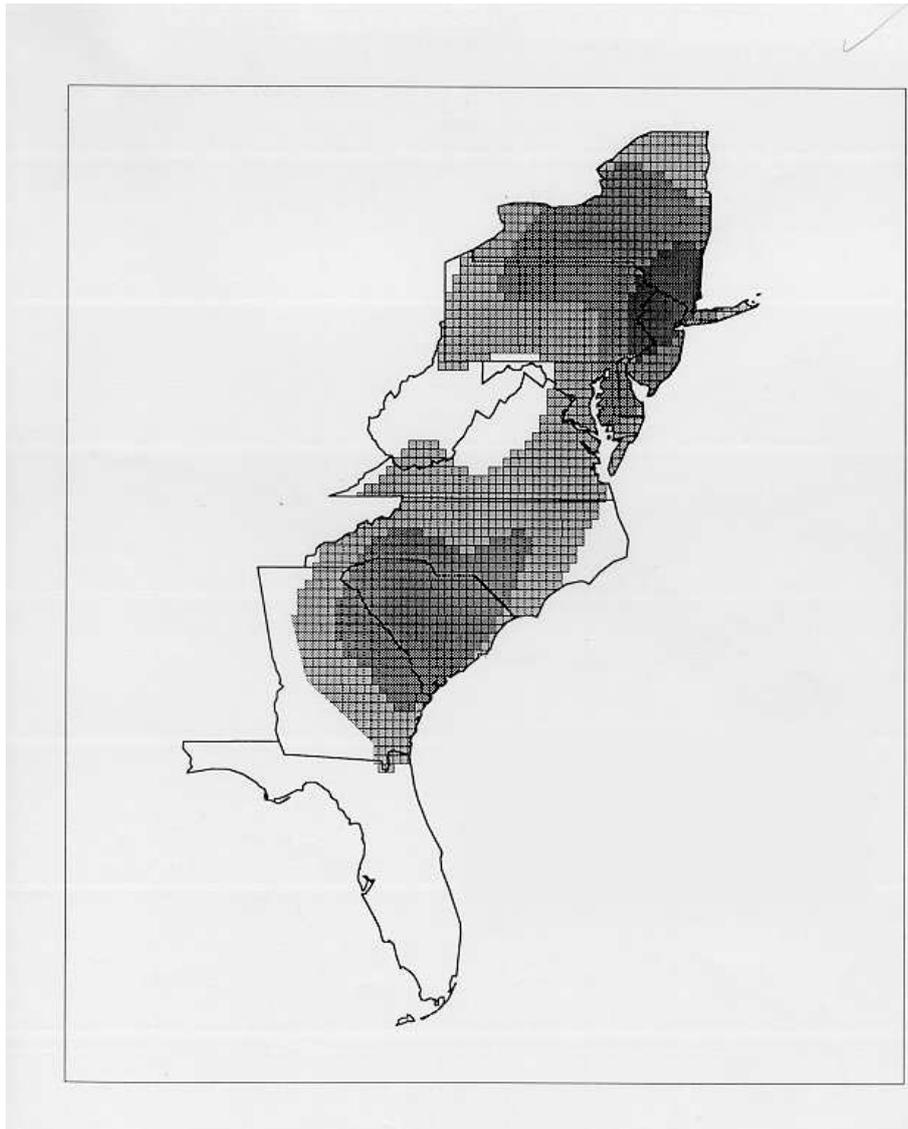


Figure 3: *Pail* responses in the Middle and South Atlantic States, processed with density estimation, kernel method for smoothing.

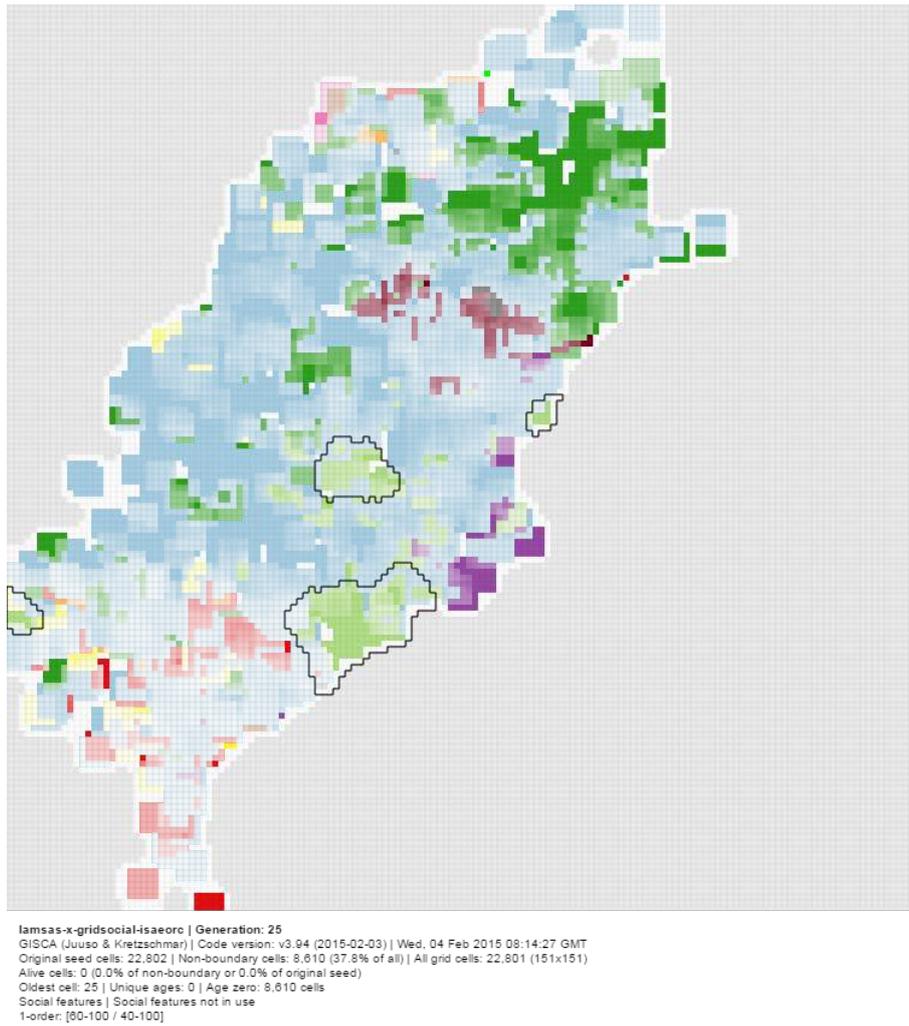


Figure 4: Multivariate density estimation, Middle and South Atlantic words for thunderstorm.

William A. Kretschmar, Jr.

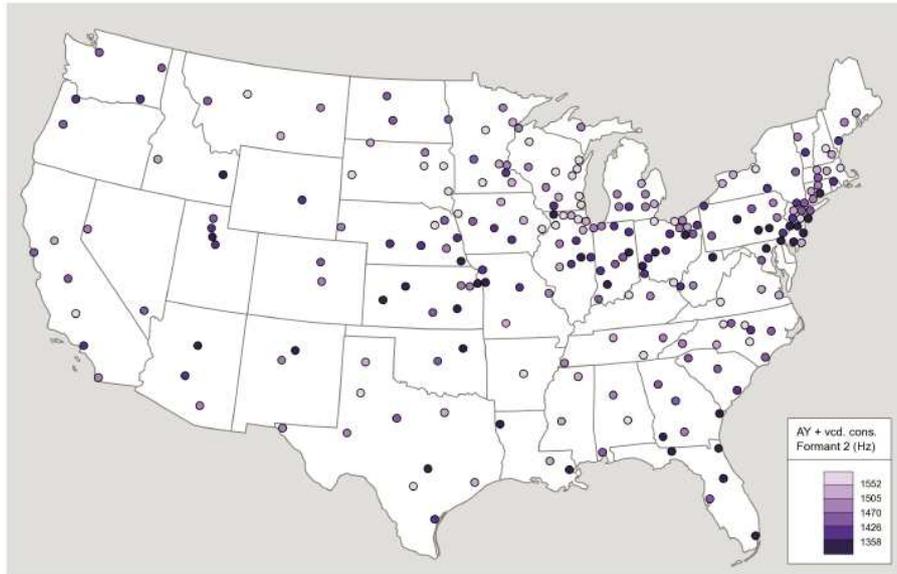


Figure 5: Raw Value Map for /ay/ before voiced consonants (e.g. *bide*) on Formant 2.

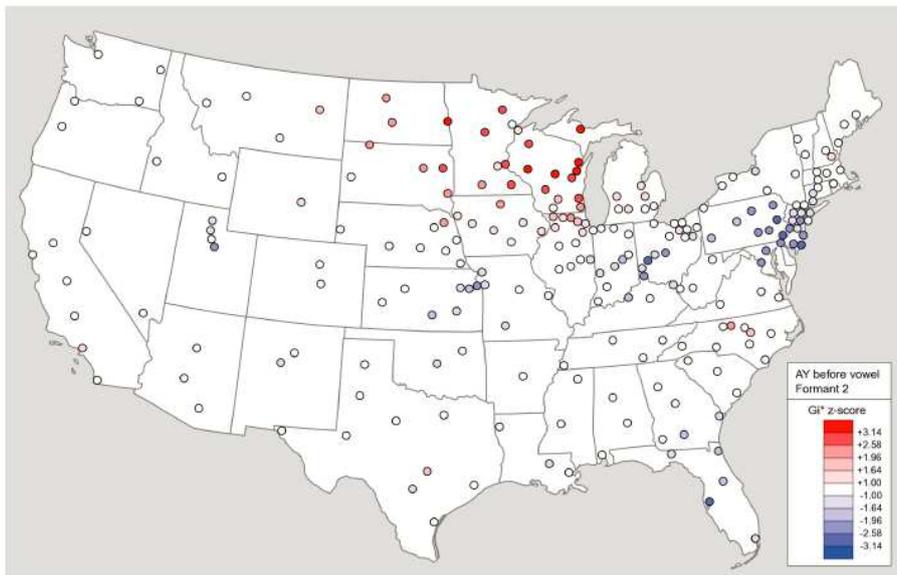


Figure 6: Local Autocorrelation Map for /ay/ before voiced consonants on Formant 2.

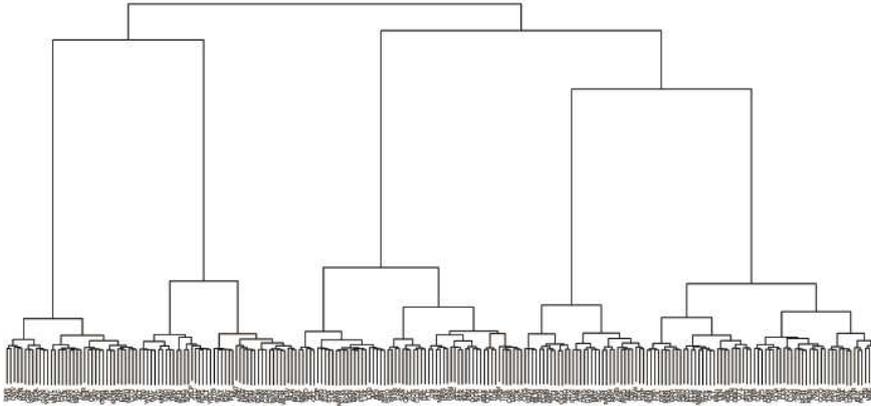


Figure 7: HCA Dendrogram based on Factors 1, 2, 3 and 4.

Grieve and colleagues interpret the HCA dendrogram as yielding five different dialect areas (Northeast, Lower Midwest, Upper Midwest, Southeast, and West), and they make attractive colored maps of the data points in each of the five HCA clusters. These areas are similar to the five areas that Labov and colleagues had named: North, Midland, South, North Central, and West. However, if we look again at the dendrogram, we see that the assignment of five areas represents a low level of agreement, quite far removed from the data point values at the bottom of the chart. It would be equally possible, based on the dendrogram, to name ten areas, since each of the five named categories originates at a bifurcation much closer to the original data points. Perhaps we should name twenty or more areas, based on the bifurcations at the next level closer to the data points. What Grieve and colleagues have done, after their modern analysis, is prefer a set of five American dialect areas, probably because that went along with what Labov had said and with what others had said before him (this is the point of Kretzschmar 2003). The last maps produced by Grieve and colleagues lose accuracy, after their careful earlier use of statistics, because they prefer a smoothed version of the data. At some point, naming too many categories from the HCA would violate the scale of the analysis, because Labov's data is not good for lower levels of scale, but the choice of five regions certainly smooths the data more than accuracy should allow.

It is certainly the case the modern linguistic maps are better than earlier linguistic maps because of their effective use of statistics. However, it is still necessary for the analyst to address all three things that make a map good: the right information, the right scale, and accuracy. Statistics alone will not make a map good, and neither is a map good because it looks similar to what previous analysts have offered. We need to maintain appropriate respect for all the aspects of the map as a model in order to make good maps.

*William A. Kretschmar, Jr.*

## References

- Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2013. Multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography* 1. 31–51.
- Kretschmar Jr., William A. 1992. Isoglosses and predictive modeling. *American Speech* 67. 227–49.
- Kretschmar Jr., William A. 2003. Mapping Southern English. *American Speech* 78. 130–149.
- Labov, William, Charles Boberg & Sherry Ash. 2006. *Atlas of North American English: phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.
- Moulton, William. 1968. Structural dialectology. *Language* 68. 451–66.

## Chapter 23

# The presentation of linguistic examples in the 1950s: an unheralded change

Charlotte Lindenberg

University of Groningen

Jan-Wouter Zwart

University of Groningen

This paper deals with a dramatic change in the presentation of linguistic examples in the linguistics literature of the twentieth century, a change coinciding (not accidentally) with the introduction of transformational generative grammar (TGG) in the 1950s. Our investigation of this change and the circumstances that gave rise to it leads us to reconsider the question of continuity and discontinuity in the history of linguistics in this crucial period, focusing on the question of the audience to which the earliest publications in TGG were directed. We argue that this was an audience of nonlinguists (information theorists and mathematical logicians) and that transformations were introduced by Chomsky, the founder of TGG, as a way to preserve, within the new paradigm of formal grammar theory, established insights of a purely linguistic nature.

### 1 The presentation of linguistic examples

The current practice (illustrated throughout this volume honoring John Nerbonne) of presenting linguistic examples (a) on a separate line, and (b) marked by continuous numbering, was virtually absent in the linguistics literature before 1950. Examples were presented either inline (Figure 1) or on a separate line without example numbering (Figure 2).

In identifying example numbering, we abstract away from the position of the number (preceding or following the example), but we do require that the numbering be continuous throughout the article (excluding numbered lists, such as Carmody 1945).

In *Language*, articles with example numbering in this sense did not appear before 1953. In *Lingua*, example numbering is introduced in 1955/1956. In neither journal

The limitation of interrogative forms to certain syntactic positions is quite common. Frequently we find them restricted to positions in the predicate of a binary sentence-type. The word-order and the plural verb-form in *who are they? what are those things?* are features of this kind. In present-day French, the non-personal *quoi?* [kwaj] ‘what?’ is scarcely ever used as actor or goal, but instead, figures as a predicate complement, appearing in the conjunct form *que* [kə], as in *qu’est-ce que c’est?* [k e s kə s eʒ] ‘what is it that this is? what’s this?’ and *qu’est-ce qu’il a vu?* [k e s k il a vyʔ] ‘what is it that he has seen? what did he see?’ In some languages the interrogative substitutes are always predicates of equational sentences, as, in Tagalog, [‘si:nu aŋ nagbi’gaj sa i’ju:] ‘who the one-who-gave to you? who gave it to you?’ or, in Menomini [awə:ɪ pə:muhnetʃ] ‘who the-one-walking-by? who is walking there?’

Figure 1: Inline presentation (Bloomfield 1933).

Tenses are not clearly distinguished. The declarative form of the verb, unless modified by the future prefix, is used to express a past action, although cases occur in which only a present can be meant.

*tiqəwgaŋno’asik* I begin to be called 94.31

In Koryak the declarative form is rarely used in narrative, while it is in common use in direct discourse.

*mai, ya’ti* halloo, have you come? Kor. 68.12  
*Valvi’mtilaʃn ti’nmim* I killed Raven-Men Kor. 20.5

In Chukchee its use in narrative is very common.

*e’nmen niki’ruʃ* then night came 36.12  
*lu’ur wɪthau’noʃ* then he began to speak 31.11

The derivative is generally used to express a present continued action, but it occurs also frequently in narrative. This use is more frequent in Koryak than in Chukchee (see § 87).

Figure 2: Separate line (Boas 1922).

does the number of articles with example numbering rise above five per year until the mid-1960s.

In both *Language* and *Lingua*, inline presentations of examples was the norm for most of the twentieth century, but there was a clear minority practice of presenting examples on separate lines without example numbering (in *Language*, around five articles per volume throughout the period 1930-1960, with example numbering kicking in only at the end of that period).

In the 1970s, two thirds of the total number of articles in *Language* had examples on separate lines, and 75% of those had example numbering. The current practice of consistent presentation of examples on separate lines with example numbering was reached in the 1980s, at least for *Language*.

We use *Language* as a test case, because its inclusive character allows us to consider the extent to which the practice of example numbering was adopted universally, regardless of theoretical affiliation. The same point could be illustrated by comparing (theory neutral) reference grammars from before 1950, where numbering was restricted to lists, and now, where continuous numbering is ubiquitous.

The universal adoption of example numbering testifies to the usefulness of the

device. Considering this, the almost complete absence of example numbering in linguistics before the 1950s is puzzling, but we will not speculate on that issue here. Our more immediate concern is to explain the introduction of example numbering in linguistics in the 1950s, and its consequences for the history of linguistics of the period, indelibly marked by the emergence of transformational generative grammar (TGG).

## 2 The influence from mathematical logic

By 1950, there was already a well-established tradition of example numbering (in the sense understood here) in the formal sciences, such as mathematics and physics. We find that in the *Annals of Mathematics*, for instance, over 60% of the articles published in 1900 had numbered formulas on separate lines, and this proportion remained constant throughout the first half of the twentieth century. Figures three and four illustrate.

THEOREM.—If  $s_1, s_2, s_3$  be the lengths of the finitely-distant infinitesimal collinear elements  $A_1B_1, A_2B_2, A_3B_3$ ; and if  $s'_1, s'_2, s'_3$  be their projections upon another right line by a pencil with any vertex; and if  $l_1, l_2, l_3$  be certain constants, viz:  $l_1 = \overline{A_1B_1} \cdot \overline{A_2B_2} \cdot \overline{A_3B_3}$ ,  $l_2 =$  etc.; and if the signs of the radicals be rightly taken, then

$$(1) \quad \left(\frac{l_1}{s'_1}\right)^{\frac{1}{2}} + \left(\frac{l_2}{s'_2}\right)^{\frac{1}{2}} + \left(\frac{l_3}{s'_3}\right)^{\frac{1}{2}} = 0.$$

So, if  $t_1, \dots, t_4$  and  $t'_1, \dots, t'_4$  be the areas of the finitely-distant infinitesimal complanar triangles  $A_1B_1C_1, \dots, A_4B_4C_4$  and of the projections, and if their constants  $m_1 = t_1 \cdot (\text{area } \overline{A_2A_3A_4})^3, m_2 =$  etc., then

$$(2) \quad \left(\frac{m_1}{t'_1}\right)^{\frac{1}{3}} + \dots + \left(\frac{m_4}{t'_4}\right)^{\frac{1}{3}} = 0.$$

So, if  $v_1, \dots, v_5$  and  $v'_1, \dots, v'_5$  be the volumes of tetrahedrons and of their projections, then

$$(3) \quad \left(\frac{n_1}{v'_1}\right)^{\frac{1}{4}} + \dots + \left(\frac{n_5}{v'_5}\right)^{\frac{1}{4}} = 0,$$

where the constants  $n_1 = v_1 \cdot (\text{vol. } \overline{A_2A_3A_4A_5})^4, n_2 =$  etc.

Figure 3: *Annals of mathematics*, 1884 (Oliver).

zu ergänzen; doch soll davon vorerst abgesehen werden, weil einerseits  $h(V, T)$  vermutlich klein ist gegenüber  $f(V)$  und  $g(T)$ , andererseits weil bei Berücksichtigung eines solchen Zusatzgliedes die Endformeln wesentlich komplizierter und unübersichtlicher werden. Aus (3) und (6) ergeben sich damit für die innere Energie  $U$  sowie für  $C_v$  die Ausdrücke:

$$(7) \quad U = F - T \left(\frac{\partial F}{\partial T}\right)_v = \Phi_0(V) + 3 N k T - 3 N k T^2 \frac{\partial g(T)}{\partial T}.$$

$$(8) \quad C_v = 3 N k \left\{1 - \frac{\partial}{\partial T} \left(T^2 \frac{\partial g(T)}{\partial T}\right)\right\}.$$

In welcher Richtung  $C_v$  bei hohen Temperaturen vom Dulong-Petitschen Wert abweicht, hängt also von  $g(T)$  ab. Es läßt sich nun qualitativ entsprechend der Überlegung von A. Eucken (a. a. O.) einsehen, daß  $g(T) > 0$  ist und mit steigender Temperatur monoton wächst. Denkt man sich ein Teilchen im Gitter schwingen, so wird es bei einem bestimmten festgehaltenen äußeren Volumen  $V$  im Zeit-

Figure 4: *Annalen der Physik*, 1935 (Damköhler).

This practice from the formal sciences had been partially adopted in philosophical circles before 1950, appearing here and there in journals like *Annalen der Philosophie* and *Erkenntnis*.

The first publications in *Language* featuring example numbering (Bar-Hillel 1953; Cherry, Halle & Jakobson 1953; Lees 1953) are clearly the result of a rapprochement of linguistics and the formal sciences, and the numbered examples are in fact mathematical formulas (see Figures five and six).

But if, instead, we have computed the various transition probabilities  $p_a(b)$ , the information conveyed by the occurrence of each successive phoneme is  $H_1(2)$ :

$$H_1(2) = -\sum p(ab) \log p_a(b) \quad (7)$$

Again, if we know the transition probabilities  $p_{ab}(c)$ :

$$H_{1,2}(3) = -\sum p(abc) \log p_{ab}(c) \quad (8)$$

Clearly these various information rates, based on different probability tables, are connected. To show this, consider equation (4); take logs of both sides and then average over all possible groups  $(ab \dots n)$ :

$$\begin{aligned} -\sum p(ab \dots n) \log p(ab \dots n) &= \\ -\sum p(ab \dots n) [\log p(a) + \log p_a(b) + \log p_{ab}(c) \dots] &\text{ or} \\ H_n = H_1 + H_1(2) + H_{1,2}(3) + H_{1,2,3}(4) \dots &\text{ bits/n-gram} \end{aligned} \quad (9)$$

This means that the information conveyed by groups of phonemes is, on the average, equal to the sum of the information obtained from each successive phoneme.

Figure 5: *Language*, 1953 (Cherry et al.).

form a string belonging to the category of sentences. That the string *Poor John sleeps* is a sentence can now be tested mechanically, without recourse to any syntactic statements, by using something like ordinary arithmetical multiplication of fractions on the INDEX-SEQUENCE corresponding to the given string, viz.

$$(1) \quad \frac{n}{\binom{n}{n}} \frac{s}{\binom{s}{n}}$$

By REDUCING the sub-sequence  $\frac{n}{\binom{n}{n}}$  to  $n$ , we obtain the FIRST DERIVATIVE

$$(2) \quad \frac{n}{\binom{s}{n}},$$

from which, by another reduction, we get the SECOND and LAST DERIVATIVE

$$(3) \quad s.$$

Let us notice immediately another important advantage of our notation: we have not only a mechanical method of testing the SYNTACTIC CONNEXITY of a

Figure 6: *Language*, 1953 (Bar-Hillel).

Around the same time, Chomsky, a linguist, published in nonlinguistic journals such as the *Journal of Symbolic Logic* (Chomsky 1953) and *IRE Transactions on Information Theory* (Chomsky 1956). In these nonlinguistic periodicals, Chomsky used example numbering in our sense, though not in his article in *Language* from around the same time (Chomsky 1955). But in his 1955 dissertation and its highly influential 1957 excerpt, *Syntactic Structures* (Chomsky 1957), example numbering was employed, and in fact not just for formulas, but for linguistic examples as well (see Figure seven and eight).

The practice of example numbering for linguistic examples was taken over by early adopters of TGG, such as Saporta (1956) and Stockwell (1960), launching a steady increase of example numbering until the current situation was reached.

### 23 The presentation of linguistic examples in the 1950s: an unheralded change

Suppose that  $a_1, \dots, a_n$  are inscriptions satisfying 'EI', and that  $A_1, \dots, A_n$  are the corresponding equivalence classes. Thus

$$(4) A_i = \hat{d}(\text{EL}x. \text{SSC}x a_i).$$

If we require merely that the syntactic categories be disjoint, we may define the syntactic categories  $\bar{A}_1, \dots, \bar{A}_n$  as

$$(5) \bar{A}_i = \hat{d}((\text{E}t)(t \in A_i. \text{SSC}xt) \cdot (y)(\text{SSC}xy. \text{E}ly. \supset y \in A_i)).$$

If we require further that no member of a syntactic category bear 'SSC' to any member of any other, then we may take them as

$$(6) \bar{A}_i = A_i \cup \hat{d}((\text{E}t)(t \in A_i. \text{SSC}xt) \cdot (y)(z)(\text{SSC}xy. \text{SSC}yz. \text{E}lz. \supset z \in A_i)).$$

In either case we can state in non-class terms the definition of 'same extended category' ('SEC'). Thus along the lines of (5) we have

$$(7) \text{'SEC}ab' \text{ for } (\text{E}t)(\text{E}u)(\text{E}l. \text{E}l u. \text{SSC}at. \text{SSC}bu. (x)(\text{SSC}ax \vee \text{SSC}bx. \text{E}lx. \supset \text{SSC}xt)).$$

From (7) we can prove that 'SEC' is transitive. We see further that two inscriptions can be in the same extended category though not related by

Figure 7: Chomsky 1953.

2.3 Second, the notion "grammatical" cannot be identified with "meaningful" or "significant" in any semantic sense. Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.

- (1) Colorless green ideas sleep furiously.
- (2) Furiously sleep ideas green colorless.

Similarly, there is no semantic reason to prefer (3) to (5) or (4) to (6), but only (3) and (4) are grammatical sentences of English.

- (3) have you a book on modern music?
- (4) the book seems interesting.
- (5) read you a book on modern music?
- (6) the child seems sleeping.

Such examples suggest that any search for a semantically based definition of "grammaticalness" will be futile. We shall see, in fact,

Figure 8: Chomsky 1957.

There is little doubt, then, that the practice of numbering examples in linguistics came about under the influence of a similar practice in the formal sciences around 1950.

## 3 Linguistics and the formal sciences

The impact of the formal sciences on linguistics in the twentieth century has been documented extensively in Tomalin (2006). Our findings regarding the practice of numbering examples are consistent with his analysis, which identifies as the origin of TGG a need felt by Chomsky to respond to the early 1950s research program (represented by Yehoshua Bar-Hillel) in which linguistics was fused with, and perhaps reduced to, mathematical logic (Tomalin 2006: 184).

However, in Tomalin's analysis, the rapprochement of linguistics and the formal sciences took place long before the origin of TGG, and is first evidenced in Bloomfield's *A set of postulates for the science of language* (Bloomfield 1926), a call for a rigorization of linguistics echoing a similar, earlier, development in mathematics. This

rigorization was achieved by introducing the axiomatic-deductive method, in which tacit assumptions are made explicit, terms are defined, and errors can be avoided (Tomalin 2006: 55). As described by Tomalin, this method did not immediately catch on, and work in this vein was not continued until Bloch (1948) and Harwood (1955) (Bloomfield himself appears to have been more interested in the question how insights from linguistics can benefit “the language of mathematics”, Tomalin 2006: 93ff).

Nevertheless, Tomalin (2006: 101) states that “Bloomfield’s [...] Formalist tendencies (whether overtly or covertly expressed), which emphasised the primacy of syntactic (rather than semantic) considerations, exerted a profound influence over a whole generation of linguists that came to maturity in the 1940s and 1950s”, such as Bloch, Hockett, Chao, Wells, Joos and Zellig Harris, who Chomsky studied linguistics with between 1947 and 1951 (Barsky 1997).

I refer to Matthews (1993: esp. pp. 111-128) for careful discussion of the form this influence on the post-Bloomfieldians took, centering on formalized discovery procedures for the structure of sentences based on the distribution of their constituents. Suffice it to say here that in Tomalin’s analysis, formalism in the post-Bloomfieldians and in Chomsky is essentially a delayed response to Bloomfield’s initial call, a derived effect of the first thrust of the impact of the formal sciences on linguistics.

From this perspective it is interesting to note that neither Bloomfield nor any of the post-Bloomfieldians, up to and including Harris, ever employed the device of continuous example numbering in their linguistic publications. But Chomsky did, right from the start, and we have to wonder why.

## 4 Chomsky and his audience

As Tomalin (2006) describes in detail, advances in mathematical logic in the first half of the twentieth century continued to impact the linguistics community, and he quotes Harris as calling the distributional methods of (post-Bloomfieldian) linguistics “hospitable” to the mathematical description of language. In this context, we must mention Carnap’s (1934/1937) *The logical syntax of language*, Ajdukiewicz’s (1936) *Syntactic connexion*, and Post’s (1944) work on recursion and generative procedures. As Tomalin (2006: 103-105) notes, Harris was thoroughly familiar with these developments and understood their relevance to linguistics.

On the other hand, Harris also “was keen to stress the differences that distinguish linguistics and logic” (Tomalin 2006: 105), emphasizing that logicians like Carnap avoid the analysis of existing languages, which are the prerogative of linguistics. We find a similar concern expressed by Hjelmslev (1948: 33), also an early adopter of the formalist approach, but in a European context, who states that “logistic language theory has been carried out without any regard to linguistics, and it is obvious that logicians, while constantly talking about language, are neglecting in a somewhat indefensible way the results of the linguistic approach to language”.

We submit that Chomsky in his earliest work, rather than being carried away on the new wave of algebraic linguistics, addresses this very problem of mathematical

### 23 *The presentation of linguistic examples in the 1950s: an unheralded change*

logic being unable to accommodate traditional linguistic insights and achievements. It is here that transformations in the Chomskyan sense make their appearance (see section 5).

This entails that Chomsky's audience, in his very first publications, was not, or not in the first place, an audience of linguists, but an audience of mathematical logicians and information theorists who, in the wake of Carnap, were taking on natural language as their object of inquiry.

From this perspective it is understandable that Chomsky sought a podium like the *Journal of Symbolic Logic* and the *IRE Transactions on Information Theory* for his first publications. Moreover, it makes sense that he addressed the readers in a format they were familiar with, including the use of continuously numbered formulas and examples.

This reconstruction presupposes a certain amount of linguistic traditionalism on the part of Chomsky. While this may be surprising in view of his role as the founder of TGG, Chomsky's respect for tradition in especially morphological analysis has been amply documented in Matthews (1993), showing that the breakdown into lexical and functional elements crucial to a key argument in *Syntactic Structures* was standard procedure in post-Bloomfieldian thinking about morphology (see also Zwart 1994). But Chomsky's position vis-à-vis linguistic tradition is also evident from the discussion transcripts of the Third Texas Conference of Problems of Linguistic Analysis in English ("The process that I use for investigating language is the one that I was taught," Chomsky 1962: 174) and the Ninth International Congress of Linguists at MIT in August 1962. At the MIT conference, Chomsky was addressed by E. M. Uhlenbeck who suggested an analysis of *The man hit the ball* via a left-to-right parsing procedure yielding *the man hit* as a constituent. Chomsky rejects this suggestion immediately by referring to the venerable NP-VP (subject-predicate) analysis (Lunt 1964: 983). Chomsky's frequent references to Jespersen and Humboldt throughout his career also underscore his respect for linguistic tradition.

In personal communication (March 31, 2015), Chomsky described his displeasure with the general tendency in post-war American academic circles to overhaul science with disregard for European accomplishments. As Chomsky noted, this affected European scientists who had immigrated to the United States before or during the war and had great problems gaining acceptance on the American scientific and cultural scene, citing Roman Jakobson as a telling example. The outcome of the war led to a certain triumphalism, in which information theory was the name of the game, and people working in that field were ignorant about language as studied in the structuralist tradition (see also Chomsky 1975: 39-40).

This is the context in which Chomsky's public discussion with Yehoshua Bar-Hillel, documented in *Language* (Bar-Hillel 1954; Chomsky 1955; see also Tomalin 2006: 125f and Hiorth 1974) must be understood. Bar-Hillel worked with Carnap as a postdoc in 1950 and headed the first machine translation lab at MIT from 1951 on (Bar-Hillel 1964: 5-6). Chomsky and Bar-Hillel were "extremely close" (Barsky 1997: 54), and Bar-Hillel appears to have been one of the few colleagues who paid attention to Chomsky's early work, including his 1951 BA-thesis (ibid.). In his reminiscences Bar-Hillel

(1964: 16) calls Chomsky “the founder of algebraic linguistics and by far the best man in this exciting new field”, adding, tellingly, “though Chomsky himself would probably claim that it is not really new at all, but just plain good old linguistics, pursued with the best means available at our time, which happens to be of algebraic nature”. Nevertheless, Chomsky and Bar-Hillel differed on the value of artificial language research for natural language linguistics, which appears to have been at the core of the discussion (Tomalin 2006: 136-137) (we are ignoring here another important point of disagreement, regarding the value of logic for the study of meaning in natural language).

At the same time, Chomsky felt alienated from the main trend in post-Bloomfieldian linguistics, championed by Harris (1951), of establishing discovery procedures (Matthews 1993: 135f; Tomalin 2006: 149f), and in this again he and Bar-Hillel came to share the same view (Tomalin 2006: 71).

We suggest that this complex situation, in which Chomsky felt at the same time a close affinity with Bar-Hillel, yet felt compelled to defend linguistic analysis against appropriation by the mathematical logicians and information theorists that his friend represented (cf. Tomalin 2006: 184; see also Chomsky 1975: 40), determined both Chomsky’s publication strategy as well as his chosen style of presentation. This style turned out to be the one common in mathematical logic, in particular in the use of numbered examples.

## 5 The origin of transformations

We have established that Chomsky directed his earliest publications at an audience of mathematical logicians and information theorists, and that this orientation provides a natural explanation for his adoption of example numbering. We believe this casts a new light on the origin of transformations in TGG. *Transformation* was a familiar term to both mathematical logicians (Carnap) and post-Bloomfieldian linguists (Harris), but as employed by Chomsky their function, of providing phrase structure grammars with the necessary linguistic sophistication, was entirely new.

As is well known, Chomsky’s main argument in his early work, including *Syntactic Structures*, was that “phrase-structure grammars are inadequate in strong generative capacity” and that “linguistic theory requires a new and more abstract level of description, the level of grammatical transformations, and a richer concept of grammar” (Chomsky 1975: 8). The resulting transformational generative grammar (emphasizing *transformational*) is able to “express rather subtle aspects of the form and interpretation of sentences” (ibid.).

The two-step process of generation and transformation was familiar to both mathematical logicians/information theorists and post-Bloomfieldian linguists. Carnap in his *The Logical Syntax of Language* (1937) describes language as a calculus, the rules of which determine both the formation (“the syntactical rules in the narrower sense”, Carnap 1937: 2) and the transformation of linguistic expressions; the transformations are just rules of logical inference among expressions (ibid.; see also Tomalin 2006: 159). In linguistics, too, the term transformation was familiar, from Harris (1951), al-

though Bar-Hillel (1954) was quick to point out that Harris' transformations were actually formation rather than transformation rules in the Carnapian sense. That the description of language structure is essentially a generative system (in the sense of Post 1944) is more or less implicit in Harris (1951, e.g. 1951: 372-373), and was apparently common thinking among structuralist linguists in the mid-1950s (Matthews 1993: 134). So here, too, the generation-transformation dichotomy was a familiar concept.

Chomsky has later (1975: 43) expressed regret for using Harris' term transformation in this different sense, potentially confusing his linguistic audience (see Tomalin 2006:159ff and Nevin 2009 for a comparison of transformations in Harris and Chomsky). But from the perspective that Chomsky was writing primarily for an audience of mathematical logicians and information theorists, such confusion could hardly have been expected (and in fact did not arise). For Chomsky introduced his transformations as "supplementary rules" in response to demonstrated limitations of phrase structure (e.g. Chomsky 1957: 44), putting a damper on expectations among his audience that the expressions of natural language could be derived just by rules of phrase structure grammar in any simple and explanatory way.

It is important to note that transformations are called upon by Chomsky to describe the most elementary properties of natural languages, which were well understood among linguists, but perhaps less so among representatives of mathematical logic and information theory. These properties include conditions on conjunction reduction, intricacies of verbal morphology, active-passive relations, negation, question formation, *do*-support, ellipsis, auxiliatation, nominalization, pronominalization, etc. (Chomsky 1957: chapters 5 and 7). Thus, while we agree with Tomalin that the origin of TGG lies in the association of linguistics with the formal sciences (2006: 186), the crucial part in that association was the introduction of transformational analysis, championing the cause of 'good old linguistics' against ill-informed logicians and information theorists.

## 6 Conclusion

We have argued for a second and decisive moment in the history of linguistics in which the language sciences underwent the influence of formal sciences, in particular mathematical logic (following the first instance, marked by Bloomfield 1926). Unlike the first instance, this renewed rapprochement changed the face of linguistics forever, in leading to the introduction of continuous example numbering in linguistic writing. This introduction of example numbering came about in linguistic publications directed at an audience of mathematical logicians and information theorists, and is first attested in journals from those fields, and then in articles in linguistics journals featuring mathematical formulas. We submit that the practice of numbering examples had its origin in Chomsky writing for an audience of nonlinguists, in a style with which his intended audience was more familiar.

In this context, we observe a parallel between Chomsky's use of example numbering and his introduction of transformations (in his sense). Both transformations and

numbered examples were formal devices familiar to his intended audience of logicians and information theorists, and in Chomsky's early work both devices served to infuse the current (mathematical) discourse with pure linguistic content. Just as transformations served as a conduit for linguistic sophistication in the new field of algebraic linguistics, so natural language examples assumed the place of mathematical formulas, presented on a separate line and continuously numbered, as they have been ever since.

## Acknowledgements

This paper benefited from discussion with Noam Chomsky, Jan Koster, and Frans Zwarts, and was earlier presented at the TABU-dag in Groningen on June 13, 2013.

## Source of illustrations

- Figure 1:** Bloomfield 1933: 260.
- Figure 2:** Bogoras 1922: 772.
- Figure 3:** Oliver 1884: 40.
- Figure 4:** Damköhler 1935: 4.
- Figure 5:** Cherry, Halle & Jakobson 1953: 45.
- Figure 6:** Bar-Hillel 1953: 48.
- Figure 7:** Chomsky 1953: 251.
- Figure 8:** Chomsky 1957: 15.

## References

- Adjukiewicz, Kazimierz. 1936. Die syntaktische Konnexität. *Studia Philosophica* 1. 1–27.
- Bar-Hillel, Yehoshua. 1953. A quasi-arithmetical notation for syntactic description. *Language* 29. 47–58.
- Bar-Hillel, Yehoshua. 1954. Logical syntax and semantics. *Language* 30. 230–237.
- Bar-Hillel, Yehoshua. 1964. Introduction. In *Language and information*, 1–16. Reading: Addison Wesley.
- Barsky, Robert F. 1997. *Noam Chomsky: a life of dissent*. Cambridge: MIT Press.
- Bloch, Bernard. 1948. A set of postulates for phonemic analysis. *Language* 24. 3–46.
- Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2. 153–164.
- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt. Image from the British edition, London: George Allen and Unwin, 1935.
- Bogoras, W. 1922. Chukchee. In F. Boas (ed.), *Handbook of American Indian languages*, vol. 2, 631–903. Washington: Bureau of American Ethnology.
- Carmody, Francis J. 1945. Syntax of the verb IS in Modern Scottish Gaelic. *Word* 1. 162–187.

23 *The presentation of linguistic examples in the 1950s: an unheralded change*

- Carnap, Rudolf. 1937. *The logical syntax of language*. London: Routledge & Kegan Paul (Translation of *Die Logische Syntax der Sprache*, Vienna: 1934).
- Cherry, E. Colin, Morris Halle & Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language* 29. 34–46.
- Chomsky, Noam. 1953. Systems of syntactic analysis. *Journal of Symbolic Logic* 18. 242–256.
- Chomsky, Noam. 1955. Logical syntax and semantics: their linguistic relevance. *Language* 31. 36–45.
- Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions of Information Theory* 2. 113–124.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
- Chomsky, Noam. 1962. A transformational approach to syntax. In *Third Texas conference on problems of linguistic analysis in English*, 124–186. Austin: The University of Texas.
- Chomsky, Noam. 1975. *The logical structure of linguistic theory*. Cambridge: MIT Press.
- Damköhler, G. 1935. Zur Theorie des festen Körpers bei hohen Temperaturen mit besonderer Berücksichtigung der Temperaturabhängigkeit von  $C_v$ . *Annalen der Physik* 416 (1). 1–30.
- Harris, Zellig. 1951. *Methods in structural linguistics*. Chicago: The University of Chicago Press.
- Harwood, F. W. 1955. Axiomatic syntax: the construction and evaluation of a syntactic calculus. *Language* 31. 409–413.
- Hiorth, Finngeir. 1974. *Noam Chomsky: linguistics and philosophy*. Oslo: Universitetsforlaget.
- Hjelmslev, Louis. 1948. Structural analysis of language. *Studia Linguistica* 1. 69–78. Reprinted in *Travaux du Cercle Linguistique de Copenhague* 12 (1959), 27–35.
- Lees, Robert B. 1953. The basis of glottochronology. *Language* 29. 113–127.
- Lunt, Horace G. (ed.). 1964. *Proceedings of the ninth International Congress of Linguists*. The Hague: Mouton.
- Matthews, Peter H. 1993. *Grammatical theory in the United States from Bloomfield to Chomsky*. Cambridge: Cambridge University Press.
- Nevin, Bruce E. 2009. More concerning the roots of Transformational Generative Grammar. *Historiographia Linguistica* 36. 461–481.
- Oliver, J. E. 1884. A projective relation among infinitesimal elements. *Annals of Mathematics* 1. 40–41.
- Post, Emil. 1944. Recursively enumerable sets of positive integers and their decision problems. *Bulletin of the American Mathematical Society* 50. 284–316.
- Saporta, Sol. 1956. A note on Spanish semivowels. *Language* 32. 287–290.
- Stockwell, Robert P. 1960. The place of intonation in a generative grammar of English. *Language* 36. 360–367.
- Tomalin, Marcus. 2006. *Linguistics and the formal sciences: the origins of generative grammar*. Cambridge: Cambridge University Press.
- Zwart, Jan-Wouter. 1994. Het ontstaan van I' en C'. *Gramma/TTT* 3. 55–70.



## Chapter 24

# Gravity, radiation, and dialectometry

Robert Malouf

San Diego State University

Gravity laws have a long history in models of human mobility. Trudgill (1974) proposed a version of the gravity model to account for the diffusion of dialect variation, but other researchers found gravity to be a poor predictor of linguistic change. In a large-scale evaluation of the gravity law across a whole dialect region, Nerbonne & Heeringa (2007) find support for some but not all of its predictions. This contribution compares the gravity model to an alternative proposed by Simini et al. (2012) which is inspired not by gravity but by the physics of radiation and absorption. In computational simulations, the radiation model appears to be a better match than the gravity model for the general patterns found in dialect diffusion. This suggests as a future direction for dialectometric work the comparison of quantitative models of human mobility and their consequences for understanding linguistic variation.

## 1 Introduction

Gravity models have a long history in geography and economics as a model of human mobility (e.g., Ravenstein 1885; Zipf 1946). Taking a form familiar from Newton's law of universal gravitation, gravity models in human geography predict that the degree of interaction between two locations  $i$  and  $j$  follows the rule:

$$I_{ij} = C \frac{m_i m_j}{d_{ij}^2} \quad (1)$$

where  $m_i$  and  $m_j$  are the populations of  $i$  and  $j$ ,  $d$  is the distance between them, and  $C$  is a constant. This makes intuitive sense in a general way: the opportunities for connections between locations increase as either of their populations increase, but decrease rapidly with distance.

Gravity laws have the appeal of being both conceptually and mathematically simple (though in its general form, the gravity law has a number of parameters that can be difficult to estimate). It also makes the correct empirical predictions (at least approximately) for a broad range of phenomena, including commuting patterns, migrations, airline traffic, and commodity flows. More recently, gravity laws have also

*Robert Malouf*

been applied to modeling mobile phone calls (Krings et al. 2009) and social media use (Noulas et al. 2012).

If social contact via communication, migration, and trade is the engine that drives the diffusion of dialect variation, then it stands to reason that dialect differences might reflect patterns of interaction predicted by the gravity law. Bloomfield (1933) offered the insight that dialect differences reflect differences in the ‘density of communication’. Trudgill (1974) expanded on this by applying a gravity law as a measure of mutual influence between cities in modeling the diffusion of London dialect features into East Anglian English.

Since that introduction into dialect studies, the gravity model has received a number of challenges and modifications. Boberg (2000), e.g., looks at diffusion of two linguistic variants across the US/Canada border and finds that (1) is a poor fit. In this and a number of other case studies, the gravity model finds mixed support and compares unfavorably with models that take social prestige and identity into account.

## 2 Gravity in dialectometry

In their study of Dutch dialects, Nerbonne and colleagues developed and extended a set of dialectometric methods for large-scale measurement of linguistic variation across space and time (Nerbonne, Heeringa & Kleiweg 1999; Heeringa 2004; Heeringa et al. 2006; Nerbonne 2009; Nerbonne & Heeringa 2010). By computing edit distance between comparable forms collected at a large number of locations, they are able to quantify the degree and nature of dialect variation across a region.

Nerbonne et al.’s metrics aggregate many dimensions of linguistic variation into a single score, which makes them ideal for evaluating the predictions of the gravity model. This is in contrast to studies like Trudgill’s and Boberg’s, which focus on one or two hand-selected linguistic variables.

Nerbonne et al. tested how well the gravity law fits against real-world (Nerbonne & Heeringa 2007; Heeringa et al. 2007) and simulated (Nerbonne 2010) dialect areas and made two findings. First, they determined that, consistent with the gravity model, geographic distance is the most important factor in predicting dialect difference. It seems that, when taken in the aggregate, linguistic variation is not as dependent on purely social factors as the post-Trudgill work might have suggested. They do find that the relationship with distance is linear rather than quadratic. However, this is consistent with the generalized form of the gravity law which is often used in studies of human mobility.

More surprisingly, though, they conclude that the influence of population on dialect difference is very small and, if anything, points weakly in the wrong direction. Sites with larger populations are slightly more likely to differ than sites with smaller populations, contrary to (1). This would seem to be straightforward falsification of the gravity model for dialect variation, but it is also puzzling. The gravity model has been successfully used to describe many different phenomena relating to social and economic connections between locations. Why should linguistic diffusion work differently?

### 3 Radiation model

A possible answer to that question comes from the observation that gravity models have come under challenge in other social domains as well. Responding to the numerous theoretical and empirical shortcomings of the gravity model, Simini et al. (2012) propose an alternative, inspired not by gravity but by the physics of radiation and absorption. They start with a model of people commuting to work and a simple assumption: commuters will search for a job as close to home as possible, and will choose the closest job which provides a benefit that is greater than that which is available in their own city. From this, they derive the fundamental equation of the radiation law, which gives the probability  $p_{ij}$  of a commuter traveling from city  $i$  to city  $j$ :

$$p_{ij} = \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})} \quad (2)$$

Here,  $m_i$  and  $m_j$  are the populations of  $i$  and  $j$ , and  $s_{ij}$  is defined as follows: if  $r_{ij}$  is the distance between city  $i$  and city  $j$ , then  $s_{ij}$  is the total population (not including the populations of  $i$  and  $j$ ) living within a circle of radius  $r_{ij}$  centered on  $i$ .

One notable property of the radiation law is that it  $p_{ij}$  is not directly dependent on distance. It depends instead on  $s_{ij}$ , the population within a certain distance. If the population density is uniform, then  $s_{ij}$  is proportional to the area of a circle of radius  $r_{ij}$  and so proportional to  $m_i d_{ij}^2$ . In general, though, the relationship between  $d$  and  $s$  depends entirely on the way the population is distributed. This allowing the radiation model to, e.g., model differences in commuting patterns in rural and urban areas, something that the gravity model had not been able to give a general account of.

Simini et al. (2012) go on to validate their radiation law empirically, showing that is a better predictor of human mobility than the gravity law based on a broad range of empirical tests, concluding: “We find that the radiation model offers an accurate quantitative description of mobility and transport spanning a wide range of time scales (hourly mobility, daily commuting, yearly migrations), capturing diverse processes (commuting, intra-day mobility, call patterns, trade), collected via a wide range of tools (census, mobile phones, tax documents) on different continents (America, Europe)” (p. 97).

While the radiation model is not without critics (e.g., Masucci et al. 2013), it has held up as a superior predictor of inter-city flows of people, goods, and communication than the older gravity model. Perhaps it would be a better model of the inter-city flow of linguistic variation as well.

### 4 Simulations

Following Nerbonne (2010), I performed simple computational simulations to compare the predictions of the gravity and radiation models for linguistic diffusion.<sup>1</sup> For

<sup>1</sup> The simulations were implemented in Python and are available at <http://github.com/rmalouf/gravity>.

each of these experiments, an artificial distribution of cities was created. The first step was to create cities of varying populations of agents. The population was seeded with 500 agents, each in its own city. Next, additional agents were added one at a time by selecting an existing agent and adding the new agent to the selected agent's city until the total population of agents reaches 100,000. This is essentially Albert & Barabási's (2002) preferential attachment algorithm, and it produces a collection of cities whose populations are distributed following an inverse power law.

The next step is to locate these cities in space. The city with the largest population is taken as the 'capital' (the source of linguistic innovations) and placed at the center. The remaining cities are then distributed uniformly within a unit circle centered on the capital.

Agents are represented as a vector of 100 binary values representing linguistic variables (Holman et al. 2007; Nerbonne 2010). Agents in the capital have the value 1 (representing the innovative value) for all variables while all other agents have the value 0. Successive iterations of the simulation model the interaction between a traveler from the capital and a resident of one of the other cities. On each iteration, a city other than the capital is selected as a destination and then an agent is selected in that city. On contact with a resident of the capital, each of the selected agent's linguistic variables have a 1% chance of being set to 1 (note that some may have already been set to 1 on a previous iteration). After 50,000 iterations, the aggregate linguistic difference between the dialects spoken in each city is calculated. A city's 'dialect' is the normalized sum of the vectors of the agents living in it, and the difference between two dialects is 1 minus the dot product of the cities' dialect vectors. This ranges from 0.0 for dialects which share all properties to 1.0 for dialects that share none.

In the first simulation, destination cities were selected for agents with a probability proportional to a modified gravity law:

$$I_{ij} = \frac{m_i m_j}{d_{ij}} \quad (3)$$

This preserves the dependency on population, but following Nerbonne & Heeringa (2007) replaces the quadratic distance term with a linear one. In the second simulation, the probability that a destination city would be selected is given by the radiation model (1).

Figures 1 and 2 give the results of the simulation. The plots show the relationship, for all of the cities except the capital, between either geographic distance from the capital or population and the difference between the local dialect and the capital's dialect.

As Figure 1 indicates, the radiation model predicts that linguistic difference should increase with geographic distance, which is consistent with Nerbonne & Heeringa's (2007) findings. The predictions for the gravity model are not quite as clear, but while the trend is not very strong there is an increase in difference as distance increases. For both models, the curves are reminiscent in their general shape of what Nerbonne (2010) calls Séguy's Law: linguistic differences increase with distance to a point beyond which further increases in distance correlate with only small increases

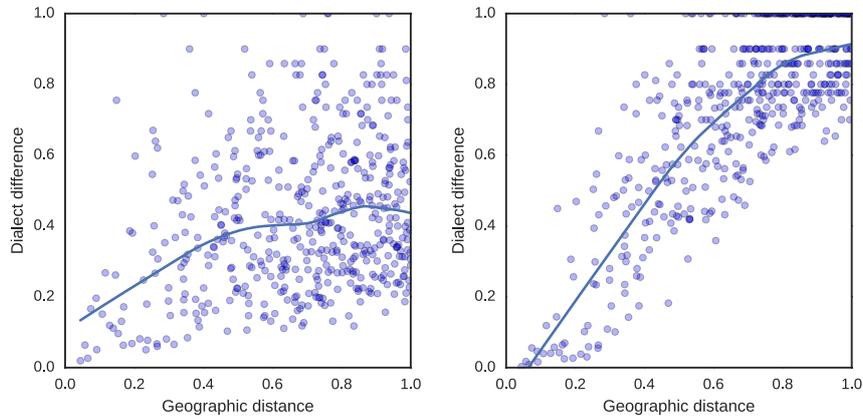


Figure 1: Geographic distance from capital vs. dialect difference, gravity (left) and radiation (right) models.

in linguistic difference. Séguý's Law leads to a sublinear increase in the dialect differences as distance increases, something both the gravity and the radiation models appear to predict.

Figure 2 shows the results for population. Based on the simulation, we can see that

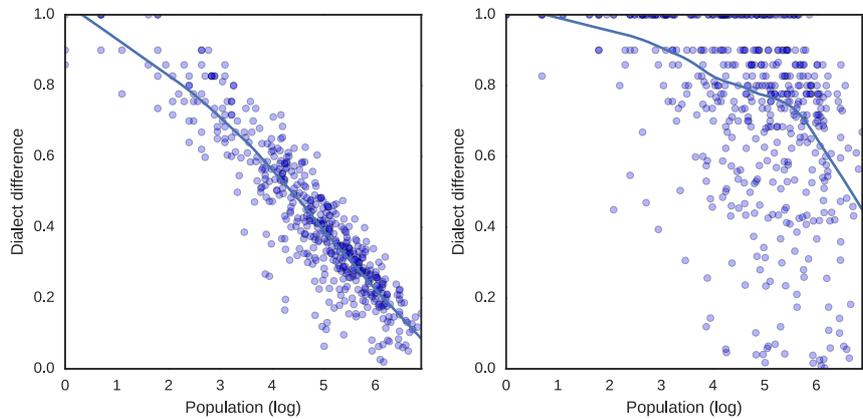


Figure 2: Population vs. dialect difference, gravity (left) and radiation (right) models.

the gravity model as specified in (1) predicts a strong inverse relationship between population size and linguistic difference. This is exactly what Nerbonne & Heeringa (2007) did **not** find in their study of Dutch dialects. For the radiation model, on the other hand, the relationship is very weak. The local regression line (fit by LOWESS) indicates a bit of a downward trend. But, many of the cities with the largest populations

*Robert Malouf*

also have the greatest linguistic distance from the capital's dialect.

## 5 Conclusions

What these results show is that at least at an impressionistic level the radiation model is a better match for aggregate patterns seen in dialect variation than the gravity model. The next step would be to go beyond simulation results and to evaluate the radiation model against the kind of dialect variation data that Nerbonne & Heeringa (2007) used to test the gravity law. The radiation model has succeeded in other areas of human mobility where the gravity model has not. If it makes the correct predictions for the diffusion of linguistic variation as well, this would add confirmation to Bloomfield's (1933) hypothesis concerning the density of communication. But, if the radiation model does not fare any better than the gravity law, that would raise profound questions about the nature of the mechanisms underlying dialect diffusion and their relation to other social processes.

More generally, we can say that without dialectometric methods these questions could not even be asked. As a complement to detailed description of ongoing linguistic changes in their social context, large-scale aggregate comparison of dialects over an entire region allow us to relate language variation and change to other aspects of human dynamics.

## References

- Albert, Réka & Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(1). 47.
- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt & Company.
- Boberg, Charles. 2000. Geolinguistic diffusion and the US-Canada border. *Language variation and change* 12(01). 1–24.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Rijksuniversiteit Groningen PhD thesis.
- Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In John Nerbonne & Erhard Hinrich (eds.), *Proceedings of the Workshop on Linguistic Distances*, 51–62.
- Heeringa, Wilbert, John Nerbonne, Renée van Bezooijen & Marco René Spruit. 2007. Geografie en inwoneraantallen als verklarende factoren voor variatie in het nederlandse dialectgebied. *Nederlandse Taal- en Letterkunde* 123(1). 70–82.
- Holman, Eric W., Christian Schulze, Dietrich Stauffer & Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology* 11. 393–421.
- Krings, Gautier, Francesco Calabrese, Carlo Ratti & Vincent D. Blondel. 2009. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* 2009(07). L07003.

- Masucci, A. Paolo, Joan Serras, Anders Johansson & Michael Batty. 2013. Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. *Physical Review E* 88(2). 022812.
- Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.
- Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3821–3828.
- Nerbonne, John & Wilbert Heeringa. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In Sam Featherston & Wolfgang Starkefeld (eds.), *Roots: linguistics in search of its evidential base*, 267–298. Berlin: Mouton de Gruyter.
- Nerbonne, John & Wilbert Heeringa. 2010. Measuring dialect differences. In Jürgen Erich Schmidt & Joachim Herrgen (eds.), *Language and space: theories and methods*, 550–566. Berlin: Mouton de Gruyter.
- Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff & Joseph Kruskal (eds.), *Time warps, string edits and macromolecules: the theory and practice of sequence comparison*. CSLI Publications.
- Noulas, Anastasios, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil & Cecilia Mascolo. 2012. A tale of many cities: universal patterns in human urban mobility. *PloS one* 7(5). e37027.
- Ravenstein, E. G. 1885. The laws of migration. *Journal of the Statistical Society of London* 48(2). 167–235.
- Simini, Filippo, Marta C. González, Amos Maritan & Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484(7392). 96–100.
- Trudgill, Peter. 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2). 215–246.
- Zipf, George K. 1946. The  $P_1 P_2 / D$  hypothesis: on the intercity movement of persons. *American Sociological Review* 11(6). 677–686.



## Chapter 25

# Exploring the role of extra-linguistic factors in defining dialectal variation patterns through cluster comparison

Simonetta Montemagni

Institute for Computational Linguistics “Antonio Zampolli”

Martijn Wieling

University of Groningen

This paper contributes to two open issues in the dialectometric literature, i.e. i) whether and how patterns of linguistic variation are influenced by extra-linguistic features such as the geomorphology of the area, or cultural, administrative and political boundaries, and ii) whether and how the influence of extra-linguistic factors remains stable across linguistically-grounded partitions of data. To investigate these issues, a case study focusing on lexical variation has been carried out on a regional lexical atlas of Tuscan dialects. A variety of extra-linguistic features was taken into account, whose impact and role has been evaluated with respect to both the whole dialectal dataset and across different semantic fields.

## 1 Introduction

In the variationist literature, it is a widely acknowledged fact that (bundles of) isoglosses correlate fairly closely with non-linguistic boundaries: in other words, language is reported to correlate with other aspects of culture (Chambers & Trudgill 1998). What underlies observed correlations remains however an important issue, which is still worth being explored. According to the notion by Bloomfield (1933) of “density of communication”, linguistic variation depends primarily on the frequency of communication within the network of speakers of a given language, which is in turn influenced by a variety of extra-linguistic factors, ranging from physical features and population density of the investigated area, to cultural and demographic patterns as well as administrative and/or political boundaries. Due to their role in

limiting or promoting communication across space, extra-linguistic factors such as these can be seen as influencing linguistic variation diatopically and diachronically.

In traditional dialectology, there is no obvious way to explore this relationship beyond superficial and impressionistic observations. In dialectometric studies, the issue of the influence of extra-linguistic features on linguistic variation has been tackled for what concerns geography and population genetics.

Whether and to what extent geographic distance influences aggregate linguistic differences among language varieties has been investigated since early dialectometry (Heeringa & Nerbonne 2001; Gooskens 2005). In these studies, geographical distance, including derivatives of the notion, such as travel time, is regarded as an operationalization of the chance of social contact. Geography has been shown to correlate strongly with linguistic variation. This type of analysis, carried out for typologically distant languages (Bantu, Bulgarian, Dutch, English, German, Norwegian) with respect to phonetic variation, showed that geography accounts for 16% to about 37% of the linguistic variation (Nerbonne 2013).

On the population genetics front, Manni et al. (2008) conclude that the social contacts reflected by dialect varieties do not seem to be related to the demographic history of the populations speaking those dialects. Linguistic and genetic patterns of variation (the latter reconstructed from people's surnames) turned out to be different, even if both are strongly conditioned by geography.

Despite the contrasting conclusions, these studies share the methodology of analysis, i.e. the comparison is carried out with respect to computed distances, be they linguistic, genetic or geographic. The variety of extra-linguistic factors potentially influencing linguistic variation, however, cannot be always modeled in terms of distances. Consider, for instance, the case of physical features (e.g., mountain ranges or river basins) of the investigated area, or its administrative or political organization (e.g., borders of the state or smaller administrative subdivisions). In these cases, the comparison cannot be carried out with respect to distances, but it is rather concerned with clusterings of the investigated locations, based e.g., on their belonging to a given state or other administrative unit, or to their being located in the valley of a given river. To our knowledge, dialectometric studies so far did not tackle this kind of analysis.

In this paper, we will investigate the relationship between linguistic and non-linguistic boundaries through a clustering comparison method, with the final aim of trying to reconstruct role and impact of extra-linguistic features in shaping patterns of linguistic variation. To this specific end, we will use an information theoretic criterion for comparing partitions of the same data set, proposed by Meilă (2003). The criterion is called VARIATION OF INFORMATION (VI), and quantifies the "distance" between the two clusters. In contrast to the modified Rand Index (Hubert & Arabie 1985) used by Prokić, Wieling & Nerbonne (2009) to compare partitions, the VI measure is a true metric (i.e. satisfying the triangle inequality).

Let us focus now on linguistic variation patterns observed with respect to a given area. The question which naturally arises at this point is whether they remain stable between and within levels of linguistic description. In dialectometry, correlation

studies focusing on dialectal variation recorded with respect to distinct linguistic description levels (e.g., phonetic, morphological, lexical) have been carried out with different methodologies and with respect to different dialects. See, for example, Goebel (2005), Spruit, Heeringa & Nerbonne (2009) and Montemagni (2008). Among them, Montemagni (2008) reports that phonetic and morpho-lexical variation in Tuscany does not appear to conform to the same pattern. This result, which is in contrast with the outcome of the other studies, is complemented by a significantly different degree of correlation between geography on the one hand and observed patterns of phonetic vs. morpho-lexical variation on the other hand ( $r = 0.1358$  vs.  $r = 0.6441$ ).

But what happens if within the same level of description different linguistically-grounded partitions of the data are considered? For instance, if the focus is on lexical variation, will the resulting partitions be the same across different semantic fields? Within a dialectometric study of Tuscan dialectal variation, Montemagni (2010) reports that patterns of variation identified with respect to different semantic domains (e.g., agriculture, weather, house, stockbreeding) differ significantly from the overall picture, suggesting the influence of extra-linguistic factors other than geography (e.g., climate, geomorphology, history). However, this is an impressionistic analysis which needs to be investigated further. Stronger evidence in this direction emerges from Franco, Geeraerts & Speelman (2015) who, building on previous studies (Speelman & Geeraerts 2008; Geeraerts & Speelman 2010), demonstrated that semantic field influences so-called ‘onomasiological heterogeneity’, i.e. the fact of being significantly more prone to variation in concept naming.

Within the wider context of this study, in this paper we also intend to contribute to the issue of whether and how the influence of extra-linguistic factors varies across linguistically-grounded partitions of data. In particular, whether there are extra-linguistic factors playing a stronger role in the definition of specific clusters. The aim of this paper can thus be summarized by the following two research questions:

- are patterns of linguistic variation influenced by extra-linguistic features such as geomorphology of the area, or cultural, administrative or political boundaries?
- To what extent does the role and impact of these features remain stable across linguistically-grounded partitions of data?

To answer these questions a case study focusing on lexical variation has been carried out on a regional lexical atlas of Tuscan dialects. A variety of extra-linguistic features was taken into account, whose impact and role has been evaluated with respect to both the whole dialectal dataset and across semantic fields.

## 2 Data

### 2.1 Dialectal data

The dialectal corpus of the *Atlante Lessicale Toscano* ('Lexical Atlas of Tuscany', ALT; Giacomelli et al. 2000) was used.<sup>1</sup> ALT is an Italian regional lexical atlas focusing on dialectal variation throughout Tuscany, where both Tuscan and non-Tuscan dialects are spoken. In this paper, we focused on Tuscan dialects only, recorded in 213 localities by a total of 2060 informants, socio-demographically selected with respect to parameters such as age, education and gender.

ALT interviews were carried out on the basis of a questionnaire including onomasiological questions, i.e. looking for concept lexicalizations, and organized into semantic domains (e.g., agriculture, food, wild animals, weather, house etc.). Out of the 460 onomasiological questions, we selected only those focusing on nominal concepts and characterized by lower 'onomasiological heterogeneity' (in the case at hand, showing 50 or fewer answer types). The resulting subset consists of 170 questionnaire items for which a total of 5,174 normalized answers types were given, corresponding to 61,496 geo-localized responses and 384,454 individual ones. Based on the results by Wieling & Montemagni (2016) who demonstrated that cluster quality improves when the analysis is based on all data, we used unfiltered data, with no pruning of infrequent variants.

To abstract away from productive phonetic variation, we used the normalized representation of ALT dialectal items (Cucurullo et al. 2006). The representativeness of the selected sample with respect to the whole set of ALT onomasiological questions was assayed using the correlation between overall lexical distances and lexical distances obtained from the selected sample, which turned out to be high ( $r = 0.94$ ). Note that the same set of questions was used in different studies, by Montemagni & Wieling (2016) on Tuscan lexical variation, and by Wieling et al. (2014) on the relationship between Tuscan dialects and standard Italian.

### 2.2 Extra-linguistic data

For this study, we focused on the following typology of extra-linguistic features:

- geomorphology of the area, described in terms of hydrographic basins;
- religious subdivisions, corresponding to dioceses in turn aggregated into arch-dioceses: these represent territorial units of administration of the Catholic Church whose origin dates back to centuries ago, when a formal church hierarchy was set up, parallel to the civil administration (whose areas of responsibility often coincided);
- political and administrative subdivisions, such as state or province.

---

<sup>1</sup> ALT is available as an online resource at the following address:  
<http://serverdbt.ilc.cnr.it/ALTWEB>.

Figure 1 shows the clusterings of Tuscan ALT locations according to the selected extra-linguistic features. Some of these were obtained from the online version of the *Geographical, physical and historical dictionary of Tuscany* by Emanuele Repetti (1833-1843), which is an encyclopedic collection of information about Tuscany published in the 19th century concerning notable places, from large towns to small villages, providing historical, archaeological and artistic information as well as physical land attributes (e.g., mountains, rivers, lakes, etc.).<sup>2</sup> Information about state, archdiocese and basin authority (the administrative counterpart of hydrographic basin) was extracted from Repetti's Dictionary. State refers to the political organization of Tuscany in 1833; archdiocese and basin authority refer to current geographical-administrative subdivisions, which were reconstructed from the diocese and valley information respectively, reported in the dictionary. The province subdivision refers to the current administrative organization of Tuscany.

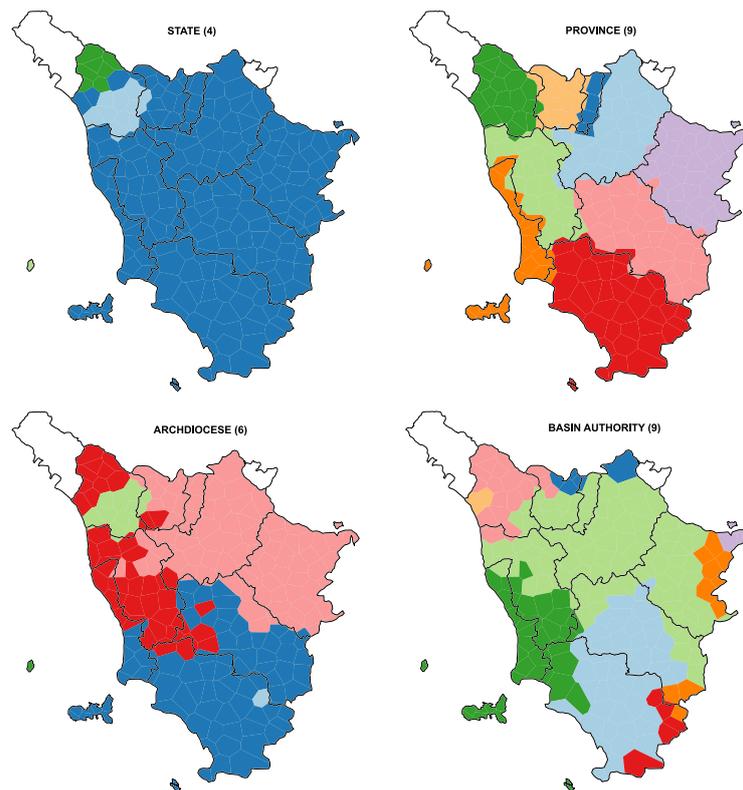


Figure 1: Geographical clustering of Tuscan ALT locations on the basis of extra-linguistic criteria (state, province, archdiocese, basin authority).

<sup>2</sup> The dictionary, published online in 2005, can be accessed at the following address: <http://www.archeogr.unisi.it/repetti>.

## 3 Methods

### 3.1 Clustering of dialectal data

For the clustering of the dialect data, we use bipartite spectral graph partitioning (Dhillon 2001). This algorithm has been used for the clustering of Tuscan dialect data before (Montemagni & Wieling 2016) and was introduced to dialectometry by Wieling & Nerbonne (2011). Bipartite spectral graph partitioning simultaneously clusters geographic locations together with their associated (characteristic) linguistic features. In short, the method functions by computing the singular value decomposition of the input matrix (of geographical locations and linguistic features), and subsequently uses  $k$ -means to (recursively) obtain a partitioning in two groups. Consequently, the clustering consists of locations together with their associated linguistic features. While Montemagni & Wieling (2016) focused on investigating and discussing the underlying features of the different clusters, here we only use the resulting geographical clustering. As we use non-linguistic data organized into either 4, 6 or 9 groups, we use linguistic clusterings having a similar number of clusters.<sup>3</sup> Besides including a clustering in 6 and 8 groups on the basis of all data, we include clusterings generated on the basis of including separate semantic fields. These consist of agriculture (6 and 9 clusters), animals (3 and 4 clusters), food (6 and 8 clusters), and house (6 and 8 clusters).

### 3.2 Clustering comparison

As indicated above, we use the VARIATION OF INFORMATION criterion created by Meilă (2003). This measure is an information theoretic criterion related to mutual information to compare two clusterings of the same data set. The advantage of the approach is that it “makes no assumptions about how the clusterings were generated” (Meilă 2003: 173) and that the measure is a “true metric on the space of clusterings” (Meilă 2003: 173). It is defined as follows:

$$V(X; Y) = -\sum_{i,j} r_{ij} (\log(r_{ij}/p_i) + \log(r_{ij}/q_j))$$

with  $p_i = |X_i|/n$  and  $q_j = |Y_j|/n$  ( $n$  equals the complete number of data points in the clustering). In the following, we will use this measure as implemented in the R package `mcclust` (Fraley & Raftery 2002).

Since there are several clusterings we need to compare, we will summarize the results visually by using multidimensional scaling (MDS) in two dimensions. In this way we are able to visualize the 10 linguistic clusters (2 on the basis of all data, and 8 on the basis of separate semantic fields) together with 4 extra-linguistic clusterings.

---

<sup>3</sup> Importantly, the VI measure we use to obtain a quantification of the difference of two clusters, does not require the clusterings to have the same number of clusters. This is fortunate as the bipartite spectral graph partitioning method does not result in a pre-specified number of clusters, as not always a division in two groups may be possible.

## 4 Results

Figure 2 visualizes the linguistic clustering for the complete data set, as well as separated by semantic domain. By visually comparing the clusterings, it is clear there are similarities. For example, the south area corresponding to the Grosseto and Siena provinces is always clustered together, but there are also clear differences at the level of the north area, going from Lucca to Arezzo (the names of the aforementioned provinces are marked in the top-left graph in Figure 2). The question which is being investigated is whether and how these clusters correlate with those reported in Figure 1, based on extra-linguistic criteria.

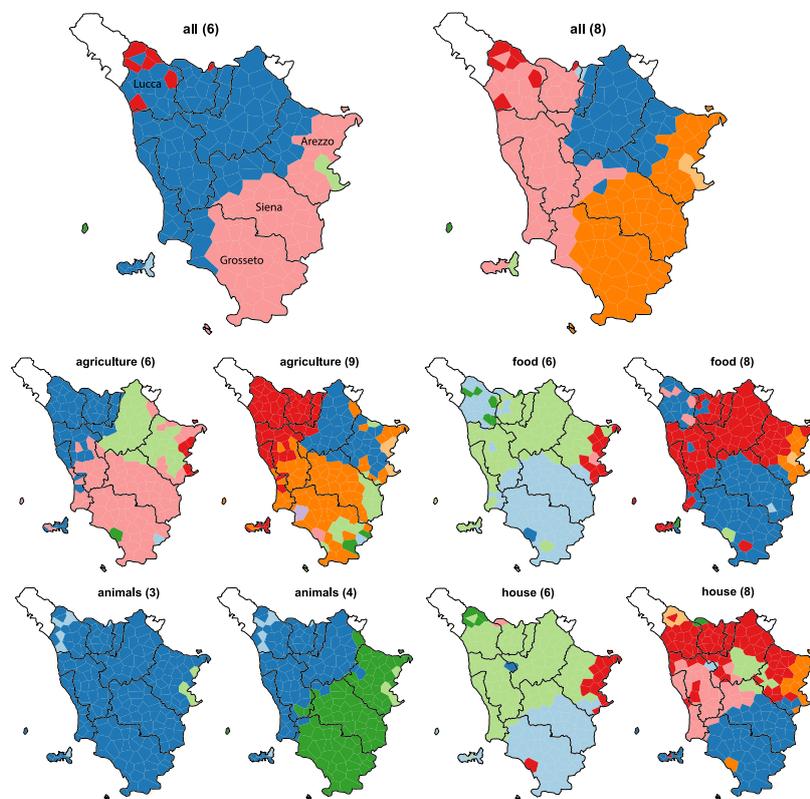


Figure 2: Geographical clustering of linguistic data.

Figure 3 represents the MDS visualization of the similarity of linguistic and extra-linguistic clusters on the basis of the VARIATION OF INFORMATION criterion: note that capitalized cluster names refer to extra-linguistic partitions of locations and that the suffixed number indicates the number of clusters. The visualization in two dimensions was adequate, as stress was only 0.16.

Consider first the relationship extra-linguistic vs. linguistic clusterings. With re-

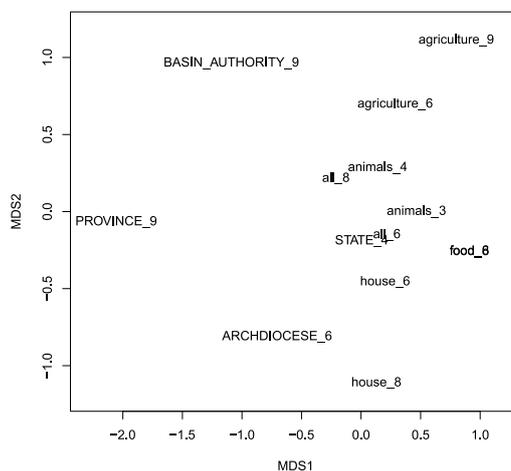


Figure 3: MDS visualization of similarity of clusterings.

spect to political and administrative subdivisions, it can be noticed that the STATE clustering seems to be most similar to the dialect results, in particular to those obtained with respect to the *animals*, *house* and *food* domains as well as to the complete dataset (named *all*). Interestingly, the current administrative subdivision, i.e. PROVINCE, seems to be far away from all linguistic subdivisions. For what concerns ARCHDIOCESE, the stronger similarity concerns the *house* and *food* domains. Last but not least, the clustering based on geomorphology, i.e. BASIN\_AUTHORITY, shows a stronger similarity with respect to the *agriculture* and *animals* domains. Among all, the closest distances are observed with respect to STATE, followed by ARCHDIOCESE, BASIN\_AUTHORITY and lastly by PROVINCE (average VI scores with respect to each of them are 1.76, 2.42, 2.59 and 2.85, respectively). For what concerns the linguistically-based partitions, it is interesting to note the stronger similarity between the clusterings in the *agriculture* and *animals* domains on the one hand, and between *food* and *house* on the other hand, with *food* functioning as a transition between nature-based vs. culture-based partitions.

## 5 Conclusion

Going back to the research questions we started with, the results of this study clearly show that the patterns of lexical variation identified through bipartite spectral graph partitioning are influenced by external factors limiting the communication across the region. These factors range from political (states and provinces) and cultural (archdioceses) subdivisions to physical ones (river basins). This provides strong support

to Bloomfield's theory of linguistic variation which is seen to be driven by so-called "density of communication", predictable from a variety of external factors.

We have seen that STATE and ARCHDIOCESE, reflecting pre-unitarian<sup>4</sup> subdivisions dating back to centuries ago, appear to play a central role in defining patterns of dialectal variation. In this respect, it is interesting to go back to Bloomfield (1933: 343), who claims that the primary correlation is with political subdivisions: "The important lines of dialectal division run close to political lines". Similarly, Kurath (1954) reports the coalescence of Middle English dialect boundaries with county lines. Interestingly, more recent administrative subdivisions (provinces) turned out to play a very small role in defining the Tuscan dialectal landscape. Again, these results are in line with the claim by Bloomfield (1933: 343) that "isoglosses along a political boundary of long standing [...] would persist, with little shifting, for some two-hundred years after the boundary had been abolished".

BASIN\_AUTHORITY, in spite of representing a sort of artificial subdivision collapsing adjacent river basins for administrative reasons, seems to play an important role in shaping patterns of lexical variation, especially for what concerns the semantic fields of *agriculture* and *animals*. This leads to the second research question, investigating whether and to what extent the impact of external features remains stable across linguistically-grounded partitions of data, semantic fields in the case at hand. On the basis of these results, the answer is positive. The influence of the different extra-linguistic factors taken into account in this study turned out to differ across semantic fields. Physical subdivisions are more relevant for what concerns the *agriculture* and *animals* domains, while political and cultural borders appear to play a stronger role with more culturally-oriented semantic domains (i.e. *house* and *food*).

## References

- Bloomfield, Leonard. 1933. *Language*. New York: Holt.
- Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge, New York: Cambridge University Press.
- Cucurullo, Nella, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi & Eva Sas-solini. 2006. Dialectal resources on-line: the ALT-Web experience. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 1846–1851. European Language Resources Association (ELRA).
- Dhillon, Inderjit. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–274. ACM New York, NY, USA.
- Fraley, Chris & Adrian E. Raftery. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97. 611–631.

---

<sup>4</sup> Italian political unification was in 1861.

- Franco, Karlien, Dirk Geeraerts & Dirk Speelman. 2015. Why dialects differ: the influence of concept features on lexical geographical variation. In *Proceedings of the International Conference on Language Variation in Europe (ICLaVE 8)*, Universität Leipzig, Leipzig, Germany, May 27-29, 2015.
- Geeraerts, Dirk & Dirk Speelman. 2010. Heterodox concept features and onomasiological heterogeneity in dialects. In D. Geeraerts, G. Kristiansen & Y. Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 23–40. Berlin, New York: De Gruyter Mouton.
- Giacomelli, Gabriella, Luciano Agostiniani, Patrizia Bellucci, Luciano Giannelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Eugenio Picchi & Teresa Poggi Salani. 2000. *Atlante Lessicale Toscano*. Roma: Lexis Progetti Editoriali.
- Goebl, Hans. 2005. La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de Linguistique Romane* 69. 321–367.
- Gooskens, Charlotte. 2005. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13. 38–62.
- Heeringa, Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13(3). 375–400.
- Hubert, Lawrence & Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2(1). 193–218.
- Kurath, Hans. 1954. *Middle English Dictionary, Plan and Bibliography*. Ann Arbor: University of Michigan Press.
- Manni, Franz, Wilbert Heeringa, Bruno Toupance & John Nerbonne. 2008. Do surname differences mirror dialect variation? *Human Biology* 80(1). 41–64.
- Meilã, Marina. 2003. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, 173–187. Springer.
- Montemagni, Simonetta. 2008. The space of tuscan dialectal variation: a correlation study. *International Journal of Humanities and Arts Computing* 2(1–2). 135–152.
- Montemagni, Simonetta. 2010. Esplorazioni computazionali nello spazio della variazione lessicale in Toscana. In *Atti del convegno 'parole. Il lessico come strumento per organizzare e trasmettere gli etnosaperi', 2–4 luglio 2009, Rende*, 619–644. Centro Editoriale e Libreria dell'Università della Calabria.
- Montemagni, Simonetta & Martijn Wieling. 2016. Tracking linguistic features underlying lexical variation patterns: a case study on tuscan dialects. In *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, 117–134. Language Science Press.
- Nerbonne, John. 2013. How much does geography influence language variation? In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: Geographical, Interactional, and Cognitive Perspectives*, 220–236. Berlin, New York: Walter de Gruyter.
- Prokić, Jelena, Martijn Wieling & John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, 18–25. Association for Computational Linguistics.

- Speelman, Dirk & Dirk Geeraerts. 2008. The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing* 2(1–2). 221–242.
- Spruit, Marco René, Wilbert Heeringa & John Nerbonne. 2009. Associations among Linguistic Levels. *Lingua* 119(11). 1624–1642.
- Wieling, Martijn & Simonetta Montemagni. 2016. Infrequent forms: noise or not? In *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, 215–224. Language Science Press.
- Wieling, Martijn, Simonetta Montemagni, John Nerbonne & R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90(3). 669–692.
- Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25(3). 700–715.



## Chapter 26

# Default inheritance and derivational morphology

Stefan Müller

Humboldt University Berlin

This paper is a contribution to the discussion whether argument structure constructions should be treated phrasally or lexically. While lexical models can explain the interaction between argument structure constructions and derivational morphology in a straightforward way, the analysis of this interaction is a desideratum for phrasal analyses. This paper deals with the question whether type hierarchies together with default inheritance can be used to describe derivational morphology. Given the challenges provided by Krieger & Nerbonne (1993) it seems impossible to do derivation without embedding (that is something like morphological phrase structure rules with mother/daughter relations or lexical rules/constructions with an input/daughter and an output/mother) and as will become clear the price for doing derivation with default inheritance is very high indeed.

## 1 Introduction

Goldberg (1995) and Goldberg & Jackendoff (2004) argue that Resultative Constructions like those in (1) are best described by phrasal rules that contribute the part of meaning that is specific to such resultative constructions.

- (1) a. The pond froze solid.

---

<sup>1</sup> I thank Ann Copestake for discussion and John Bateman, Dorothee Beermann, Hans-Ulrich Krieger, and Andrew McIntyre for discussion and for comments on earlier versions of Müller (2006), research that is related to the present paper. An earlier version of this paper was presented at the 2nd International Workshop on Constraint-Based Grammar, which was held in 2005 in Bremen. I thank all participants for discussion.

While preparing this paper I extended the explanations and added and updated references. I thank Antonio Machicao y Priemer for detailed comments which resulted in many clarifications in the revised paper. The revised paper is 21 pages long, which is way outside of the page limit of the present collection. So I decided to publish the earlier version with all its shortcomings. The interested reader is referred to the extended version, which can be downloaded at <http://hpsg.fu-berlin.de/~stefan/Pub/default-morph.html>.

- b. The gardener watered the flowers flat.
- c. They drank the pub dry.

Authors who work in the framework of Construction Grammar (CxG) usually capture generalizations about Constructions in inheritance hierarchies (Kay & Fillmore 1999; Goldberg 2003). Goldberg assumes that generalizations regarding active, passive, and middle variants of the Resultative Construction can be expressed this way. However, there are other ways to realize Resultative Constructions and it is not obvious how such patterns should be treated. The German examples in (2) show, that resultative constructions and derivational morphology interact (Müller 2002; 2003; 2006):

- (2) a. *-ung* nominalizations:  
*Leerfischung*<sup>2</sup> ‘empty.fishing’, *Kaputterschließung*<sup>3</sup> ‘broken.development’,  
*Kaputtmilitarisierung*<sup>4</sup> ‘broken.militarization’, *Gelbfärbung*<sup>5</sup>  
‘yellow.dyeing’
- b. *-er* nominalizations:  
*Totschläger*<sup>6</sup> ‘dead.beater’ or ‘cudgel’, *SFB-Gesundbeter*<sup>7</sup>  
‘SFB.healthy.prayer’,  
*Ex-Bierflaschenleertrinker*<sup>8</sup> ‘ex.beer.bottles.empty.drinker’
- c. marginally in *Ge-* *-e* nominalizations:  
*Totgeschlage*<sup>9</sup> ‘beating.to.death’

So if all generalizations about resultative constructions are captured in inheritance hierarchies, the derivational facts should be covered that way too.

Krieger & Nerbonne (1993) showed that derivational morphology cannot be modeled using (simple) inheritance hierarchies since recursion as for example in *Vorvorversion* ‘preprepreversion’ cannot be covered in inheritance networks (Krieger & Nerbonne 1993). Since information about the prefix *vor-* is contained in *Vorversion* inheriting a second time from *vor-* would not add anything. Secondly, in an inheritance-based approach to derivation, it cannot be explained why *undoable* has the two readings that correspond to the two bracketings in (3), since inheriting information in different orders does not change the result.

- (3) a. [un- [do -able]]
- b. [[un- do] -able]

Proponents of CxG often refer to default inheritance (Goldberg 1995; Michaelis & Ruppenhofer 2001) and Michaelis & Ruppenhofer (2001) explicitly suggest an analysis of derivational morphology that is based on default inheritance. For the class

<sup>2</sup> taz, 20-06-1996, p. 6.

<sup>3</sup> taz, 02-09-1987, p. 8.

<sup>4</sup> taz, 19-04-1990, p. 5.

<sup>5</sup> taz, 14-08-1995, p. 3.

<sup>6</sup> taz, Bremen, 24-05-1996, p. 24 and taz, Hamburg 21-07-1999, p. 22

<sup>7</sup> taz, 25-08-1989, p. 20.

<sup>8</sup> taz, 13/14-01-2001, p. 32.

<sup>9</sup> Fleischer & Barz (1995: 208).

of *be*-Verbs they assume the Applicative Construction which derives for instance *be-laden* from *laden* by overriding incompatible information of the base verb by material inherited from the *be*- Construction (p. 59).

This paper deals with the question whether type hierarchies together with default inheritance can be used to describe derivational morphology. Given the challenges provided by Krieger & Nerbonne (1993) it seems impossible to do derivation without embedding and as will become clear the price for doing derivation with default inheritance is very high indeed.

A full account of derivational morphology has to explain the following facts:

- derivation may change the phonological form (*Les+bar+keit* ‘readability’);
- derivation may change the syntactic category:  
*Les+bar+keit* ‘readability’ =  $V \rightarrow A \rightarrow N$ ;
- derivation changes the semantic contribution:  
 $\text{lesen}(x,y) \rightarrow \text{modal}(\text{lesen}(x,y)) \rightarrow \text{nominal}(\text{modal}(\text{lesen}(x,y)))$ .

The crucial point that has to be captured by every analysis is that there are productive morphological patterns. This means that it is not sufficient to specify two types in a type hierarchy and introduce explicitly a new subtype of the latter two types. One could do this for instance for *Lesbarkeit*. The category of the verbal stem and of all affixes would be specified as a default and a subtype for *lesbar* and *Lesbarkeit* is stipulated. In such a setting, the values that override defaults are stipulated for all lexemes, they do not follow from any rule. This is not adequate since it does not capture the productive aspect of many morphology patterns. What is needed is some way to automatically compute subtypes that correspond to stems or words. This can be done by an automatic closure computation that introduces new types for all compatible types specified in a type hierarchy. Such online type computation was suggested by Kay (2002) for phrasal Constructions in CxG and by Koenig (1999) in the framework of HPSG. However, such online computation makes it necessary to specify the default information in appropriate ways and to have some way to determine automatically in which ways default information may be overridden. In the remainder of the paper, I show how a default-based analysis has to be set up in order to capture the productive aspects of derivational morphology. I use the default logic described in Lascarides & Copestake (1999). In this formalization default information is explicitly marked, so the hierarchy may be set up in a way that it is clear which information overrides which other information.

## 2 Derivational morphology with default inheritance

The following subsection deals with changes in phonology, Subsection 2.2 deals with changes in syntactic category, and Subsection 2.3 captures semantics.

## 2.1 Changes in phonology

Villavicencio (2000: 86–87) suggested a way to extend the length of a list by using default unification. The trick is to mark the end of a list as default information. This information maybe overridden by inheritance from another type that specifies a conflicting value for the list end. For instance, the stem *les* can be specified as follows:

$$(4) \left[ \begin{array}{l} les \\ \text{PHON-H} \left[ \begin{array}{l} ne\text{-list} \\ \text{HD } les \\ \text{TL } / e\text{-list} \end{array} \right] \end{array} \right]$$

*e-list* stands for empty list and *ne-list* for non-empty list.

Combining this stem with the suffix *-bar* in (5a) yields the representation in (5b):

$$(5) \quad \text{a.} \left[ \begin{array}{l} bar \\ \text{PHON-H} \left[ \begin{array}{l} ne\text{-list} \\ \text{TL} \left[ \begin{array}{l} ne\text{-list} \\ \text{HD } bar \\ \text{TL } / e\text{-list} \end{array} \right] \end{array} \right] \end{array} \right] \quad \text{b.} \left[ \begin{array}{l} les \wedge bar \\ \text{PHON-H} \left[ \begin{array}{l} ne\text{-list} \\ \text{HD } les \\ \text{TL} \left[ \begin{array}{l} ne\text{-list} \\ \text{HD } bar \\ \text{TL } / e\text{-list} \end{array} \right] \end{array} \right] \end{array} \right]$$

The PHON-H|TL value of (4) is overridden by the value in (5a). The new end of the list is in turn marked as default.

There is one little problem and one big problem with this approach. The little problem is that the elements in the PHON-H list are in the wrong order if the affix is a prefix. Consider the noun *Vorversion*. Since *Vor-* is a prefix it should appear before *Version*, but if we use the mechanism described above, affixes are appended at the end of the PHON-H list. This problem could be solved by making the values in the PHON-H list more complex: if new material is added at the end of the list, information about prefix/suffixhood is added as well. For a noun like *Anfahrbarkeit* as in *Die Anfahrbarkeit des Flughafens muß gewährleistet sein* ‘The accessibility of the airport by car must be guaranteed’, one would get  $\langle \textit{fahr}, \textit{an}\text{-prefix}, \textit{bar}\text{-suffix}, \textit{keit}\text{-suffix} \rangle$ . We would then use the following relational constraint that maps this list onto the actual phonological realization.

$$(6) \quad \text{compute\_phon}(\langle \rangle, \underline{1}, \underline{1}).$$

$$\text{compute\_phon}(\langle \left[ \begin{array}{l} prefix \\ \text{PHON } \underline{1} \end{array} \right] \rangle \oplus \underline{2}, \underline{3}, \underline{4}) \text{ if}$$

$$\text{compute\_phon}(\underline{2}, \langle \underline{1} \rangle \oplus \underline{3}, \underline{4}).$$

$$\text{compute\_phon}(\langle \left[ \begin{array}{l} suffix \\ \text{PHON } \underline{1} \end{array} \right] \rangle \oplus \underline{2}, \underline{3}, \underline{4}) \text{ if}$$

$$\text{compute\_phon}(\underline{2}, \underline{3} \oplus \langle \underline{1} \rangle, \underline{4}).$$

The symbol ‘ $\oplus$ ’ stands for *append*, a relation that concatenates two lists. `compute_phon` is defined recursively. The second and the third clause take one element – a prefix or a suffix, respectively – from the beginning of a list and then call `compute_phon` with a shorter list (2). The first clause ends the recursion. If the first argument contains an empty list, all affixes are processed and the second and the third argument are identified (1). When `compute_phon` is called initially the second argument is the root, e.g., *fahr* in *Anfahrbarkeit*. The list of affixes is passed to `compute_phon` as the first argument. If this list starts with a prefix, the PHON value of the prefix (1) is appended to the value in the second slot (3) and the result  $\langle 1 \rangle \oplus 3$  is the second argument of the recursive call of `compute_phon`. The third clause is for suffixes and works parallel to the second clause, the only difference being that the phonology of the suffix is appended to the second argument at the end. By recursively working through the affix list in the first slot the list gets shorter while the list in the second slot gets longer. When the recursion ends due to exhaustion of the list of affixes, the phonological information is in the second slot. Clause one of `compute_phon` identifies the second and the third slot of the relational constraint. Since the third slot is just passed on in the second and third clause (4), the third slot will contain the result of the PHON computation.

The value that is determined by the relational constraint is declared to be the PHON value of the sign (1):

$$(7) \left[ \begin{array}{l} \text{PHON} \quad [1] \\ \text{PHON-H} \quad \left[ \begin{array}{l} \text{HD} \quad [\text{PHON} \quad 2] \\ \text{TL} \quad [3] \end{array} \right] \end{array} \right] \wedge \text{compute-phon}(3, 2, 1)$$

`compute_phon` takes the phonology of the root (2) as its second argument and the remainder of the PHON-H list, which contains all the affixes, as its first argument (3).

This analysis also gets the bracketing problem in (3) right: for [un [do able]], we get  $\langle do, able\text{-suffix}, un\text{-prefix} \rangle$  and for [[un do] able], we get  $\langle do, un\text{-prefix}, able\text{-suffix} \rangle$ . The phonology is the same in both cases, but the semantics differs. See Section 2.3 for the meaning representation.

While this solves the problem of ordering prefixes and suffixes, there remains an even bigger problem. This problem has to do with the question when the PHON value is determined. The computation of the PHON value has to happen at the interface between the lexicon and syntax, that is, at the moment when it is clear that no further affixes will be unified with the existing description. If one would have a PHON value for the stem *les*, this PHON value would also be part of the unification of *les* and *-bar*. If one computes the PHON value for *lesbar* one would get a conflict between the value of *les* and the computed value *lesbar*. It would not be an option to leave PHON values of stems underspecified and let the affix determine the PHON value of the whole construction since it is possible to have more than one affix.

## 2.2 Changes in part of speech

The change in part of speech can be explained in an analogous fashion: the category information is stored in an auxiliary list (CAT-H) that is extended by the affix. A path equation is used to identify the last element of the auxiliary list with the actual category value (CAT). This path equation is specified to be default information. An affix can override this information with an explicit path inequation and add a new default path equation that points to a newly introduced element at the end of the auxiliary list.

## 2.3 Changes in semantics

For the meaning representation similar tricks can be applied, a difference being that we need embedding. Let us consider the noun *Vorversion*. The lexical item for *Version* is given in (8):

$$(8) \left[ \begin{array}{l} \text{lex-version} \\ \text{PHON-H} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD } \textit{version} \\ \text{TL } / \textit{e-list} \end{array} \right] \\ \text{SEM} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD } \textit{version-rel} \\ \text{TL } / \textit{e-list} \end{array} \right] \end{array} \right]$$

This description says that the value of PHON-H is  $\langle \textit{Version} \rangle$  and that the value of SEM is  $\langle \textit{version-rel} \rangle$ . The important part of the definition above is that the TL value is marked to be a default specification, i.e., this value may be overridden. Using lists for the representation of semantic information is also crucial for the mechanisms described below.

The following type for the prefix *Vor-* can be unified with the type *lex-version*. *pref-vor* contains information about the second element of the PHON-H list and the SEM list. The result of the unification of *lex-version* and *pref-vor* is given in (9b):

$$(9) \text{ a. } \left[ \begin{array}{l} \text{pref-vor} \\ \text{PHON-H} \left[ \begin{array}{l} \text{TL} \left[ \begin{array}{l} \text{HD } \textit{vor} \\ \text{TL } / \textit{e-list} \end{array} \right] \end{array} \right] \\ \text{SEM} \left[ \begin{array}{l} \text{HD } \boxed{1} \\ \text{TL} \left[ \begin{array}{l} \text{HD} \left[ \begin{array}{l} \textit{vor-rel} \\ \text{ARG1 } \boxed{1} \end{array} \right] \\ \text{TL } / \textit{e-list} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right] \text{ b. } \left[ \begin{array}{l} \text{lex-version} \wedge \text{lex-vor} \\ \text{PHON-H} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD } \textit{version} \\ \text{TL} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD } \textit{vor} \\ \text{TL } / \textit{e-list} \end{array} \right] \end{array} \right] \\ \text{SEM} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD } \boxed{1} \textit{version-rel} \\ \text{TL} \left[ \begin{array}{l} \text{ne-list} \\ \text{HD} \left[ \begin{array}{l} \textit{vor-rel} \\ \text{ARG1 } \boxed{1} \end{array} \right] \\ \text{TL } / \textit{e-list} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$$

So the prefix *Vor-* extends the PHON-H list and adds its phonological information. It also adds its semantic contribution and embeds the semantic contribution of the

type it was combined with (1). Of course the semantic representation in (9) is not satisfying, since the SEM list contains both a version-rel(X) and an embedded version-rel(X), but only the representation vor-rel(version-rel(X)) is appropriate for *Vorversion*; version-rel(X) is not. In the framework of *Minimal Recursion Semantics* (2005) pointers are used to identify the main semantic contribution of a linguistic object. This pointer would identify version-rel in the entry for *Version* and vor-rel in the lexical entry of *Vorversion*. The path equation pointing to the semantic contribution has to be specified as default information. An affix overrides this equation with a non-default inequation and adds a new default equation pointing to the meaning representation that it contributes.

### 3 Comments on defaults and recursion

Note that this analysis is basically a misuse of defaults. In classical knowledge representation defaults are used to say things like the following: birds have wings and they can fly. A penguin is a bird, it inherits the property of having wings from the super concept, but it overrides the property of being able to fly. In the analysis of derivation given above, a crucial property of a word, namely how it is pronounced, is overridden. This overriding can occur an unbounded number of times. This amounts to making a statement like the following: *Vorvorvorversion* is essentially *Version* except that it is pronounced differently and means something different.

Whether or not one sees ways around this problem, there is a more serious problem, namely that the types above do not account for *Vorvorversion*. To account for *Vorvorversion* we have to have a type that can be combined with structures that have two elements on their PHON list:

$$(10) \left[ \begin{array}{l} \text{pref-vor-2} \\ \text{PHON-H} \left[ \begin{array}{l} \text{TL} | \text{TL} \left[ \begin{array}{l} \text{HD } \text{vor} \\ \text{TL } / \text{e-list} \end{array} \right] \end{array} \right] \\ \text{SEM} \left[ \begin{array}{l} \text{TL} \left[ \begin{array}{l} \text{HD } \boxed{1} \\ \text{TL} \left[ \begin{array}{l} \text{HD } \left[ \begin{array}{l} \text{vor-rel} \\ \text{ARG1 } \boxed{1} \end{array} \right] \\ \text{TL } / \text{e-list} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$$

This means that we have to have infinitely many entries for each prefix, since we do not know the length of the lists of the stem the prefix combines with. For instance if *Vor-* combines with *An+kündig+ung* ‘announcement’, we would need a version of *Vor-* that attaches itself to the end of a three element list. To form *Vorvorankündigung* we need another *Vor-* that attaches to an even longer list.

One can imagine a way to fix this: instead of giving a fixed path to the end of the list in the definition of *pref-vor* or *pref-vor-2*, one could extend the formalism and allow for regular expressions in type declarations. The star after TL in the following

definition would mean that the feature description that follows it can be unified in after any sequence of  $TL$ s. Since the  $HD$  of the description following  $TL^*$  is specified to be non-default information and all other  $PHON-H$  values are also non-default information, the expression following  $TL^*$  can only be unified in at the very end of the list.

$$(11) \left[ \begin{array}{l} \textit{pref-vor} \\ PHON-H \left[ TL^* \left[ \begin{array}{l} HD \textit{ vor} \\ TL / e\textit{-list} \end{array} \right] \right] \end{array} \right]$$

Thus we get *Vorversion* and *Vorankündigung*. In addition to the  $PHON-H$  specification in (11), we would need a similar expression for the computation of the semantics.

But note what such an extension would lead to: if we unify the structure in (9b) with a prefix *Vor-* that contains a regular expression as the one above, we do not get a unique result. One possible result of unification would be the structure in (9b) itself, i.e.,  $TL^*$  is expanded as  $TL$  and another possible result would be the structure that corresponds to *Vorvorversion*. This is the case where  $TL^*$  is expanded as  $TL|TL$ . This means that if we “apply” the prefix *Vor-* to *Vorversion* we get two results, one being spurious.

It could be argued that the spurious unification result mentioned above does not do any harm if we use a closure computation, since elements computed twice are not represented twice in the closure. But note that we have two independent regular expressions: one for extending the  $PHON-H$  list and one for extending the  $SEM$  list. Therefore using such regular expressions would not only result in spurious unification results it would also result in unwanted structures, for instance in a structure with  $PHON$  value *Vorversion* and meaning  $\text{vor-rel}(\text{vor-rel}(\text{version-rel}(X)))$ . To fix this, one would have to introduce another extension of the formalism that allows one to use a certain expansion of a regular expression at various places in a feature description.

## 4 Conclusion

This discussion has shown that default inheritance can be used to model derivation only at a very high cost. To achieve this we need:

- an auxiliary list in which the affixes are collected in the order of application;
- a special marking of the elements in the list that indicates whether the item is a prefix or a suffix;
- a complex relational constraint that walks through the auxiliary list and computes the actual phonological form;
- regular expressions in type definitions that basically break everything we know about unification;

- variables that help us to use the same regular expression in a feature description;
- and some sort of automatic unification of lexical types to get recursion.
- In addition to all this additional machinery, we have misused the concept of defaults.

I consider this too high a price to pay. If one compares this approach to the simplicity of embedding constructions like those usually used in lexical rule-based approaches, it is clear which approach should be preferred. Such constructions can state whether they are prefix or suffix constructions simply by putting constraints on the order in which the phonology of the embedded object and the phonology contributed by the construction are concatenated.

Concluding this paper, we can state that derivation cannot be done without embedding in a reasonable way, the techniques developed here may be used to implement other things in inheritance networks, though.

The analysis is implemented in the LKB system (Copestake 2002) and the code of the implementation is available at <http://hpsg.fu-berlin.de/~stefan/Pub/default-morph.html>.

## References

- Copestake, Ann. 2002. *Implementing typed feature structure grammars* (CSLI Lecture Notes 110). Stanford, CA: CSLI Publications.
- Copestake, Ann, Daniel P. Flickinger, Carl J. Pollard & Ivan A. Sag. 2005. Minimal recursion semantics: an introduction. *Research on Language and Computation* 4(3). 281–332.
- Fleischer, Wolfgang & Irmhild Barz. 1995. *Wortbildung der deutschen Gegenwartssprache*. 2nd edn. Tübingen: Max Niemeyer Verlag.
- Goldberg, Adele E. 1995. *Constructions: a Construction Grammar approach to argument structure* (Cognitive Theory of Language and Culture). Chicago/London: The University of Chicago Press.
- Goldberg, Adele E. 2003. Words by default: the Persian complex predicate construction. In Elaine J. Francis & Laura A. Michaelis (eds.), *Mismatch: form-function incongruity and the architecture of grammar* (CSLI Lecture Notes 163), 117–146. Stanford, CA: CSLI Publications.
- Goldberg, Adele E. & Ray S. Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80(3). 532–568.
- Kay, Paul. 2002. An informal sketch of a formal architecture for Construction Grammar. *Grammars* 5(1). 1–19.
- Kay, Paul & Charles J. Fillmore. 1999. Grammatical Constructions and linguistic generalizations: the What's X Doing Y? Construction. *Language* 75(1). 1–33.
- Koenig, Jean-Pierre. 1999. *Lexical relations* (Stanford Monographs in Linguistics). Stanford, CA: CSLI Publications.

*Stefan Müller*

- Krieger, Hans-Ulrich & John Nerbonne. 1993. Feature-based inheritance networks for computational lexicons. In Briscoe, Copestake & de Paiva (eds.), *Inheritance, defaults, and the lexicon*, 90–136. Cambridge University Press.
- Lascarides, Alex & Ann Copestake. 1999. Default representation in constraint-based frameworks. *Computational Linguistics* 25(1). 55–105.
- Michaelis, Laura A. & Josef Ruppenhofer. 2001. *Beyond alternations: a Constructional model of the German applicative pattern* (Stanford Monographs in Linguistics). Stanford, CA: CSLI Publications.
- Müller, Stefan. 2002. *Complex predicates: verbal complexes, resultative constructions, and particle verbs in German* (Studies in Constraint-Based Lexicalism 13). Stanford, CA: CSLI Publications.
- Müller, Stefan. 2003. Solving the bracketing paradox: an analysis of the morphology of German particle verbs. *Journal of Linguistics* 39(2). 275–325.
- Müller, Stefan. 2006. Phrasal or lexical constructions? *Language* 82(4). 850–883.
- Villavicencio, Aline. 2000. The use of default unification in a system of lexical types. In Erhard W. Hinrichs, Walt Detmar Meurers & Shuly Wintner (eds.), *Proceedings of the ESSLLI-2000 Workshop on Linguistic Theory and Grammar Implementation*, 81–96. Birmingham, UK.

## Chapter 27

# Licensing resultative phrases: the case of locatum subject alternation verbs in Japanese

Tsuneko Nakazawa

University of Tokyo

This paper addresses the problem of identifying the predication relation between resultative phrases and arguments they describe. Resultative phrases in Japanese are generally believed to conform to the Direct Object Restriction just like English: that is, they describe the direct object of the transitive verbs. This paper focuses on resultative phrases licensed by the locatum subject alternation verbs in Japanese, and shows that resultative phrases can be predicated of not only the direct object expressing the location argument of the verbs but also the locatum argument. Thus, it is claimed that the predication relation is not constrained on the syntactic ground as generally believed, but rather by the lexical semantics of the verbs.

## 1 Introduction

Some verbs allow alternative syntactic structures to express the same event. The locative alternation verb *load*, for example, gives rise to syntactic structures in which the locative argument appears either as the direct object or as a PP. If the admissibility of resultative construction is sensitive to the syntactic structures, a resultative phrase admitted in one of the alternative structures may be ungrammatical in the other even though they denote the same event. The prediction is borne out as shown in (1). (In the following examples, resultative phrases are underlined while the arguments described by resultative phrases are in bold.)

- (1) (Williams 1980: 204)
- a. John loaded **the wagon** full with hay.
  - b. \*John loaded the hay into **the wagon** full.

While the same constraint is generally believed to hold in the Japanese resultative construction, the present paper focuses on the verbs of locatum subject alternation,

and demonstrates that the syntactic alternation does not affect the admissibility of resultative phrases. Consequently, it is claimed that the admissibility of resultative phrases is not syntactically constrained, but determined by the lexical semantics of the verb.

## 2 The resultative construction in Japanese

It is generally claimed (e.g. Tsujimura 1990; Kageyama 1996) that resultative phrases in Japanese are similar to those in English in that they obey the Direct Object Restriction (Simpson 1983; Levin & Rappaport Hovav 1995), i.e. they are predicated of the direct object of transitive verbs, or of the subject of unaccusative intransitive verbs as exemplified in (2) and (3).

- (2) object-oriented resultative with a transitive verb  
Taro-ga **kabin-o** konagona-ni kowasi-ta.  
Taro-NOM vase-ACC pieces-NI break-PST  
'Taro broke a vase into pieces.'
- (3) subject-oriented resultative with an unaccusative intransitive verb  
**hune-ga** huka-ku sizun-da.  
ship-NOM deep-KU sink-PST  
'(lit.) The ship sank (and ended up) deep.'

The resultative construction in Japanese, however, lacks the third type in Simpson's analysis of English resultative phrases, which is predicated of a post-verbal NP following unergative intransitive verbs, e.g. *I laughed myself sick*, or transitive verbs which do not subcategorize for it, e.g. *I ate him out of house and home* (Simpson 1983: 146-47). This paper is mostly concerned with resultative phrases licensed by transitive verbs subcategorizing for an object NP.

Since resultative phrases describe a state of an argument resulting from the event denoted by the verb, such verbs must generally express an event involving a change of state. Various authors (e.g. Koizumi 1994; Kageyama 1996; 2001) claim that the Japanese resultative construction requires the verbs that specify a change of state of an argument as part of their lexical semantics, rather than just express an event likely to be associated with such a change. For example, Kageyama (1996) and Washio (1997) argue that, unlike the English counterpart, the Japanese verb of applying force *tatak-* 'hit, pound', does not allow a resultative phrase because, although a state change of the theme argument is likely, such a change is not encoded in the lexical semantics of the verb.

The Japanese resultative construction is even more constrained than English in that it only describes a predictable result of the event, which Washio (1997) calls 'weak resultatives'. Thus, as an example of 'strong resultatives' in English, i.e. resultatives which express an unpredictable result, the sentence *The horses dragged the logs smooth* has no well-formed Japanese equivalent because, it is claimed, logs' being

smooth is not a result predictable from horses' dragging them. As discussed above, Japanese lacks resultative phrases predicated of non-subcategorized arguments, and according to Wechsler (1997), they coincide with the resultative phrases not required to express a 'canonical or generic' result in English. Thus, the resultative construction in Japanese requires the verbs with a lexical specification of a state change, and furthermore, of a predictable result of such a change.

At the same time, both in English and Japanese, the resultative construction is not totally productive: not all verbs of state change allow a resultative phrase to co-occur, and not all predictable results can be expressed as a resultative phrase. The collocations of particular verbs and resultative phrases are to some extent conventionalized, or idiomatic.

Morphologically, the head of resultative phrases in Japanese can be an adjective such as *huka-ku* 'deep' in (3), a noun such as *konagona-ni* 'pieces' in (2), or an 'adjectival noun' such as *ike-no-you-ni* 'like a pond' in (6b) below. The syntactic and semantic functions of adjectival nouns are the same as those of adjectives, but their declension is similar to that of nouns rather than adjectives: both nouns and adjectival nouns are suffixed by *-ni* while adjectives are suffixed by *-ku*.

These morphological forms are, however, not unique to the resultative construction, and they also mark, for example, derived adverbials. In fact, some authors uniformly call all resultative phrases 'adverbs of result' (e.g. Nitta 2002). While in English, there are true adverbs of result, e.g. *They decorated the room beautiful\*(ly)* (Geuder 2000), Japanese lacks the morphological distinction between resultative phrases and derived adverbs. Semantically, however, resultative phrases are distinct from adverbs in that they are predicates of individuals, rather than predicates of events as adverbs are generally treated in Davidsonian event semantics (Davidson 1967). This paper does not concern itself with the categorial status of resultative expressions, but defines a resultative phrase as a predicate of individuals which denotes the resultant state of an individual involved in the event denoted by the main verb.

### 3 The locatum subject alternation verbs

The locatum subject alternation (Levin 1993: 81) is invoked by such verbs as *fill*, *decorate*, *cover* and *surround*, and those verbs involve two arguments: one refers to what undergoes motion and the other to the goal of motion. Throughout the present paper, what undergoes motion is called a locatum argument, and the goal is called a location argument, borrowing the terms from E. V. Clark & H. H. Clark (1979), a study of zero-derived denominal verbs. Those terms are used here simply to identify the participants in the event described by the verb, and no theoretical claim is intended that 'location' and 'locatum' are thematic role labels, or associated with specific grammatical functions.

The locatum subject alternation verbs allow two alternative syntactic structures in which the locatum argument appears as a PP headed by *with*, or the subject NP, while the location argument is expressed as the direct object, as shown in (4).

- (4) the locatum subject alternation
- a. He filled a bottle with water. (locatum-PP variant)
  - b. Water filled a bottle. (locatum-subject variant)

The alternation does not change the transitivity of verbs, but the agent, expressed as the subject in (4a), is not expressed in (4b), and consequently (4b) involves one less argument.

What makes the locatum subject alternation possible is two semantic structures associated with the alternation verbs, which are manifested in different syntactic structures. The meaning of the locatum-PP variant in (4a) can be schematized as ‘X causes Y to change state by causing Z to go into/onto Y’, where the variable X stands for the causer of a state change, Y for the location argument which undergoes a state change, and Z for the locatum argument which undergoes motion. The causer X, or the agent, is absent from the locatum-subject variant in (4b), which expresses an inchoative motion: ‘Z goes into/onto Y’.

Some Japanese verbs also invoke the locatum subject alternation similar to that in English. The Japanese verb *mitas-* ‘fill’, for example, appears in alternative syntactic structures in (5).

- (5) the locatum subject alternation in Japanese
- a. *kare-ga bin-o mizu-de mitas-ita.* (locatum-PP variant)  
he-NOM bottle-ACC water-with fill-PST  
‘He filled the bottle with water.’
  - b. *mizu-ga bin-o mitas-ita.* (locatum-subject variant)  
water-NOM bottle-ACC fill-PST  
‘Water filled the bottle.’

The locatum argument *mizu* ‘water’ appears with the suffix *-de* ‘with’ in (5a), and as the subject NP marked by the nominative suffix *-ga* in (5b). Although the expression of the locatum argument in (5a) constitutes an oblique NP marked by *-de*, rather than a PP as is the case with English, the term locatum-PP variant is retained to refer to Japanese as well as English. In both variants, the location argument *bin* ‘bottle’ appears as direct object marked by the accusative suffix *-o*.

## 4 Object-oriented resultative phrases

The DOR requires the resultative phrase to be predicated of the direct object. Consequently, if a resultative phrase appears with locatum subject alternation verbs, it is predicted to describe the location argument. The prediction is borne out as shown in (6), taken from the language corpus BCCWJ. Example (6a) is an instance of the locatum-PP variant, and (6b) is an instance of the locatum-subject variant. The resultative phrase *ippai-ni* ‘full’ describes the direct object *atasi* ‘me,’ and *ike-no-you-ni* ‘like a pond’ describes the resultant state of *kubon-da tokoro* ‘a hollow (in the ground)’ respectively.

- (6) locatum subject alternation verb *mitas-* ‘fill’
- a. (locatum-PP variant)
- kare-wa sutekina utau-youna koe-de [...] atasi-o ippai-ni  
 he-TOP wonderful singing-like voice-with me-ACC full-NI  
 mitasite-kure-ru.  
 fill-give-NONPST  
 ‘He fills me full with his wonderful singing-like voice.’ [Joyce 2003]
- b. (locatum-subject variant)
- ookina sizuku-ga [...] **kubon-da** **tokoro-o** ike-no-you-ni  
 big drop-NOM subside-PST place-ACC pond-GEN-appearance-NI  
 mitas-i, ...  
 fill-and  
 ‘(lit.) Big drops (of water) fill a hollow (in the ground) (so that it becomes) like a pond, and ...’ [Zola 2003]

The verb *oow-* ‘cover’ is another example of locatum subject alternation verbs. Example (7a) is a passivized instance of the locatum-PP variant; the passive subject *yane-no zenbu* ‘the entire roof’ is functionally the direct object of the verb *oow-* ‘cover’, and the resultative phrase *siro-ku* ‘white’ describes it. The agentive argument is suppressed as a result of passivization.

- (7) locatum subject alternation verb *oow-* ‘cover’
- a. (locatum-PP variant)
- huyu-no asa **yane-no zenbu-ga**  
 winter-GEN morning roof-GEN all-NOM  
siro-ku simo-de oow-are-tei-ru ...  
 white-KU frost-with cover-PASS-STATIVE-NONPST  
 ‘On a winter morning, the entire roof is covered white with frost...’ [Okada 1986]

The verb *oow-* ‘cover’ also appears in the locatum-subject variant in (7b), where the locatum argument *mikkabun-no busyouhige* ‘three day’s worth of stubble’ is expressed as subject.

- (7) b. (locatum-subject variant)
- mikkabun-no busyouhige-ga **ganmen-o** kitanarasi-ku  
 three.days-GEN stubble-NOM face-ACC dirty-KU  
 oot-tei-ru.  
 cover-STATIVE-NONPST  
 ‘(lit.) Three days’ growth of stubble covers his face dirty. (He has three days’ growth of stubble on his dirty face.)’ [Forsyth 1989]

The resultative phrase *kitanarasi-ku* ‘dirty’ describes the location argument *ganmen* ‘face’ expressed as direct object. Both examples (7a) and (7b) conform to the DOR

which states that resultative phrases are predicated of either the surface or the deep object (e.g. the passive subject).

Examples in (6) and (7) show that, regardless of which variant the verbs of locatum subject alternation appear in, they express an event in which the referent of direct object NP, i.e. the location argument, undergoes a change of state. The verb *mitas*-‘fill’ lexically specifies that the location argument comes to be filled as a result of the filling event. The verb *oow*-‘cover’ denotes an event in which the location argument comes to be covered. Resultative phrases, if co-occur, further elaborate the resultant state of the location argument which is lexically predetermined, conforming to the DOR.

## 5 Resultative phrases predicated of the locatum argument

As shown in Section 4, the locatum subject alternation verbs which allow a resultative phrase necessarily denote an event in which the location argument undergoes a change of state. Since the state change is a result of motion of the locatum argument, it is also possible to view the locatum argument as undergoing a change of state as well if motion is viewed as a change of location or spatial state. The locatum argument is, however, expressed either as *de*-marked oblique NP or the subject of the alternation verbs, and the DOR predicts that it cannot be modified by a resultative phrase.

Contrary to the DOR, the following examples in (8) show that the verb *oow*-‘cover’ also allows a resultative phrase which is predicated of the locatum argument in either variant. Example (8a) is a passivized example of the locatum-PP variant; the location argument *iwaiwa* ‘rocks’ appears as passive subject while the agentive argument is suppressed. The resultative phrase *atu-ku* ‘thick’ describes the locatum argument expressed as *de*-marked oblique NP, i.e. the (layer of) leaves of blue poppies, rather than the passive subject as predicted by the DOR.

(8) locatum subject alternation verb *oow*-‘cover’

a. (locatum-PP variant)

*iwaiwa-ga sono madara-no ha-de atu-ku*  
rocks-NOM their mottle-GEN leaf-with thick-KU

*oow-are-tei-ru ...*

COVER-PASS-STATIVE-NONPST

‘(lit.) Rocks are covered with mottled leaves (of the blue poppies) thick ...’  
[Kingdon-Ward 1999]

In (8b), a locatum-subject variant, the resultative phrase *usu-ku* ‘thin, sparse’ describes the (density of) clouds, expressed as the locatum subject, rather than the location argument *sora* ‘sky’ expressed as direct object.

- (8) b. (locatum-subject variant)  
 takai **iwasigumo**-ga usu-ku sora-o oot-tei-ru.  
 high mackerel.cloud-NOM thin-KU sky-ACC cover-STATIVE-NONPST  
 ‘(lit.) High mackerel-like clouds cover the sky sparse(ly).’ [Kawabata 2001]

Syntactically, there is no clue as to which argument of the verb is described by the resultative phrases in examples (7) and (8). As discussed in Section 2, however, Japanese allows only ‘weak’ resultatives, i.e. they only describe a predictable result of the argument that undergoes a change of state, and consequently their predication relation can be determined on the semantic ground. If a roof is covered with frost, it is the roof that becomes white in (7a), and if leaves cover rocks, it is the layer of leaves that becomes thick in (8a). No ambiguity arises in any of syntactic variants in (7) and (8).

The verb *tutum-* ‘surround, wrap’, another verb of locatum subject alternation, provides more examples of resultative phrases describing the locatum argument in (9). Example (9a), a locatum-PP variant, is cooking instruction, and the resultative phrase *nizyuu-ni* ‘in two layers’ describes the state of boiled cabbage, the locatum-PP, after wrapping meat in it.

- (9) locatum subject alternation verb *tutum-* ‘surround, wrap’  
 a. (locatum-PP variant)  
 aibikiniku, [...] kosyo-o yoku maze-te, [...]  
 minced.pork.and.beef pepper-ACC well mix-and  
**yude-ta kyabetu**-de nizyuu-ni tutun-de ...  
 boil-PST cabbage-with two.layers-NI wrap-and  
 ‘(lit.) (You) mix minced pork and beef, [...] and pepper well, and wrap the mixture in two layers of boiled cabbage...’ [Yahoo 2008]

In (9b), an instance of the locatum-subject variant appears in a relative clause headed by the verb *tutum-* ‘surround, wrap.’ The head noun of the relative clause *hantoumei pinku* ‘semitransparent pink (substance)’ expresses the locatum argument which is functionally the subject of the verb. The resultative phrase *atu-ku* ‘thick’ describes the locatum argument surrounding the jelly monster.

- (9) b. (locatum-subject variant)  
 kaibutu zerii-no karada-o atu-ku tutum-u  
 monster jelly-GEN body-ACC thick-KU surround-NONPST  
**hantoumei pinku**-wa ...  
 semitransparent pink-TOP  
 ‘(lit.) The semitransparent pink (substance) which surrounds the body of the jelly monster thick...’ [Miura 1995]

In all examples in (8) and (9), resultative phrases are predicated of the locatum argument and describe the result of its change of location and spatial configuration. Syntactically, those locatum arguments are not expressed as direct object, as required by

the DOR, and whether they are expressed as *de*-marked oblique NP or the subject is irrelevant to the predication relation of resultative phrases and what they modify.

## 6 Conclusion

The locatum subject alternation verbs denote an event which involves two entities, one that undergoes a change of state and the other that undergoes a change of location. To express the event, the verbs provide two semantic structures, 'X causes Y to change state by causing Z to go into/onto Y' and 'Z goes into/onto Y', where the variable X stands for the agent, Y for the location argument that undergoes a state change, and Z for the locatum argument that undergoes a location change. Those semantic structures represent alternative views to interpret the same event as a change of state or as a change of location. Those alternative views are evidenced by the occurrence of resultative phrases that describe the result of either change.

Those semantic structures give rise to distinct syntactic structures. The distribution of resultative phrases, however, remains the same regardless which variant of syntactic structures they appear in as shown in Sections 4 and 5. If the predication relation between resultative phrases and what they modify is determined on the syntactic ground such as the DOR, resultative phrases would be predicted to only modify the location argument expressed as the direct object. However, the occurrence of resultative phrases which are predicated of the locatum argument whether expressed as the subject or an oblique NP, indicates that the predication relation is determined on the semantic ground.

## References

- Clark, Eve V. & Herbert H. Clark. 1979. When nouns surface as verbs. *Language* 55(4). 767–811.
- Davidson, Donald. 1967. The logical form of action sentences. In Nicholas Rescher (ed.), *The logic of decision and action*, 81–95. Pittsburgh, PA: University of Pittsburgh Press.
- Geuder, Wilhelm. 2000. Oriented adverbs: Issues in the lexical semantics of event adverbs. Tübingen: Universität Tübingen PhD thesis.
- Kageyama, Taro. 1996. *Doshi imiron: gengo-to ninchi-no setten* [Semantics of verbs: the interface between language and cognition]. Tokyo: Kuroshio.
- Kageyama, Taro. 2001. Kekka kobun [The resultative construction]. In Taro Kageyama (ed.), *Doshi-no imi-to kobun* [Semantics and constructions of verbs], 154–181. Tokyo: Taishukan.
- Koizumi, Masatoshi. 1994. Secondary predicates. *Journal of East Asian Linguistics* 3(1). 25–79.
- Levin, Beth. 1993. *English verb classes and alternation*. Chicago: University of Chicago Press.

- Levin, Beth & Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, MA: MIT Press.
- Nitta, Yoshio. 2002. *Fukushi-teki hyogen-no shoso* [Aspects of adverbial expressions]. Tokyo: Kuroshio.
- Simpson, Jane. 1983. Resultatives. In L. Levin, M. Rappaport & A. Zaenen (eds.), *Papers in Lexical-Functional Grammar*, 143–157. Bloomington: Indiana University Linguistics Club.
- Tsujimura, Natsuko. 1990. Unaccusative nouns and resultatives in Japanese. In Hajime Hoji (ed.), *Japanese/Korean linguistics*, 335–349. Stanford, CA: Center for the Study of Language and Information.
- Washio, Ryuichi. 1997. Resultatives, compositionality and language variation. *Journal of East Asian Linguistics* 6(1). 1–49.
- Wechsler, Stephen. 1997. Resultative predicates and control. In Ralph C. Blight & Michelle J. Moosally (eds.), *Texas Linguistic Forum 38: Proceedings of the 1997 Texas Linguistics Society Conference*, 307–321. Austin: University of Texas Department of Linguistics.
- Williams, Edwin. 1980. Predication. *Linguistic Inquiry* 11(1). 203–238.

## Source of examples

- Forsyth, Frederick. 1989. *The negotiator*. Trans. by Makoto Shinohara. Tokyo: Kadokawa.
- Joyce, James. 2003. *Ulysses*. Trans. by Saiichi Maruya. Tokyo: Shueisha.
- Kawabata, Hiroto. 2001. *Nicotiana*. Tokyo: Kadokawa.
- Kingdon-Ward, Frank. 1999. *Pilgrimage for plants*. Trans. by Hirokazu Tsukaya. Tokyo: Iwanami.
- Miura, Toshihiko. 1995. *Mitsurin resu* [Jungle race]. Tokyo: Kawade Shobo Shinsha.
- Okada, Masakazu. 1986. *Hyomen-no kagaku* [Science of surface]. Tokyo: Otsuki.
- Yahoo! Blog. 2008.
- Zola, Émile. 2003. *Le Ventre de Paris* [The belly of Paris]. Trans. by Koji Asahina. Tokyo: Fujiwara.



## Chapter 28

# Free word order and MCFLs

Mark-Jan Nederhof

University of St Andrews

It was recently shown that the MIX language, over a three-symbol alphabet, is generated by a multiple context-free grammar. This paper investigates generalizations of the MIX language to alphabets of any size, as well as generalizations of the above-mentioned grammar. Presented are theoretical results that shed new light on the relation between these languages and these grammars. Precise conjectures are formulated that would further narrow down this relation. It is explained that validity of these conjectures would greatly enhance our understanding of the abilities of grammatical formalisms to describe free word order.

### 1 Introduction

Since the earliest attempts to build formal grammars describing the syntactic structure of natural languages, a central aim has been to identify formalisms with the appropriate generative power. If a formalism is too weak, it cannot describe all languages, or requires inelegant or an unreasonably large number of grammar rules to describe natural language phenomena. If a formalism is too strong, and allows description of phenomena that are unlike those in any natural language, then this causes its own problems. For example, the formalism may offer too little guidance to a linguist building a grammar by hand, and the search space may be too large for an algorithm to effectively learn a grammar from examples. Moreover, parsing and recognition algorithms of very powerful formalisms tend to have time complexities that are too high to be useful for practical purposes.

Against the backdrop of the Chomsky hierarchy (Chomsky 1959), the notion of *mildly context-sensitive grammars* was an attempt to identify properties required of an appropriate formalism for describing syntax, and to motivate tree adjoining grammars as a prime example of such a grammar formalism (Joshi 1985). The specified properties included formal requirements and informal characterizations.

In the years that followed, other formalisms were shown to be equivalent to tree adjoining grammars (Vijay-Shanker & Weir 1994), adding to the evidence that the tree adjoining languages are a natural class. In addition, more powerful formalisms have

been found that clearly satisfy the formal requirements of mildly context-sensitive grammars and also appeared to satisfy the informal characterizations. The most notable are the multiple context-free grammars (MCFGs) (Seki et al. 1991), generating the multiple context-free languages (MCFLs). For an overview, see Kallmeyer (2010).

Joshi (1985) singles out one particular artificial language as posing a potential challenge to the theory of mildly context-sensitive grammars. This is the language of strings over  $\{b_1, b_2, b_3\}$  such that, for some  $m$ , each of the three symbols occurs exactly  $m$  times, in any order. It is most commonly known as the MIX language (Gazdar 1988), referred to below as  $\text{MIX}_3$  to be able to put it in a broader context later. It is also known as the Bach language, after Bach (1981), although its properties have been studied at least since Aho & Ullman (1972: Exercise 2.6.3c).

The  $\text{MIX}_3$  language represents an extreme case of free word order, which appears to be irrelevant to any natural language. Joshi (1985) conjectured that  $\text{MIX}_3$  was not a tree adjoining language, consistent with the idea that tree adjoining grammars constitute an appropriate restriction of the power of context-sensitive grammars in order to model natural languages. The conjecture was finally proved by Kanazawa & Salvati (2012).

However, this leaves open the question whether  $\text{MIX}_3$  can be generated by other formalisms that are generally considered to be mildly context-sensitive, such as the MCFGs. A recently published result by Salvati (2015) shows the answer to be positive, by a proof via the  $\text{O}_2$  language.  $\text{MIX}_3$  and  $\text{O}_2$  are rationally equivalent, which means that if one is an MCFL then so is the other.

In this paper we will broaden the investigation to the  $\text{O}_n$  languages (Fischer & Rosenberg 1968). For fixed  $n \geq 1$ , there are  $2n$  symbols  $a_1, \dots, a_n, \bar{a}_1, \dots, \bar{a}_n$ . A string is in  $\text{O}_n$  if, for each  $i$ , the number of occurrences of  $a_i$  equals the number of occurrences of  $\bar{a}_i$ . Similarly, a string is in the generalized  $\text{MIX}_n$  language over  $\{b_1, \dots, b_n\}$  if, for some  $m$ , each of the  $n$  symbols occurs exactly  $m$  times. For each  $n$ ,  $\text{O}_n$  and  $\text{MIX}_{n+1}$  are rationally equivalent. Formally,  $T_1(\text{O}_n) = \text{MIX}_{n+1}$  and  $T_2(\text{MIX}_{n+1}) = \text{O}_n$ , where  $T_1$  and  $T_2$  are two rational transductions. (For language  $L$  and rational transduction  $T$  we let  $T(L) = \{w \mid \exists v \in L (v, w) \in T\}$ .) We can specify  $T_1$  and  $T_2$  by finite-state transducers  $M_1$  and  $M_2$ , each with a single state. For  $M_1$ , the transitions are labeled  $a_i : b_i$ , for  $1 \leq i \leq n$ , and  $\bar{a}_1 \cdots \bar{a}_n : b_{n+1}$ . For  $M_2$ , the transitions are labeled  $b_1 \cdots b_i : a_i$  and  $b_{i+1} \cdots b_{n+1} : \bar{a}_i$  for  $1 \leq i \leq n$ .

In this paper, we formulate the family of MCFGs  $G_n$  that are conjectured to generate  $\text{O}_n$ . If our conjectures are true, this would imply the remarkable finding that the MCFLs include all permutation closures of regular languages; by the *permutation closure* of a language  $L$  we mean the set of all strings that are permutations of strings in  $L$ . This implication was mentioned before by Salvati (2015) on the basis of Latteux (1979).<sup>1</sup> Note that this would also mean that the MCFLs, unlike the context-free languages, are closed under the operation of permutation closure.

<sup>1</sup> The relevant result is Proposition III.12 of Latteux (1979), which states that the permutation closure of a regular language with  $n$  symbols is in the closure of  $\text{MIX}_{n+1}$  under homomorphism, inverse homomorphism and intersection with regular language. Note that MCFLs are closed under these three operations. The proof appears to require the following correction in the sixth line: “ $g(h^{-1}(c(w^*)) \cap R)$  où  $R = \{a_1, \dots, a_k\}^* w_1^* \cdots w_p^*$ ”.

The importance of our investigation is that it sheds more light on the apparent incongruence between the generative power of some formalisms commonly considered to be mildly context-sensitive, and the observation that extreme free word order does not seem to occur in natural languages.

## 2 The $O_n$ languages

Let  $n$  be a positive integer. The alphabet  $\Sigma_n$  consists of the  $2n$  pairwise distinct symbols  $a_1, \dots, a_n, \bar{a}_1, \dots, \bar{a}_n$ . The length of a string  $w$  is denoted by  $|w|$ . For a string  $w$  over  $\Sigma_n$  and symbol  $a \in \Sigma_n$ ,  $|w|_a$  denotes the number of occurrences of  $a$  in  $w$ .

The *imbalance* of  $w \in \Sigma_n^*$ , denoted by  $\text{imb}(w)$ , is the  $n$ -tuple  $(|w|_{a_1} - |w|_{\bar{a}_1}, \dots, |w|_{a_n} - |w|_{\bar{a}_n})$ . In other words, for each  $i$  ( $1 \leq i \leq n$ ), we match the number of occurrences of  $a_i$  against the number of occurrences of  $\bar{a}_i$ , and the difference, which can be positive or negative, is one value in the imbalance. The expression  $0^n$  denotes the tuple of  $n$  zeros. If  $w$  is such that  $\text{imb}(w) = 0^n$ , then we say that  $w$  is *balanced*.

For each  $n$ , the language  $O_n$  is defined to be the set of balanced strings, or formally  $O_n = \{w \in \Sigma_n^* \mid \text{imb}(w) = 0^n\}$ .

## 3 MCFGs

We use terminology related to the MCFGs from Seki et al. (1991). This formalism is largely equivalent to the string-based LCFRSs from Vijay-Shanker, Weir & Joshi (1987).

A *multiple context-free grammar* (MCFG) is a tuple  $G = (\Sigma, N, S, R)$ , where  $\Sigma$  is a finite set of *terminals*,  $N$  is a finite set of *nonterminals* ( $\Sigma \cap N = \emptyset$ ), and  $S \in N$  is the *start symbol*. Each nonterminal is associated with a positive integer, called its *fanout*. The start symbol has fanout 1.

Further,  $R$  is a finite set of *rules*, each of the form:

$$A_0(s_1, \dots, s_{k_0}) \rightarrow A_1(x_1, \dots, x_{m_1}) A_2(x_{m_1+1}, \dots, x_{m_2}) \cdots A_r(x_{m_{r-1}+1}, \dots, x_{m_r})$$

where each  $A_i$  ( $0 \leq i \leq r$ ) has fanout  $k_i$ ,  $m_i = \sum_{j:1 \leq j \leq i} k_j$  ( $1 \leq i \leq r$ ),  $x_1, \dots, x_{m_r}$  are pairwise distinct variables, and each  $s_j$  ( $1 \leq j \leq k_0$ ) is a string consisting of variables and terminals. Moreover, each variable  $x_i$  ( $1 \leq i \leq m_r$ ) occurs exactly once in the *left-hand side*  $A_0(s_1, \dots, s_{k_0})$  and the left-hand side contains no other variables. The value  $r$  is called the *rank* of the rule.

An *instance* of a rule is obtained by choosing a string over  $\Sigma^*$  for each variable in the rule, and then replacing both occurrences of each variable by the chosen string. Let  $\hat{\Sigma}_G$  denote the set of symbols of the form  $A(w_1, \dots, w_k)$ , where  $k$  is the fanout of  $A \in N$  and  $w_1, \dots, w_k \in \Sigma^*$ ; we refer to such a symbol as an *instance* of a nonterminal. The binary ‘derives’ relation  $\Rightarrow_G$  over  $\hat{\Sigma}_G^*$  is defined by  $\phi_1 \hat{A} \phi_2 \Rightarrow \phi_1 \phi \phi_2$  if  $\hat{A} \rightarrow \phi$  is a rule instance, with  $\hat{A} \in \hat{\Sigma}_G$  and  $\phi_1, \phi_2, \phi \in \hat{\Sigma}_G^*$ . The reflexive,

transitive closure of  $\Rightarrow_G$  is  $\Rightarrow_G^*$ . The language *generated* by  $G$  is  $L(G) = \{w \in \Sigma^* \mid S(w) \Rightarrow_G^* \varepsilon\}$ , where  $\varepsilon$  is the empty string.

The largest fanout of any nonterminal in a given MCFG is called the fanout of the grammar, and the largest rank of any rule is called the rank of the grammar. We call a MCFG *binary* if the rank is at most 2. Every MCFG can be brought into binary form, at the expense of a higher fanout (Rambow & Satta 1999). We assume that all considered MCFGs are *reduced*, which means that each nonterminal is involved in at least one derivation  $S(w) \Rightarrow_G^* \varepsilon$ . We say a MCFG is in *normal form* if terminals occur only in rules with rank 0, or in other words if a rule contains variables or terminals, but not both at the same time. A grammar can be brought into normal form without increasing the fanout (Seki et al. 1991: Lemma 2.2).

## 4 MCFGs and the $\mathbf{O}_n$ languages

In order to relate the  $\mathbf{O}_n$  languages to the languages generated by MCFGs, we start with a negative result.

**Theorem 1** *For any  $n \geq 2$ , the language  $\mathbf{O}_n$  is not generated by any MCFG with fanout strictly smaller than  $n$ .*

For the proof, assume that  $\mathbf{O}_n$ , for some  $n$ , is generated by a MCFG  $G$  with fanout  $n - 1$  or smaller. Without loss of generality, assume  $G$  is in normal form. We first show that if there are a nonterminal  $A$  and strings  $w_1, \dots, w_k, w'_1, \dots, w'_k$  such that  $A(w_1, \dots, w_k) \Rightarrow_G^* \varepsilon$  and  $A(w'_1, \dots, w'_k) \Rightarrow_G^* \varepsilon$ , then  $\text{imb}(w_1 \cdots w_k) = \text{imb}(w'_1 \cdots w'_k)$ . A sketch of the proof is as follows. Suppose that  $\text{imb}(w_1 \cdots w_k) \neq \text{imb}(w'_1 \cdots w'_k)$ . Then let  $w \in \Sigma^*$  be such that  $S(w) \Rightarrow_G^* \phi_1 A(w_1, \dots, w_k) \phi_2 \Rightarrow_G^* \varepsilon$ ; such a derivation must exist as we assumed grammars are always reduced. Moreover  $w \in \mathbf{O}_n$  by our initial assumption. Now replace the subderivation of  $A(w_1, \dots, w_k)$  by the subderivation of  $A(w'_1, \dots, w'_k)$ ; this is possible by the context-freeness of MCFGs. Thereby we obtain  $S(w') \Rightarrow_G^* \phi_1 A(w'_1, \dots, w'_k) \phi_2 \Rightarrow_G^* \varepsilon$ , for some  $w'$  such that  $\text{imb}(w) - \text{imb}(w') = \text{imb}(w_1 \cdots w_k) - \text{imb}(w'_1 \cdots w'_k)$ . But since  $\text{imb}(w_1 \cdots w_k) - \text{imb}(w'_1 \cdots w'_k) \neq 0^n$  and  $\text{imb}(w) = 0^n$ , we must have  $\text{imb}(w') \neq 0^n$ , which violates the assumption that  $L(G) = \mathbf{O}_n$ .

We conclude that each nonterminal  $A$  can be associated with a unique  $n$ -tuple  $\tau_A$  such that  $A(w_1, \dots, w_k) \Rightarrow_G^* \varepsilon$  implies  $\text{imb}(w_1 \cdots w_k) = \tau_A$ . Let  $d$  be  $\max_{A, \tau_A=(d_1, \dots, d_n), i} |d_i|$ , or in other words, the largest absolute point-wise imbalance in any of the  $n$  pairs  $(a_i, \bar{a}_i)$  of symbols for any nonterminal  $A$ . Let  $\rho$  be the rank of the grammar.

Now consider the balanced string:

$$w = \bar{a}_1^{(n-1)m} (a_1 \cdots a_n)^m \bar{a}_2^{(n-1)m} (a_1 \cdots a_n)^m \cdots (a_1 \cdots a_n)^m \bar{a}_n^{(n-1)m}$$

for  $m = \rho(2d + n) + d$ . In a derivation of  $w$ , from the root downwards, consider the first nonterminal instance of the form  $A(w_1, \dots, w_k)$  where  $2d + n \leq$

$|w_1 \cdots w_k|_{\bar{a}_1} < \rho(2d + n)$ ; such an instance always exists, as the number of occurrences of  $\bar{a}_1$  in the left-hand side of a rule instance and a nonterminal instance in its right-hand side can differ by at most a factor  $\rho$ . This implies  $d + n \leq |w_1 \cdots w_k|_{\bar{a}_1} < \rho(2d + n) + d = m$ , by the definition of  $d$ . As  $w$  contains  $n - 1$  substrings of the form  $(a_1 \cdots a_n)^m$  and  $a_1$  occurs nowhere else, it follows that  $w_1, \dots, w_k$  must contain at least some non-empty parts of these substrings  $(a_1 \cdots a_n)^m$ , including at least  $(d + n) - (n - 1) = d + 1$  occurrences of each of  $a_2, \dots, a_n$ , due to  $k \leq n - 1$  by our assumptions. No substring  $(a_1 \cdots a_n)^m$  can be entirely included in any of  $w_1, \dots, w_k$  however, as  $|w_1 \cdots w_k|_{\bar{a}_1} < m$ . Moreover,  $w_1 \cdots w_k$  must include at least  $(d + 1) - d = 1$  occurrence of each of  $\bar{a}_2, \dots, \bar{a}_n$ , so that some non-empty part of each of the  $n$  substrings  $\bar{a}_i^{(n-1)m}$  of  $w$  must be included in  $w_1, \dots, w_k$ . This is impossible because  $k < n$ , which completes the proof. ■

As the above proof of non-existence fails for fanout greater than or equal to  $n$ , one may suspect the following.

**Conjecture 1** *For any  $n \geq 2$ , the language  $\mathbf{O}_n$  is generated by a binary MCFG of fanout  $n$ .*

For  $n \geq 2$ , let the binary MCFG  $G_n$  of fanout  $n$  with alphabet  $\Sigma_n$  be defined by the following rules:

$$\begin{aligned}
S(x_1 \cdots x_n) &\rightarrow A(x_1, \dots, x_n) \\
A(w_1, \dots, w_n) &\rightarrow \varepsilon, \text{ for all } w_1, \dots, w_n \in \Sigma_n \cup \{\varepsilon\} \text{ such that} \\
&\quad |w_1 \cdots w_n| \leq 2 \text{ and } \text{imb}(w_1 \cdots w_n) = 0^n \\
A(s_1, \dots, s_n) &\rightarrow A(x_1, \dots, x_n) A(y_1, \dots, y_n), \text{ for all non-empty} \\
&\quad s_1, \dots, s_n \text{ such that } |s_1 \cdots s_n| = 2n, s_1 \text{ starts with } x_1, \\
&\quad x_1 \cdots x_n \text{ and } y_1 \cdots y_n \text{ are subsequences of } s_1 \cdots s_n, \\
&\quad \text{and no } s_i \text{ has a substring of the form } x_j x_{j+1} \text{ or } y_j y_{j+1}
\end{aligned}$$

We now wish to strengthen Conjecture 1 to:

**Conjecture 2** *For any  $n \geq 2$ ,  $L(G_n) = \mathbf{O}_n$ .*

Even stronger is the following:

**Conjecture 3** *For any  $n \geq 2$ , and  $w_1, \dots, w_n \in \Sigma_n^*$  such that  $\text{imb}(w_1 \cdots w_n) = 0^n$ , we have  $A(w_1, \dots, w_n) \Rightarrow_{G_n}^* \varepsilon$ .*

It is clear that if Conjecture 3 is true, then so is Conjecture 2. The added value of Conjecture 3 is that it would exclude the possibility of making a ‘wrong’ derivation step, as long as all nonterminal instances contain arguments that together are balanced. For example, how a string  $w \in \mathbf{O}_n$  is divided into  $n$  parts in the first step using an instantiated rule  $S(w_1 \cdots w_n) \rightarrow A(w_1, \dots, w_n)$ , with  $w = w_1 \cdots w_n$ , would not affect whether we can complete the derivation.

A proof of Conjecture 3 would likely be by induction, first on the length of  $w_1 \cdots w_n$ , and second on the number of arguments among  $w_1, \dots, w_n$  that are  $\varepsilon$ . The base case is obviously  $w_1 = \dots = w_n = \varepsilon$ .

The inductive step is straightforward if at least one of the arguments, say  $w_i$ , is  $\varepsilon$ . Two subcases can be distinguished. In the first, at least one of the remaining arguments, say  $w_j$ , has length 2 or greater. Assume without loss of generality that  $1 < i < j$ . We can then use a rule of the form  $A(s_1, \dots, s_n) \rightarrow A(x_1, \dots, x_n) A(y_1, \dots, y_n)$ , where:

$$s_k = \begin{cases} x_k y_k & \text{if } 1 \leq k < i \\ y_k & \text{if } k = i \\ x_{k-1} y_k & \text{if } i < k < j \\ x_{k-1} y_k x_k & \text{if } k = j \\ x_k y_k & \text{if } j < k \leq n \end{cases}$$

In the required rule instance, we would replace each  $y_k$  ( $1 \leq k \leq n$ ) by  $\varepsilon$ , replace each  $x_k$  by  $w_k$  if  $k < i$  or  $j < k$ , replace each  $x_{k-1}$  by  $w_k$  if  $i < k < j$ , and replace  $x_{j-1}$  and  $x_j$  by non-empty strings  $w'$  and  $w''$ , respectively, such that  $w_j = w'w''$ . We also use the rule  $A(\varepsilon, \dots, \varepsilon) \rightarrow \varepsilon$ , together with the inductive hypothesis for a nonterminal instance in which one argument fewer is  $\varepsilon$ .

In the second subcase, all the non-empty arguments are of length 1. There must then be two arguments, say  $w_i$  and  $w_j$ , that are  $a_\ell$  and  $\overline{a_\ell}$ , respectively, for some  $\ell$  ( $1 \leq \ell \leq n$ ). We can then use a rule of the form  $A(s_1, \dots, s_n) \rightarrow A(x_1, \dots, x_n) A(y_1, \dots, y_n)$ , where each  $s_k$  ( $1 \leq k \leq n$ ) is  $x_k y_k$ . In the required rule instance, we would replace each  $x_k$  and  $y_k$  ( $1 \leq k \leq n$ ) by  $w_k$  and  $\varepsilon$ , respectively, if  $k \notin \{i, j\}$ , and by  $\varepsilon$  and  $w_k$ , respectively, if  $k \in \{i, j\}$ . We also use a rule of the form  $A(v_1, \dots, v_n) \rightarrow \varepsilon$ , with  $v_i = a_\ell$  and  $v_j = \overline{a_\ell}$ , and  $v_k = \varepsilon$  for  $k \notin \{i, j\}$ , together with the inductive hypothesis.

The inductive step is less straightforward if  $w_1, \dots, w_n$  are all non-empty. We then need to show that there is a sequence of  $2n$  strings  $v_1, \dots, v_{2n}$  such that:

- there is a sequence of positive integers  $k_1, \dots, k_n$  such that for each  $i$  ( $1 \leq i \leq n$ ) we have  $w_i = v_{m_{i-1}+1} \cdots v_{m_i}$ , where  $m_i = \sum_{j:1 \leq j \leq i} k_j$  ( $0 \leq i \leq n$ ), and  $m_n = 2n$ , and
- there is a permutation  $u_1, \dots, u_n, u'_1, \dots, u'_n$  of  $v_1, \dots, v_{2n}$ , such that  $\text{imb}(u_1 \cdots u_n) = \text{imb}(u'_1 \cdots u'_n) = 0^n$ ,  $|u_1 \cdots u_n| > 0$  and  $|u'_1 \cdots u'_n| > 0$ .

In words, a balanced string divided into  $n$  non-empty parts can be further divided into  $2n$  smaller parts, and in particular the  $i$ -th part is divided into  $k_i$  smaller parts, and the  $2n$  smaller parts can be partitioned to form two other balanced (but non-empty) strings, each again divided into  $n$  parts.

Special treatment can be given to cases where  $w_1, \dots, w_n$  are all non-empty, but  $\text{imb}(v_1 \cdots v_k) = 0^n$  for a proper non-empty subset  $\{v_1, \dots, v_k\}$  of  $\{w_1, \dots, w_n\}$ , by deriving:

$$A(w_1, \dots, w_n) \Rightarrow A(v_1, \dots, v_k, \varepsilon, \dots, \varepsilon) A(u_1, \dots, u_n)$$

for some  $u_1, \dots, u_n$ , which allows use of the inductive hypothesis. Hence the interesting case that remains is where  $\text{imb}(v_1 \cdots v_k) = 0^n$  does *not* hold for any proper non-empty subset  $\{v_1, \dots, v_k\}$  of  $\{w_1, \dots, w_n\}$ .

## 5 Special cases

### 5.1 The $\mathbf{O}_1$ language

The case  $n = 1$  has been ignored in the above. It is straightforward to show that  $\mathbf{O}_1$  is generated by an MCFG of fanout 1, but this grammar  $G_1$  has a slightly different structure from the grammars  $G_n$  ( $n \geq 2$ ) that were defined above. The rules of  $G_1$  are  $S(xy) \rightarrow S(x)S(y)$ ,  $S(a_1x\bar{a}_1) \rightarrow S(x)$ ,  $S(\bar{a}_1xa_1) \rightarrow S(x)$ , and  $S(\varepsilon) \rightarrow \varepsilon$ .

The central observation in the proof by induction concerns strings in  $\mathbf{O}_1$  of the form  $a_1wa_1$  or of the form  $\bar{a}_1w\bar{a}_1$ . In the first case, the imbalance of the prefix  $a_1$  is a positive number and the imbalance of the prefix  $a_1w$  is a negative number. This implies that there must be a proper prefix  $a_1w'$  of  $a_1w$  whose imbalance is  $0^1$ , which means we can use the rule  $S(xy) \rightarrow S(x)S(y)$  and the inductive hypothesis for two shorter strings. The second case is symmetric.

### 5.2 The $\mathbf{O}_2$ language

Conjecture 3 restricted to  $n = 2$  was proved by Salvati (2015), using arguments involving considerable sophistication. The proof is geometric in nature, interpreting the imbalance of a series of prefixes of a string in  $\mathbf{O}_2$  of increasing length as a path in 2-dimensional space. The use of the complex exponential function seems to make the proof difficult to generalize to higher dimensions.

An alternative proof is due to Nederhof (2016). It is similarly geometric in nature, but avoids the complex exponential function. Its core argument divides 2-dimensional space into an ‘above’ and a ‘below’. We will refer to this as the ‘partition argument’. Before the argument can be applied, the paths must first be brought into a normal form.

The proof requires all four binary rules of  $G_2$ :

$$\begin{aligned} A(x_1y_1, x_2y_2) &\rightarrow A(x_1, x_2) A(y_1, y_2) \\ A(x_1y_1, y_2x_2) &\rightarrow A(x_1, x_2) A(y_1, y_2) \\ A(x_1y_1x_2, y_2) &\rightarrow A(x_1, x_2) A(y_1, y_2) \\ A(x_1, y_1x_2y_2) &\rightarrow A(x_1, x_2) A(y_1, y_2) \end{aligned}$$

### 5.3 The $O_3$ language

Nederhof (2016) also sketches a potential generalization of the proof to  $n = 3$ . The partition argument now relies on the (three plus six) rules:

$$\begin{aligned}
 A(x_1y_1, x_2y_2, y_3x_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1y_1, y_2x_2, x_3y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(y_1x_1, x_2y_2, x_3y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1y_1x_2, x_3y_2, y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1y_1x_2, y_2, x_3y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1y_1, x_2y_2x_3, y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1y_1, y_2, x_2y_3x_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1, y_1x_2y_2, y_3x_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1, y_1x_2, y_2x_3y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3)
 \end{aligned}$$

In addition, we need a ‘corkscrew argument’, which requires three further rules:

$$\begin{aligned}
 A(x_1y_1x_2y_2, x_3, y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1, y_1x_2y_2x_3, y_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3) \\
 A(x_1, y_1, y_2x_2y_3x_3) &\rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3)
 \end{aligned}$$

It is remarkable that these 12 rules are only a portion of the 22 binary rules of  $G_3$ . If however we remove any of the last three rules, then we cannot always use the inductive hypothesis. For example, if we remove:

$$A(x_1y_1x_2y_2, x_3, y_3) \rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3)$$

then we can no longer handle  $A(\overline{a_3} \overline{a_2} \overline{a_2} \overline{a_2} a_3 a_1 a_1 a_2, a_2 \overline{a_1}, a_2 \overline{a_1})$ . The same applies to the group of six rules. For example, if we remove:

$$A(x_1y_1x_2, x_3y_2, y_3) \rightarrow A(x_1, x_2, x_3) A(y_1, y_2, y_3)$$

then we can no longer handle  $A(a_3 a_2 a_2 a_1 \overline{a_3} a_1 \overline{a_2}, \overline{a_1} \overline{a_3} \overline{a_1}, a_3 \overline{a_2})$ . This can be verified by mechanically matching these nonterminal instances against the remaining rules. We have not been able to ascertain that more than one of the first three rules is necessary to always be able to apply the inductive hypothesis. Hence we cannot exclude the possibility at this time that only 10 binary rules would suffice.

The main difficulty in obtaining a complete proof of Conjecture 3 restricted to  $n = 3$  pertains to the partition argument. This would again depend on a normal form for paths, this time in 3-dimensional space. It seems much more difficult than before to show that the normal form can be obtained while preserving appropriate invariants.

## 6 Conclusions

It is trivial to show that  $O_1$  is generated by  $G_1$ , whereas it took great efforts to find the first proof that  $O_2$  is generated by  $G_2$ . The second proof of the same result seems to create realistic prospects that a proof may one day be found that (a subgrammar of)  $G_3$  generates  $O_3$ , but considerable challenges lie ahead. Very little is known for  $n \geq 4$ .

## Acknowledgements

This work came out of correspondence with Giorgio Satta. Gratefully acknowledged are also fruitful discussions with Sylvain Salvati, Vinodh Rajan, and Markus Pfeiffer.

## References

- Aho, A. V. & J. D. Ullman. 1972. *Parsing*. Vol. 1 (The Theory of Parsing, Translation and Compiling). Englewood Cliffs, N.J.: Prentice-Hall.
- Bach, E. 1981. Discontinuous constituents in generalized categorial grammars. In *Proceedings of the eleventh Annual Meeting of the North Eastern Linguistic Society*, 1–12.
- Chomsky, N. 1959. On certain formal properties of grammars. *Information and Control* 2. 137–167.
- Fischer, M. J. & A. L. Rosenberg. 1968. Real-time solutions of the origin-crossing problem. *Mathematical Systems Theory* 2. 257–263.
- Gazdar, G. 1988. Applicability of indexed grammars to natural languages. In U. Reyle & C. Rohrer (eds.), *Natural language parsing and linguistic theories*, 69–94. D. Reidel Publishing Company.
- Joshi, A. K. 1985. Tree adjoining grammars: how much context-sensitivity is required to provide reasonable structural descriptions? In D. R. Dowty, L. Karttunen & A. M. Zwicky (eds.), *Natural language parsing: psychological, computational, and theoretical perspectives*, 206–250. Cambridge University Press.
- Kallmeyer, Laura. 2010. *Parsing beyond context-free grammars*. Springer-Verlag.
- Kanazawa, M. & S. Salvati. 2012. MIX is not a tree-adjoining language. In *50th Annual Meeting of the Association for Computational Linguistics, proceedings of the conference*, 666–674. Jeju Island, Korea.
- Latteux, M. 1979. Cônes rationnels commutatifs. *Journal of Computer and System Sciences* 18. 307–333.
- Nederhof, Mark-Jan. 2016. A short proof that  $O_2$  is an MCFL. In *54th Annual Meeting of the Association for Computational Linguistics, proceedings of the conference*, vol. 1, 1117–1126. Berlin.
- Rambow, O. & G. Satta. 1999. Independent parallelism in finite copying parallel rewriting systems. *Theoretical Computer Science* 223. 87–120.
- Salvati, S. 2015. MIX is a 2-MCFL and the word problem in  $\mathbb{Z}^2$  is captured by the IO and the OI hierarchies. *Journal of Computer and System Sciences* 81. 1252–1277.

*Mark-Jan Nederhof*

- Seki, H., T. Matsumura, M. Fujii & T. Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science* 88. 191–229.
- Vijay-Shanker, K. & D. J. Weir. 1994. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory* 27. 511–546.
- Vijay-Shanker, K., D. J. Weir & A. K. Joshi. 1987. Characterizing structural descriptions produced by various grammatical formalisms. In *25th Annual Meeting of the Association for Computational Linguistics, proceedings of the conference*, 104–111. Stanford, California, USA.

## Chapter 29

# Keystroke dynamics for authorship attribution

Barbara Plank

University of Groningen

We examine to what extent information from keystroke dynamics reflects individual style for authorship attribution. We compare models that use keystroke dynamics to more traditional authorship attribution methods. Our results show that biometric features are more predictive of authorship than stylometric features.

Authorship attribution is the task of identifying the author of a text. It can be viewed as a special form of text classification in the field of stylometry, which more broadly speaking includes the identification of author traits (like identity, or gender, age, personality etc).

As noted by Nerbonne (2007): “A key question in authorship attribution has been to determine what sorts of *evidence* might bear on determining authorship.” Traditionally, authorship studies focused on finding evidence in the *text* produced by authors, and examined, e.g., high-frequency elements.

However, as people produce text, they unconsciously produce loads of cognitive by-product. Can we use such meta-data as additional evidence of authorship? Examples of cognitive processing data include brain activations, gaze pattern, or keystroke dynamics. In this paper we focus on the latter. *Keystroke dynamics* concerns a user’s typing pattern and keystroke logs are the recordings of a user’s typing dynamics. When a person types on a keyboard, the latencies between successive keystrokes and their duration reflect the presumably unique typing behavior of a person. Assuming access to keystroke logs, to what extent are they informative for authorship attribution, i.e., do they help identifying the author of a piece of text?

Keystroke logs are studied mostly in cognitive writing and translation process research to gain insights into the cognitive load involved in the writing process. Only very recently this source has been explored as information in natural language processing (NLP), in particular for shallow syntactic parsing (Plank 2016). Keystroke logs have been used in computer security for user verification, however, combining keystroke biometrics with traditional stylometry metrics has not yet been proven successful (Stewart et al. 2011). In this paper we examine to what extent keystroke

dynamics are informative for authorship attribution. We compare them to more traditional stylometry features, and investigate various ways to combine them.

## 1 Keystroke dynamics

Keystroke dynamics provide a complementary view on a user's style beyond the linguistic signal.

A major scientific interest in keystroke dynamics arose in writing research, where it has developed into a promising non-intrusive method for studying cognitive processes involved in writing (Sullivan & Lindgren 2006; Nottbusch, Weingarten & Sahel 2007; Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012). In these studies time measurements—pauses, bursts and revisions—are studied as traces of the recursive nature of the writing process. *Bursts* are defined as consecutive chunks of text produced and defined by a 2000ms time of inactivity (Wengelin 2006).

Keystroke logs have the distinct advantage over other cognitive modalities like eye tracking or brain scanning that they are readily available and can be harvested easily. They do not rely on special equipment beyond a keyboard. Moreover, they are non-intrusive, inexpensive, and have the potential to offer continuous adaptation to specific users. Imagine integrating keystroke logging into (online) text processing tools.

In its raw form, keystroke logs contain information on which key was pressed for how long (key, time press, time release). Research on keystroke dynamics typically considers a number of timing metrics, such as *holding time* and *time press* and *time release* between keystrokes, e.g.,  $p$  in Figure 1, inspired by the figure in (Goodkind & Rosenberg 2015). An example of raw keystroke log data is shown in Table 1.

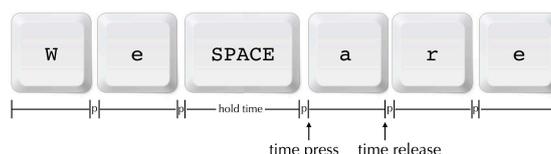


Figure 1: Keystroke logs illustrated:  $p$  are pauses between keystrokes.

Raw keystroke log data can be used to calculate keystroke pause durations, such as pre-word pauses. However, if we examine the literature we find different ways to define the duration of pauses. Stewart et al. (2011) and Goodkind & Rosenberg (2015) use the difference between release time of the previous key and the timepress of the current key to calculate keystroke (or pre-word durations). In contrast, writing research (Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012) defines pauses as the start time of a keystroke until the start time for the next keystroke. In this paper we follow user authentication studies (Stewart et al. 2011) and use the former definition of pause duration.

Table 1: Example of the raw keystroke logging data.

user	session	timepress	timerelease	keycode	keyname
1	1	1304433167859	1304433168307	16	shift
1	1	1304433168227	1304433168371	67	c
1	1	1304433168291	1304433168451	79	o
1	1	1304433170051	1304433170179	69	e
1	1	1304433170451	1304433170531	70	f
1	1	1304433170579	1304433170675	70	f
1	1	1304433170675	1304433170851	73	i
1	1	1304433171171	1304433171299	67	c
1	1	1304433172179	1304433172275	8	backspace

A challenge when using keystroke log data is that the typing behavior of users typically differ. For instance, Figure 2 plots the distribution of keystroke durations for two different users. Keystroke logs are presumably idiosyncratic. In fact, they were successfully used for author verification in computer security research (Stewart et al. 2011; Monaco et al. 2013; Locklear et al. 2014). In this paper we study how predictive biometric features are for authorship, as compared to more traditional features obtained from the text alone, and whether combining the two sources aids authorship attribution.

## 2 Experiments

Given a dataset with keystroke logs from 38 authors, the aim of our experiments is to classify who of the authors wrote a piece of text.

### 2.1 Dataset

The keystroke logging data stems from students taking an actual test on spreadsheet modeling in a university course (Stewart et al. 2011). The dataset was collected during an exam, and as such represents free-text input. The dataset contains data from 38 users for several sessions.<sup>1</sup> We take the first two sessions as development data (resulting in 76 instances), session 3-5 as test section (114 instances), and the remaining session as training sections (total 856 instances).

### 2.2 Setup

As classification system we use Support Vector Machines (SVM), implemented in `sk-learn`.<sup>2</sup> For all experiments we use the same hyperparameters, i.e., SVM with

<sup>1</sup> Following Stewart et al. (2011) some users were discarded due to issues with logging.

<sup>2</sup> The code for our experiments is available at: <https://github.com/bplank/festschrift>.

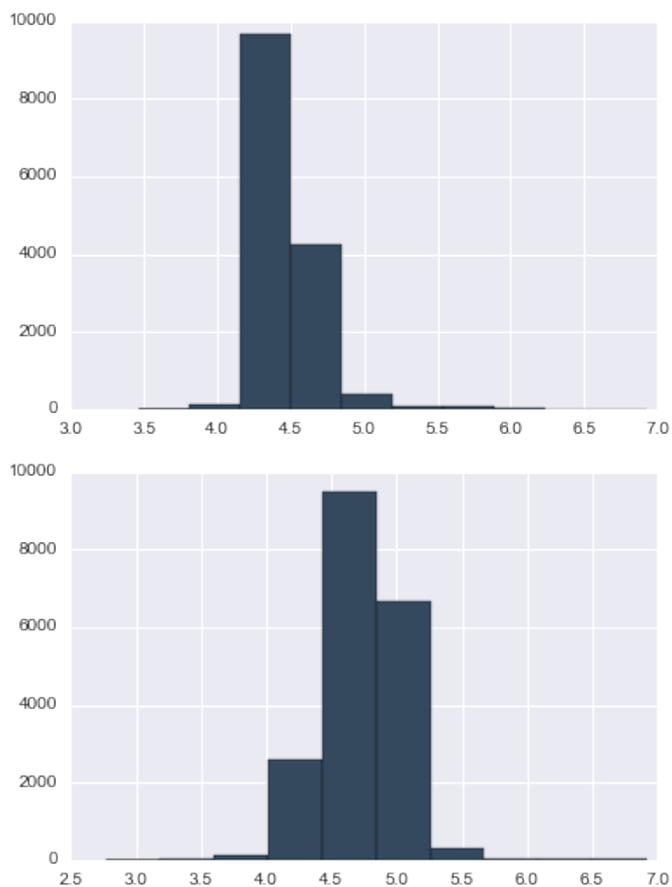


Figure 2: Distribution of pauses for two users (plotted in log space): user 3 (top), user 20 (bottom).

default  $C$  and a linear kernel. SVMs were chosen in preliminary experiments as they outperformed alternative approaches (logistic regression, naive Bayes).

### 2.3 Features

We use 218 biometric features following Stewart et al. (2011), who in turn follow Tappert et al. (2010). These biometric features include duration features (mean and standard deviation) and are grouped roughly into: duration features of individual letters (which we later refer to as keystroke basic), and transition features between letters or groups of letters, between letters and non-letters and overall percentage features.

We use the freely available feature extractor<sup>3</sup> and test two configurations: using only letter durations (52 keystroke basic features) and all duration features (keystrokes extended, 218 features).

For the textual features, we extracted the final text from the keystroke logging data (using revisions to alter the text to obtain the final output). We then use commonly used authorship attribution features, binary indicator features for character n-grams and word n-grams. We also evaluated only pronouns, however, that resulted in worse performance, thus we do not further elaborate on it. In addition, we examine *word embedding* features estimated from a large English Wikipedia dump (Al-Rfou, Perozzi & Skiena 2013). We represent each text as the mean average activation over all word embeddings (Collobert et al. 2011), resulting in 64 features. In contrast to the previous *sparse* n-gram feature representations, this represents adding a *dense* feature vector that represents the text, and is more similar to the standardized keystroke features (both in terms of value and number of features).

### 3 Results

The results of training a classifier to predict the identify on an author are given in Table 2. A random baseline obtains an accuracy rate of only 2% on this dataset (38 authors). The stylistic features based on the text obtain an accuracy of around 28-37%.

Table 2: Accuracy for authorship attribution (38 authors), comparison of stylometry features (word and character n-grams) versus biometric stylometry (keystroke dynamics) and combined (embeds: word embeddings).

FEATURES	num features	DEV	TEST
<i>Stylistic:</i>			
character 3grams	8.3k	23.68	28.07
character 2+3grams	9.8k	25.00	31.58
word unigrams	8.9k	27.64	30.70
word unigrams +char 2+3grams	18.7k	25.00	37.72
<i>Biometrics:</i>			
keystrokes (basic)	52	72.37	71.05
keystrokes (extended)	218	81.58	77.19
<i>Combined:</i>			
keystrokes (basic)+word unigrams	8.9k	55.26	50.88
keystrokes (ext.)+word unigrams	9.1k	65.79	67.65
keystroke (basic)+embeds	116	73.68	71.93
keystrokes (extended) +embeds	282	80.26	78.07

<sup>3</sup> <https://bitbucket.org/vmonaco/keystroke-feature-extractor>.

*Barbara Plank*

Using keystroke dynamics results in substantial performance gains. Already the basic feature set using 52 letter duration features clearly outperforms the stylistic features, reaching an accuracy of 71% on the test data. Adding keystroke transition durations further boost performance to 77%. These are remarkable results in light of the low baseline given by the rather large number of candidate authors. In fact, as the number of authors increases, authorship attribution becomes increasingly more difficult (Luyckx & Daelemans 2008).

## 4 Discussion

Our results show that the biometric keystroke features are more predictive of authorship than the stylometric features. This confirms earlier findings (Stewart et al. 2011), however, they used a simpler setup (binary classification). However, it is not straightforward to combine these two sources of information. Adding plain n-gram features (character or word n-grams) results in performance drops. In that case we add a high-dimensional sparse feature space to the dense duration feature, most probably these large amount of features swamp the feature space. In contrast, if we use word embeddings, we model the user's text as average point in a high-dimensional space and effectively add a dense low-dimensional vector to the keystroke dynamics data. This gives at times slight improvements, albeit not significant on our relatively small test and development set.

To examine which kind of features are highly predictive, we train a logistic regression model on our best configuration (extended keystrokes and embeddings) and examine the most predictive features. We see that mostly duration features of non-letter symbols are among the most predictive features, in particular punctuation symbols and spaces. This is intuitively pleasing, as users exhibit different behavior at word and sentence boundaries (Wengelin 2006).

## 5 Related Work

Authorship attribution has a long tradition dating back to early works in the 19th century. The most influential work on authorship attribution goes back to Mosteller & Wallace (1964), who construe it as a text classification problem (Nerbonne 2007). For a long time statistical approaches to authorship attribution focused on distributions of *function words*, high-frequency words that are presumably not consciously manipulated by the author (Nerbonne 2007; Pennebaker 2011). For example, in the well-known Federalist papers, *enough* and *while* were used exclusively by Hamilton, while *whilst* was prototypical for Madison. An early study using neural networks to infer the author of the disputed documents of the Federalist papers used 11 function words as predictive features (Tweedie, Singh & Holmes 1996). Recent work also includes authorship studies on microblog texts (Rappoport & Koppel 2013). An excellent recent summary is Stamatatos (2009).

Keystroke logging has developed into a promising tool for research into writing (Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012), as time measurements can give insights into cognitive processes involved in writing (Nottbusch, Weingarten & Sahel 2007) or translation studies. In fact, most prior work that uses keystroke logs focuses on experimental research. For example, Hanouille, Hoste & Remael (2015) study whether a bilingual glossary reduces the working time of professional translators. They consider pause durations before terms extracted from keystroke logs and find that a bilingual glossary in the translation process of documentaries reduces the translators' workload. Other translation research has combined eye-tracking data with keystroke logs to study the translation process (Carl et al. 2016). An analysis of users' typing behavior was studied by Baba & Suzuki (2012). They collect keystroke logs of online users describing images to measure spelling difficulty. They analyzed corrected and uncorrected spelling mistakes in Japanese and English and found that spelling errors related to phonetic problems remain mostly unnoticed.

It has been shown that pauses reflect the planning of the unit of text itself (Baaijen, Galbraith & de Glopper 2012) and that they correlate with clause and sentence boundaries (Spelman Miller & Sullivan 2006). Goodkind & Rosenberg (2015) investigate the relationship between pre-word pauses and multi-word expressions. They found that within MWE pauses vary depending on the cognitive task. Taking writing research as a starting point, a recent study postulated that keystrokes contain fine-grained information that aids the identification of syntactic chunks (Plank 2016). They integrated automatically derived labels from keystroke logs as auxiliary task in a multi-task setup (Plank, Søgaard & Goldberg 2016) with promising results. Instead, this paper focuses on the idiosyncrasy of keystroke patterns. Our results show that keystroke biometrics are far superior to that of a stylometry-based approach to authorship attribution. At the same time it is challenging to combine the two sources of information. This confirms earlier findings by the most related study (Stewart et al. 2011). They combine keystroke log features and linguistic stylometry features for user verification using a  $k$ -nearest neighbor approach. Their study differs from ours in two aspects. First, they use different stylometric features, i.e., the number of a specific set of characters, number of words of a certain length, average word length and number of punctuation symbols, see the full list in the appendix of their paper. Second, they target user authentication, thus their setup is a binary classification task (authenticated vs. not-authenticated), while we here focus on a multi-class classification setup (who wrote the piece of text out of all possible authors).

## 6 Conclusions

We have shown that keystroke dynamics contain highly indicative information to predict the authorship of a text. We compared keystroke dynamics to more traditional authorship attribution features and found keystroke biometrics to be superior. In particular, duration features of punctuation and spaces are highly predictive of authorship. However, combining keystrokes and linguistic features, two very different

Barbara Plank

feature spaces, proves difficult. Some promising initial results are obtained by using word embeddings, however, further investigations are needed to test the robustness of this direction.

## References

- Baaijen, Veerle M., David Galbraith & Kees de Glopper. 2012. Keystroke analysis: reflections on procedures and measures. *Written Communication*.
- Baba, Yukino & Hisami Suzuki. 2012. How are spelling errors generated and corrected?: a study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, vol. 2.
- Carl, Michael, Isabel Lacruz, Masaru Yamada & Akiko Aizawa. 2016. Measuring the translation process. In *The 22nd Annual Meeting of the Association for Natural Language Processing*.
- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa & Michael Collins. 2011. Natural language processing (almost) from scratch.
- Goodkind, Adam & Andrew Rosenberg. 2015. Muddying the multiword expression waters: how cognitive demand affects multiword expression production. In *Proceedings of MWE 2015*.
- Hanoulle, Sabien, Véronique Hoste & Aline Remael. 2015. The translation of documentaries: can domain-specific, bilingual glossaries reduce the translators' workload? an experiment involving professional translators. *New Voices in Translation Studies* (13).
- Locklear, Hilbert, Sathya Govindarajan, Zdenka Sitova, Adam Goodkind, David Guy Brizan, Andrew Rosenberg, Vir V. Phoha, Paolo Gasti & Kiran S. Balagani. 2014. Continuous authentication with cognition-centric text production and revision features. In *Biometrics (IJCB), 2014 IEEE International Joint Conference*.
- Luyckx, Kim & Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1, 513–520. Association for Computational Linguistics.
- Monaco, John V., John C. Stewart, Sung-Hyuk Cha & Charles C. Tappert. 2013. Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In *Biometrics: theory, applications and systems (BTAS), 2013 IEEE Sixth International Conference on*.
- Mosteller, Frederick & David Wallace. 1964. Inference and disputed authorship: the Federalist.
- Nerbonne, John. 2007. The exact analysis of text. *Foreword to the 3rd edition of Frederick Mosteller and David Wallace Inference and Disputed Authorship: The Federalist Papers CSLI: Stanford*.

- Nottbusch, Guido, Rüdiger Weingarten & Said Sahel. 2007. From written word to written sentence production. *Writing and cognition: Research and applications*. Mark Torrance, Luuk van Waes & David W. Galbraith (eds.). 31–54.
- Pennebaker, James W. 2011. Using computer analyses to identify language style and aggressive intent: the secret life of function words. *Dynamics of Asymmetric Conflict* 4(2). 92–102.
- Plank, Barbara. 2016. Keystroke dynamics as signal for shallow syntactic parsing. In *The 26th International Conference on Computational Linguistics (COLING)*.
- Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL*.
- Rappoport, Roy Schwartz Oren Tsur Ari & Moshe Koppel. 2013. Authorship attribution of micro-messages. In *EMNLP*.
- Al-Rfou, Rami, Bryan Perozzi & Steven Skiena. 2013. Polyglot: distributed word representations for multilingual NLP. In *CoNLL*.
- Spelman Miller, Kristyan & Kirk P. H. Sullivan. 2006. *Keystroke logging: an introduction*. Elsevier.
- Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3). 538–556.
- Stewart, John C., John V. Monaco, Sung-Hyuk Cha & Charles C. Tappert. 2011. An investigation of keystroke and stylometry traits for authenticating online test takers. In *Biometrics (IJCB), 2011 International Joint Conference*.
- Sullivan, Kirk P. H., Eva Lindgren, et al. 2006. *Computer keystroke logging and writing: methods and applications*. Elsevier.
- Tappert, Charles C., Sung-Hyuk Cha, Mary Villani & Robert S. Zack. 2010. A keystroke biometric system for long-text input. *International Journal of Information Security and Privacy (IJISP)* 4. 32–60.
- Tweedie, Fiona J., Sameer Singh & David I. Holmes. 1996. Neural network applications in stylometry: the Federalist papers. *Computers and the Humanities* 30(1). 1–10.
- van Waes, Luuk, Mariëlle Leijten & Daphne van Weijen. 2009. Keystroke logging in writing research: observing writing processes with inputlog. *GFL-German as a foreign language* 2(3). 41–64.
- Wengelin, Åsa. 2006. Examining pauses in writing: theory, methods and empirical data. *Computer key-stroke logging and writing: methods and applications*. (Studies in Writing) 18. 107–130.



## Chapter 30

# Quantitative diachronic dialectology

Jelena Prokić

Philipp University of Marburg

In this paper quantitative analysis of sound correspondences in German dialects is presented. This method combines geographical spread of sound change and regularity of sound change within a lexicon. By combining the two, language diffusion processes can be successfully modeled, enabling the researchers not only to identify main dialect groups, but also to look into the individual processes of sound change and track their geographic spread.

### 1 Introduction

Languages change constantly and this has already been observed by scholarly people in ancient times. While in ancient times language change was considered a ‘decay’, or ‘corruption of a holy speech’, modern linguistics sees it as a fact, i.e., an everlasting process that does not lead to ‘better’ or ‘worse’ languages, but tries to explain how and why languages change. Unlike historical linguistics, whose main goal is to determine relatedness between languages and language families, dialectology is a study of subdivisions of a particular language and deals with the variation at the micro-level. Apart from determining main dialect groups within a given language area, the focus of dialect research is determining the underlying linguistic processes that lead to the observed distributions of linguistic features. One of the central research questions in dialectology is to discover patterns of dialect interaction, e.g., if a change is spreading in a continuous or discontinuous pattern, determine *focal*, *relic* and *transition* areas of language change and main factors that lead to observed change and its spread.

In the past three decades there has been an increasing interest in using quantitative methods to address these questions in dialect research. Computational and statistical methods have successfully been applied in tasks that, among many, include automatic detection of dialect groups (Nerbonne et al. 1996; Gooskens & Heeringa 2004; Bolognesi & Heeringa 2002), identification of linguistic features responsible for observed dialect divisions (Wieling & Nerbonne 2011; Prokić, Çöltekin & Nerbonne 2012) and analysis of social determinants of dialect variation (Wieling, Nerbonne &

Baayen 2011). Despite significant progresses in the field of dialectometry, diachronic studies are a rather recent initiative.

In this paper, I will present quantitative analysis of German dialects, in which a network model of dialect change is combined with the automatically extracted regular sound correspondences, in order to model the diffusion process of dialect change and detect centers from which language innovations started spreading from. It is a novel approach in dialectometry proposed by Prokić & Cysouw (2013), that is mainly concerned with shedding more light on the evolutionary processes in dialect formation. In the next section I will first address the importance of network models in dialect evolution, as opposed to tree models used both in traditional and quantitative historical linguistics to model language change.

## 2 Tree and network models of language evolution

The main assumption of both the comparative method used in traditional historical linguistics and phylogenetic methods exploited in quantitative historical linguistics, is that the underlying model of language change is tree-like. Tree-like model of evolution is based on the assumption that proto-language was completely uniform, without any internal variation, that this proto-language split suddenly into two or more daughter languages, which do not have further contact once they split, and that sound change is completely regular (Bloomfield 1973; Campbell 2004). However, historical linguists are aware that these assumptions are ‘reasonable idealizations’ (Campbell 2004), since abundant linguistic material shows that proto-languages were not uniform and the split of daughter languages is never sudden. A family tree model alone is not sufficient for dealing with all aspects of relationships between languages, but it is capturing the genetic relatedness between languages, the most important factor in language diversification at the macro-level.

Interaction between neighboring dialects is characterized by an intensive contact on a day-to-day basis, and easy exchange of linguistic material at all levels thanks to the mutual intelligibility of dialects. Language innovations can occur at any location and spread in various geographic directions, sometimes even opposite (Hock 1991). Regions with a long settlement history are characterized by the so-called ‘diffusion model’ of language change (Chambers & Trudgill 2004: Ch. 11). The diffusion model assumes that innovations spread gradually in a wave-like manner. The area in which the change originates is the so-called *focal* area, where the change is regular. The area in which the change does not happen is the *relic* area. The third area, found in between, is the *transition* area, characterized by less regular change. The consequence of the diffusion model is that dialects form continuums, rather than clearly separated groups, and their historical relationship cannot be described using a tree-model. In order to model processes involved in diachronic dialect change, a network-model is much more suitable, as shown in Prokić & Cysouw (2013). In the network model proposed by Prokić & Cysouw (2013), each site in the data set represents a node in a network and is connected only to the neighboring sites (i.e. nodes). This kind of network representation allows us to explore the spread of language change via diffu-

sion, where changes spread in a wave-like manner from one site to the neighboring sites.

In this paper, German dialect data is analyzed using the described network model. A network representation of German sites can be seen in Figure 1, in which all localities are connected in a such way that each site is connected only to the neighboring sites. In the next section I give a brief overview of the German dialect data used in this paper, including previous quantitative analysis of the data.

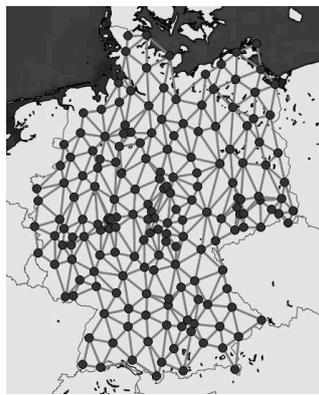


Figure 1: Network model – each site is connected to the neighboring sites.

### 3 Data

The data presented in this paper comes from the *Phonetischer Atlas von Deutschland* (PAD) which will be published as a part of the *Digitaler Wenker-Atlas* at the Forschungszentrum Deutscher Sprachatlas at the Philipps-Universität Marburg.<sup>1</sup> The PAD data consists of the pronunciations of the Wenker sentences recorded in West Germany during the 1960s and 1970s and recordings of the same sentences made in the former East Germany during the 1990s. There are in total 183 sites included in this survey and they can be seen in Figure 1.<sup>2</sup>

Recorded sentences were segmented into words and transcribed by the phoneticians at the Philipps-Universität Marburg. There are in total pronunciations of 202 words in the digitized version of the PAD data. The data was first digitized using the X-SAMPA phonetic alphabet during the project Visualisation of Language Variation, a joint project between the University of Groningen and the Forschungszentrum Deutscher Sprachatlas at the Philipps-Universität Marburg. X-SAMPA transcribed

<sup>1</sup> <http://www.diwa.info/titel.aspx>.

<sup>2</sup> With the exception of Docelles, which is on the territory of France, all other sites are on the territory of Germany. Data from Docelles was excluded from the present study.

data was later automatically converted into IPA at the Forschungszentrum Deutscher Sprachatlas.<sup>3</sup>

### 3.1 Previous work

Previous quantitative analysis of the PAD data set includes work by Nerbonne & Siedle (2005), in which the authors rely on the so-called aggregate approach in order to determine the most important dialect groups in Germany. Linguistic distances between each two sites in the data set are estimated using Levenshtein distance which compares pronunciation of two words by calculating the smallest number of operations (insertions, deletions and substitutions) needed to transform one pronunciation into another. Pronunciations of all corresponding words recorded in two sites are compared by means of Levenshtein distance, resulting in the aggregate distance between these two sites. The procedure is repeated for all pairs of sites in the data set, resulting in an  $n \times n$  distance matrix, where  $n$  is a total number of sites. A detailed description of the Levenshtein approach in dialectometry can be found in Heeringa (2004). The obtained distance matrix is further analyzed using a weighted average clustering algorithm and multidimensional scaling. Both techniques have confirmed that the main dialect split is between High and Low German dialects, splitting the country in two main dialect areas. Using a clustering technique, the authors have detected a further split into five dialect groups, which to some extent correspond to the traditional dialects: Low German, East Central German and Upper German and a heterogeneous area in the west that is divided into two smaller groups. (Figure 2 (left)). Using multidimensional scaling, the authors have detected, next to a two-way split into High and Low German dialects, also Upper German, East Central German, East Low German, and West Low German areas (Figure 2 (right)). Nerbonne & Siedle (2005) have shown that quantitative approaches can successfully be applied on traditional dialect data to detect main dialect areas.

## 4 Regular sound correspondences

In order to examine diachronic processes responsible for the observed dialect divisions in Germany, in this paper geographic distribution of regular sound correspondences is examined with the help of a network model. Detection of regular sound correspondences between two or more languages is the essential part of the comparative method that aims to postulate regular sound changes, determine if examined languages are genetically related and, if so, reconstruct the proto-language. In traditional historical linguistics, extraction of regular sound correspondences is done manually by comparing lists of potential cognate words. In quantitative historical linguistics and dialectometry, this process is automatized by using algorithms to align cognate words and extract sound correspondences from the alignments (Prokić 2007; List 2012a). In order to estimate the strength of the association (i.e., if a given correspondence is regular or irregular) between any two aligned sounds, Wieling, Prokić

<sup>3</sup> In this paper the IPA version of the data was used.

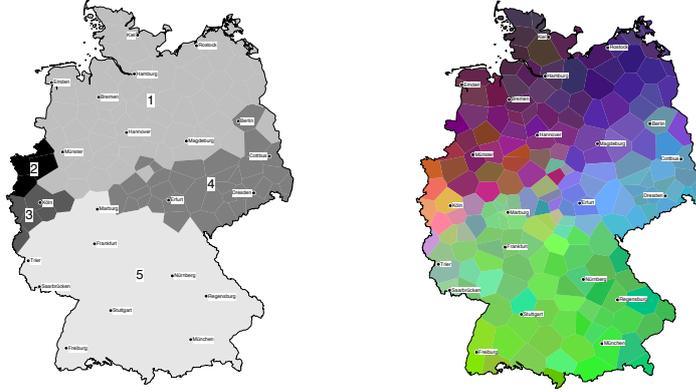


Figure 2: Results of the aggregate analysis of German dialects by Nerbonne & Siedle (2005) using clustering (left) and multidimensional scaling (right).

& Nerbonne (2009) rely on pointwise mutual information (PMI) and show that introduction of this association measure in the alignment procedure leads to improved alignments. Prokić & Cysouw (2013) employ an association measure based on Poisson distribution to detect regular sound correspondences, since this measure does not assume a normal distribution of the data and at the same time gives better estimates for the less frequent sounds. In this paper, I follow Prokić & Cysouw (2013) and use the Poisson association measure to estimate regularity of sound correspondences in the PAD data:

$$\text{Poisson Association} = \text{sign}(O - E) * \left( O * \log \left( \frac{O}{E} \right) - (O - E) \right) \quad (1)$$

where  $O$  is the observed co-occurrence and  $E$  is the expected co-occurrence of two sounds. The observed co-occurrence represents the number of times two sounds are found aligned in the data, while the expected co-occurrence is the number of times two sounds are expected to be aligned based on their frequency in the data. The expected frequencies  $E$  are calculated using the following formula:

$$E_{xy} = \frac{n_x \cdot n_y}{N} \quad (2)$$

where  $n_x$  represents the number of times sound  $x$  is aligned with any other sound,  $n_y$  represents the number of times sound  $y$  is aligned with any other sound, and  $N$  is the total number of aligned phone pairs in the data.

## 5 Analysis

Analysis of German dialects presented in this paper relies on a network model shown in Figure 1 to investigate geographic distribution of regular sound correspondences and their association strength. Employment of this network model enables us to compare only geographically neighboring sites and examine, for each sound correspondence individually, if it is regular or not. In traditional dialectology, areas where sound change is regular are considered to be areas where sound change originates. Less regular sound changes are characteristic for the so-called *transition areas*. The method presented in this paper detects areas with regular and irregular sound correspondences from synchronic dialect data, which allows us to infer geographic spread of sound changes. The analysis of German dialect data was performed according to the following steps.

Step 1: phonetic transcriptions were automatically multiple aligned using `lingpy` Python library (List & Moran 2013). Unlike in pairwise alignment employed by Nerbonne & Siedle (2005), in multiple alignment all strings are aligned and compared at the same time. Multiple string alignment is a good technique for discovering patterns in the aligned strings and the advantages of multiple over pairwise string alignment have relatively recently started being recognized in linguistics (Prokić, Wieling & Nerbonne 2009; List 2012b).

Step 2: all sites in the data were plotted on the map and connected to the neighboring sites, forming a network presented in Figure 1.

Step 3: for each pair of neighboring sites, all pairs of corresponding sounds were extracted from the alignments. The association strength (i.e., regularity) for all pairs of sounds was calculated using the Poisson association measure (Equation 1).

Step 4: for each site, the number of irregular correspondences between that site and all neighboring sites was calculated by counting all sound correspondences which association value is under a certain threshold. Following Prokić & Cysouw (2013), all correspondences which association value is smaller than 5 are taken to be irregular. It is a very conservative cut-off, which ensures that only highly irregular correspondences are counted.

Step 5: the number of irregular correspondences for each site was plotted on a map of Germany and analyzed using the Inverse Distance Weighting (IDW) interpolation method as implemented in the Quantum GIS software (QGIS Development Team 2012). Interpolation methods are used in spatial analyses to predict unknown values for any geographic point data based on a limited number of sample data points. They are used for the analysis of continuous spatial phenomena. The Inverse Distance Weighted method estimates unknown values by averaging the values of sample data points in the neighborhood of each processing cell.

## 5.1 Results

The results of the analysis can be seen on the map in Figure 3. Areas with a high number of irregular sound correspondences are colored red, while the areas with a small number of irregular sound correspondences are represented with a dark blue color. The map shows that areas in the north-west and south-west, colored dark blue, are the areas where many sound correspondences show a high degree of regularity. Two areas in the central part of the country and one in the north-east show a high number of irregular sound correspondences, which suggests that these areas are the so-called *transition areas*. The map in Figure 3 shows a historical split of dialects into northern and southern varieties, which corresponds well with the traditional scholarship that divides German dialect continuum into Low German found in the north and High German found in the south of the country. This split was also detected by method used in Nerbonne & Siedle (2005) that focuses on the synchronic dialect divisions.

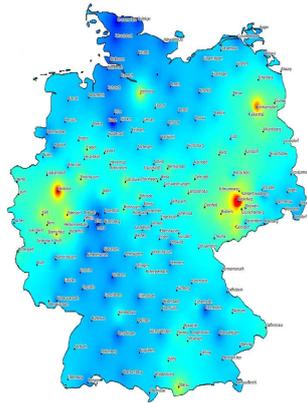


Figure 3: Map of Germany showing two areas with a high number of regular sound correspondences: in the north-west and south-west. The area in the middle of the country shows a high number of irregular sound correspondences.

## 6 Discussion

The method presented in this paper is the first step toward quantitative diachronic dialectology, that tries to use quantitative and statistical methods in order to examine diachronic processes that are responsible for the observed dialect divisions. The method relies on the network representation of the sites, combined with automatically extracted sound correspondences and their association strength. By analyzing regularity of all sound correspondences in the data set, it is possible to detect historically important areas when it comes to the spread of sound changes. In this paper,

all sound correspondences found in the data are analyzed, resulting in an aggregate diachronic map. Applying this method on a specific sound change would give a much sharper picture of the geographic origin of a sound change in question and its geographic spread. This type of analysis is beyond the scope of this paper and remains future work.

## References

- Bloomfield, Leonard. 1973. *Language*. London: Allen & Unwin.
- Bolognesi, Roberto & Wilbert Heeringa. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap* 9(1). 45–84.
- Campbell, Lyle. 2004. *Historical linguistics*. 2nd edn. Edinburgh: Edinburgh University Press.
- Chambers, J. K. & Peter Trudgill. 2004. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.
- Gooskens, Charlotte & Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16. 189–207.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. University of Groningen PhD dissertation.
- Hock, Hans Henrich. 1991. *Principles of historical linguistics*. 2nd edn. Berlin: Mouton de Gruyter.
- List, Johann-Mattis. 2012a. LexStat: automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125.
- List, Johann-Mattis. 2012b. Multiple sequence alignment in historical linguistics. In Enrico Boone, Kathrin Linke & Maartje Schulpen (eds.), *Proceedings of ConSOLE XIX*, 241–260.
- List, Johann-Mattis & Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the ACL 2013 System Demonstrations*, 13–18.
- Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van de Kooi, Simone Otten & Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans & S. Gills (eds.), *CLIN VI, papers from the sixth CLIN meeting*, 185–202. University of Antwerpen, Antwerpen.
- Nerbonne, John & Christine Siedle. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 2(72). 129–147.
- Prokić, Jelena. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, 61–66. Prague, Czech Republic: Association for Computational Linguistics.
- Prokić, Jelena, Çağrı Çöltekin & John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 72–80. Avignon: Association for Computational Linguistics.

- Prokić, Jelena & Michael Cysouw. 2013. Combining regular sound correspondences and geographic spread. *Language Dynamics and Change* 3.2. Special issue, “Phylogeny and Beyond: Quantitative Diachronic Approaches to Language Diversity”, edited by Michael Dunn.
- Prokić, Jelena, Martijn Wieling & John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELTER 2009)*, 18–25. Association for Computational Linguistics.
- QGIS Development Team. 2012. *QGIS Geographic Information System*. Open Source Geospatial Foundation. <http://qgis.osgeo.org>.
- Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25. 700–715.
- Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS One* 6(9).
- Wieling, Martijn, Jelena Prokić & John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELTER 2009)*, 26–34. Association for Computational Linguistics.



## Chapter 31

# Write as you speak? A cross-linguistic investigation of orthographic transparency in 16 Germanic, Romance and Slavic languages

Anja Schüppert

University of Groningen

Wilbert Heeringa

Fryske Akademy

Jelena Golubovic

University of Groningen

Charlotte Gooskens

University of Groningen

### 1 Introduction

Ever since the invention of the printing press, the creation of a written standard represented an important cornerstone in the development of most languages. For many language varieties, this meant taking the step from a dialect or regiolect to a language, and for most national languages, the presence of a written standard was a necessary prerequisite. The orthographies of most European languages were developed from a set of unstandardized conventions which usually served as the basis of the new norm. For some languages the spelling norms were established with the first official translation of the Bible (e.g. Czech), while others set the norm through publishing a dictionary and describing the new spelling used (Dutch, French, Spanish). The spelling was then updated through a series of reforms, most of them passed

in the 18th and 19th centuries. These reforms often followed significant historical changes and were seen as a vital part of language standardisation, which at the time was often an important element in nation building.

Naturally, since the speed at which the orthographic reforms follow the changes in speech varies substantially per language, this generally means that the extent to which ancient pronunciation is preserved in the current spelling also varies. Sometimes spelling is also deliberately kept unchanged in order to illustrate word etymology. This variety in decisions and the very fact that spelling rules are arbitrary, in turn leads to cross-linguistic differences in how accurately contemporary pronunciation is reflected in the orthography.

All official EU languages are written in alphabetic systems, which use roughly one character (or grapheme) for one sound (or phoneme). However, the correspondence of one character to one sound is not always a strict rule. This is partly because pronunciation has changed since the establishment of a written standard and the spelling has not been (fully) reformed to reflect these changes. Another reason is the fact that in many alphabetical orthographies, the grapheme repertoire is restricted to some 20-40 graphemes. The sound repertoires, however, are usually somewhat larger, which means that not every grapheme can be assigned to exactly one phoneme.

More specifically, the same grapheme can represent different phonemes, such as the letter <o> that can represent /ɔ/ as in <not>, or /oʊ/ as in <go>, or even /ɪ/ as in <women> in English. Analogically, the same phoneme can be represented by different graphemes, such as the consonant /i:/ that can be spelled <e> as in <here>, or <ea> as in <beat>, or <ee> as in <beet>. Similar examples can be found in other European languages, such as the pronunciation of the Dutch letter <a> as /ɑ/ in <pad>, but as /a:/ as in <paden>, or the spelling of the Swedish consonant /ɸ/ as <sk> in the word <skär>, or as <skj> in the word <skjorta>, or as <stj> in the word <stjärna>, or as <g> in the word <giraf>.

Not only can the same phoneme be represented by different graphemes and vice versa, but in many European orthographies, a single phoneme might also be represented by a grapheme *cluster* (such as the clusters <ph> or <ea> in English e.g. <phonetic measurement> /fə'netɪk məʒəmənt/). To be able to read the written word <phonetic> correctly, the reader needs to be familiar with the rule that the grapheme cluster <ph> is not pronounced as the sum of its parts /ph/ or /p<sup>h</sup>/, but as a separate phoneme /f/. The same is true for the cluster <ea> in <measurement>, which is generally transcribed as /e/ (British English) or /ɛ/ (American English). There are numerous other examples from different European languages, such as the grapheme cluster <sch> in German, which is pronounced /ʃ/ in <syntaktisch> /zʏn'taktɪʃ/ or the clusters <gn> and <ent> in French <alignement> /aliɲmɑ̃/ which are pronounced with the single phonemes /ɲ/ and /ɑ̃/, or the spelling <Groningen> with the grapheme cluster <ng>, for the pronunciation with a single phoneme /ŋ/. In other words, a reader who is aware of the fact that a single letter can be pronounced in different ways, but not of these additional rules for the pronunciation of grapheme clusters, will have even more difficulties to read written language properly – although arguably, the degree to which these missing rules create problems in the reading process vary a

lot across different languages and their orthographies. And again, in a similar way, phoneme clusters can be spelled with a single grapheme, such as the affricate /dʒ/ in the English name /dʒɒn/ (Am. /dʒɑn/) which is spelled <John>, or the phoneme cluster /ks/ in /d̥arələkt p̥rɔksɪməti/ which is spelled <dialect proximity>. These intransparencies create huge problems for beginning readers and writers, be it (usually young) native speakers or (young as well as older) second-language learners.

And as if the sheer presence of rules for grapheme (and phoneme) clusters was not enough to confuse the poor reader (and writer), in some orthographies similar grapheme clusters have different etymological backgrounds. This means that they follow different pronunciation rules and thus are not clearly decipherable, even if a reader is aware of the fact that grapheme clusters ought to be treated differently than single graphemes. One of many examples for this is the English grapheme cluster <gh> which can be pronounced as /f/ (as in <laugh> /lɑ:f/), or as a zero-ending as in <dough> /dɒʊ/. To be able to read words such as <laugh> and <dough> correctly, the reader needs to apply not only the knowledge that <gh> is pronounced differently than the sum of its parts, but also be familiar with the etymology of the word in order to choose the correct pronunciation (alternatively, have learned proper pronunciation for every specific word separately, which might be the most frequent strategy). Analogically, again, if every /f/ sound in English would be spelled /gh/, this would restrict the opaqueness and hence the difficulties to the process of reading. But what makes things even worse is the fact that the phoneme /f/ can also be spelled <f> or <ff>, and that the phoneme cluster /oʊ/ also can be spelled <o> (as in <go>), <ow> (as in <low>), or <oe> (as in <toe>). This extends the scope of intransparency from reading (due to an opaque grapheme-phoneme correspondence) to writing (due to an opaque phoneme-grapheme correspondence).

If you are able to read this paper until here, you are most likely at least partly familiar with English spelling. You might even have come across the notion of English orthography belonging to the orthographies that are particularly intransparent – an assumption which strongly contradicts the statement made by Chomsky & Halle (1968: 49) who claimed that “[t]here is [...] nothing particularly surprising about the fact that conventional orthography is [...] a near optimal system for the lexical representation of English words”. There is not only common-sense and speculative evidence that Chomsky & Halle (1968) were wrong, but also growing scientific support for the objection against their claim. Borgwaldt, Hellwig & de Groot (2005) conducted entropy (= uncertainty) measurements for letter-to-phoneme mappings in Dutch, English, French, German, Hungarian, Italian and Portuguese. They reported that English had the highest entropy value of all included languages, which means that the pronunciation predictability for English letters is lower than for any of the other six languages. Their algorithm models an advanced reader, in that it analyses the consistency of spelling and pronunciation of rimes and onsets, and not of single graphemes and phonemes. In particular, some basic knowledge of consonant and grapheme clusters is presumed, such as the English rule that <gh> is only pronounced /f/ word-medially and word-finally, but never word-initially. In contrast to Borgwaldt et al.’s (2005) bottom-up approach, Nicolai & Kondrak (2015) used a top-

down approach to scrutinise Chomsky & Halle's (1968) claim. Instead of quantifying the uncertainty of different existing orthographies, they developed an algorithm modelling a 'better fit' for the spelling of English pronunciation, i.e. a more transparent orthography for English. The two investigations have two things in common: The conclusion that English orthography is far from being optimal, and the fact that they restrict themselves to phoneme-to-grapheme correspondences (quantifying spelling problems in English), but do not investigate grapheme-to-phoneme correspondences (which would quantify reading problems).

It becomes evident that English orthography is situated close to one end of the continuum between transparent (or *shallow*) and intransparent (or *deep*) orthographies. Some of the European orthographies that are said to belong to the other end of this continuum are Finnish and several Slavic languages, in which most graphemes are pronounced in only one way, and most phonemes are spelled in only one way – at least in careful speech. According to the *Orthographic Depth Hypothesis* (ODH) (Katz & Frost 1992), decoding a deep orthography requires a different way of reading than decoding a shallow one. The ODH suggests that when reading a shallow orthography, the reader can focus on phonological (non-lexical) information, while the reader has to focus on larger units (lexical information) when reading a deep orthography. In line with the ODH, the *Psycholinguistic Grain Size Theory* (PGST), put forward by Ziegler & Goswami (2005), postulates that reading shallow orthographies allows the reader to process smaller units (*grain sizes*) than reading deep orthographies, since the predictability of phoneme-grapheme correspondences in deep orthographies increases if the grapheme or phoneme context is taken into account. In other words, readers and writers of deep orthographies are more likely to use logographic entities than learners of shallow orthographies are, at least when they have reached a certain level of literacy.

Seymour, Aro & Erskine (2003) were among the first to conduct a cross-linguistic study on literacy acquisition on a broad range of European languages. They use *syllable complexity* and *orthographic depth* as two independent predicting factors. Although the outcomes of such a study have to be interpreted with caution (as the orthographic system naturally is not the only factor that differs between 1st-graders in Scotland, Iceland, or Greece), their study showed that children acquiring an orthography that has been described as relatively deep (such as English) are learning to read twice as slowly as children acquiring an orthography that is traditionally seen as shallow (such as Finnish). However, the classification into 'deep' or 'shallow' (on a five-pointscale) is a "hypothetical classification" (Seymour, Aro & Erskine 2003: 146). It is not clear what the basis of the placement of every orthography on the continuum between 'deep' and 'shallow' in their study is based on. Also earlier cross-linguistic studies investigating the effect of orthographic depth on reading development lay a cloak of silence on the question how orthographic depth was determined (cf. Wimmer & Goswami 1994; Frith, Wimmer & Landerl 1998; Goswami, Porpodas & Wheelwright 1997; Goswami 2008; Ziegler & Goswami 2005; 2006). However, in a study comparing children's reading speed and accuracy in three languages (English, French and Spanish), Goswami, Gombert & de Barrera (1998) refer to inves-

tigations on the consistency of English spelling conducted by Treiman et al. (1995), as well as a similar investigation by Peereman & Content (1996) on French orthographic consistency. No measurements are presented for Spanish, however. Ziegler, Stone & Jacobs (1997) present a database of phoneme-to-grapheme and grapheme-to-phoneme consistency for 2,694 English monosyllabic words and report that 72.3% of the monosyllabic words could theoretically be spelt in more than one way, and that 30.7% of the words could be pronounced in more than one way. In a recent paper, Ziegler et al. (2010) conduct a thorough investigation of reading skills in Finnish-, Hungarian-, Dutch-, Portuguese-, and French-speaking children, and discuss their findings in the light of the mean orthographic depth of the word onsets in every language as established by Borgwaldt, Hellwig & de Groot (2005).

Crucially, however, most studies that compared cross-linguistic literacy acquisition in a large number of languages seem to categorise the involved orthographies on the basis of ‘common sense’ or speculation. In the present paper the orthographies of 16 European languages are compared using a uniform methodology applied to the same set of 100 words. Importantly, the method models a completely illiterate reader, as the entity of our analysis are phonemes and graphemes. In other words, any rules that are reflected in larger entities of written language, such as clusters or rimes, are treated as the sum of their parts. By providing entropy values for both grapheme-to-phoneme and phoneme-to-grapheme correspondences, the results can be used as a basis for the prediction of writing as well as reading development in very early beginners, namely practically illiterate beginners who are only familiar with the ‘names’ of the letters in their alphabet.

## 2 Method

Entropy measures the uncertainty in a random variable. In this study, the uncertainty between phoneme-to-grapheme mappings and grapheme-to-phoneme mappings are at focus: a letter may correspond to one or several phonemes. If the letter can correspond to more than one phoneme, then for a beginning reader, who knows nothing more than the fact that s/he is confronted with an alphabetical orthography as well as the names of the letters of the specific alphabet, there exists uncertainty about which phoneme corresponds to the specific letter in the word that s/he is reading. Reversely, when the same person is listening to a language in which the same phoneme in different words is transcribed by the different letters, there will be uncertainty about which letter represents the sound s/he hears when writing down the word.

We quantify uncertainty as Shannon’s entropy (Shannon 1948). Given a grapheme  $x$ , and given variable  $Y$  being a random variable with  $m$  possible pronunciations  $y_i$  with probabilities  $p(y_i)$  for grapheme  $x$ , then the entropy, i.e. the uncertainty about which pronunciation  $y_i$  will correspond with  $x$  is:

$$H(Y) = - \sum_{i=1}^m p(y_i) \log_2 p(y_i) \quad (1)$$

Given a phoneme  $y$ , and given variable  $X$  being a random variable with  $m$  possible spellings  $x_i$  with probabilities  $p(x_i)$  for grapheme  $y$ , then the entropy, i.e. the uncertainty about which spelling  $x_i$  will correspond with  $y$  is:

$$H(X) = - \sum_{i=1}^m p(x_i) \log_2 p(x_i) \quad (2)$$

An entropy of 0 represents a fully predictable correspondence. The larger the entropy, the less predictable the correspondences.

The average grapheme-to-phoneme entropy for a language  $L$  with a grapheme inventory consisting of  $v$  different graphemes is calculated as the average of  $v$  entropy values of the  $v$  individual graphemes. Similarly, we calculate the average phoneme-to-grapheme entropy for a language  $L$  with a phoneme inventory of  $w$  different phonemes as the average of  $w$  entropy values corresponding with the  $w$  individual phonemes.

In order to measure the average grapheme-to-phoneme entropy and the average phoneme-to-grapheme entropy for each language, an R script was developed (R Core Team 2016).

### 3 Corpora and alignment

Due to practical reasons, we decided to include only the 16 Germanic, Romance and Slavic languages that are official EU-languages in our study. The languages in question are: Bulgarian, Croatian, Czech, Danish, Dutch, English, French, German, Italian, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish. Unfortunately, this meant excluding one of the languages that is said to have one of the most transparent orthographies, namely Finnish.

#### 3.1 Graphemic corpus

The graphemic corpus, providing us with the official spelling in all 16 languages, consisted of 100 words per language. The words were among the 109 most frequent nouns from the British National Corpus (BNC; Leech, Rayson & Wilson 2001). We excluded words for semantically too similar concepts for our goal, such as the nouns *sort* and *kind*. The remaining 100 nouns were translated into 16 official EU-languages. To make sure that the translators had the same concept in mind, the English source words were presented to them embedded in a sentence. For example, the English noun *state* could be translated into Dutch as ‘toestand’, ‘land’, ‘staat’, or ‘deelstaat’. The provided context was supposed to reduce the ambiguity of the concept behind the source word. All translators were native speakers of the target language. Every list of 100 translations was checked by a second native speaker and, if necessary, corrected. In a few cases of doubt, a third native speaker was consulted.

### 3.2 Phonemic corpus

The phonemic corpus, providing us with the most standard pronunciation in all 16 languages, consisted of the very same 100 words per language. For every language, the word list was transcribed phonemically using X-SAMPA (Wells 1995) using pronunciation dictionaries.

### 3.3 Alignments

In a first step, the graphemic and the phonemic transcriptions were aligned word by word in two different ways: (1) As phoneme-to-grapheme alignment, and (2) as grapheme-to-phoneme alignment. These two different alignment tables for all 16 languages serve as the input data for the two entropy measurements, i.e. writing entropy (1) and reading entropy (2).

For (1), three rules were formulated: rule (i) demanded that every grapheme ought to be put in a separate cell (modelling a reader who is unaware of potential rules for grapheme clusters such as for English <ph>); rule (ii) stated that phonemes should be aligned in such a way that vowel graphemes represent vowel phonemes and consonant graphemes represent consonant phonemes (modelling the usage of knowledge of the names of the letters by a reader).

Combined, these two rules meant that some graphemes could not be aligned to a phoneme (e.g. the word-final <e> in the English word <time> which is aligned to /taɪm/). We presume that the reader who is completely unaware of any phonological rules of the rime <ime> being pronounced VVC rather than VCV is confused by this unpronounced final letter and might try to map it to the preceding consonant.

Therefore, rule (iii) was formulated: no empty cells were allowed in either of the columns, but in cases where no specific phoneme could be aligned to a vowel grapheme, the preceding or following vowel was extended, and when no specific phoneme could be aligned to a consonant grapheme, the preceding or following consonant was 'prolonged'. This decision aligns 'silent' letters to many different phonemes, and thereby results in a higher entropy value for languages with many silent letters. Although we think that this captures the opaqueness of such orthography well, this is arguably not the only way to model the opaqueness of silent letters. Similarly to the problem of unalignable phonemes, there were instances when two phonemes had to be aligned to one single grapheme (recall rule (i) that every grapheme had to be put in a separate cell in the phoneme-to-grapheme alignments). An example of this is the letter <i> in the English word <time> which is aligned to /aɪ/.

In a very similar way, the phoneme-to-grapheme alignments were conducted. Again, rule (i) required that every phoneme was put in a separate cell, rule (ii) demanded that vowel phonemes were aligned to vowel graphemes and consonant phonemes to consonant graphemes, and rule (iii) stated that empty cells were allowed in none of the two columns. Table 1 shows an example of the alignments for English.

Table 1: Grapheme-to-phoneme alignments (left), which formed the basis of the reading entropy, and phoneme-to-grapheme alignments (right), which formed the basis of the writing entropy in English.

Reading uncertainty		Writing uncertainty	
grapheme	phoneme	phoneme	grapheme
t	t	t	t
i	aɪ	a	i
m	m	ɪ	i
e	m	m	me

## 4 Results

The results from the entropy measurements are given in Figure 1. It becomes obvious that English is the language with the least predictable orthography from a beginners' point of view, both in writing and reading. French has a very opaque orthography for a beginning reader, and it is still rather hard to spell properly if nothing more is known than the names of the letters. Regarding the uncertainty of spelling, German and Danish share a third place with fairly opaque phoneme-to-grapheme representations, and Danish grapheme-to-phoneme representations are even more opaque than the French. On the other end of the continuum the differences in transparency are less pronounced. Apart from Swedish, only Romance and Slavic languages have entropy values of less than 0.5. Among the 16 included languages, Bulgarian is the language that is easiest to read and Croatian has the orthography which is easiest to write.

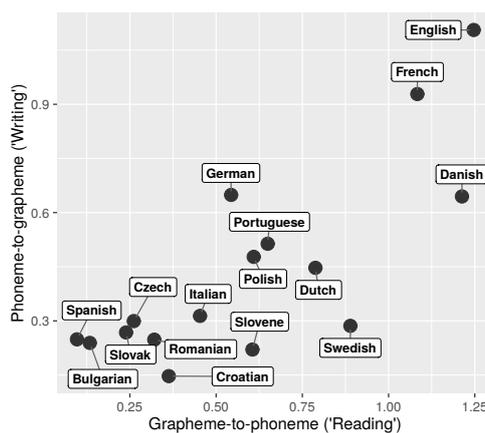


Figure 1: Grapheme-to-phoneme ('reading') entropy values plotted against phoneme-to-grapheme ('writing') entropy values for all 16 languages.

## 5 Conclusion

Using an R-script that calculates the entropy (uncertainty) of grapheme-to-phoneme and phoneme-to-grapheme correspondences, we modelled a beginning reader and writer, who is familiar with nothing more than the letter names in every orthography. For this type of illiterate learner, English and French are by far the most opaque orthographies to read, and English and Danish are the most opaque orthographies to write, while most Slavic and Romance languages are far less opaque.

The main reason why Slavic languages as a group are characterized by relatively predictable grapheme-to-phoneme and phoneme-to-grapheme correspondences are the orthographic reforms mostly carried out in the 19th century. Czech was the first language to standardize its orthographic system and the orthographies of Slovak, Slovene and Serbo-Croatian used it as a model when creating their own orthographies. The rule attributed to German philologist Adelung “write as you speak and read as it is written” was often explicitly invoked by Slavic language experts carrying out the reforms. Therefore, the current orthographic consistency in Czech, Slovak, Slovene and Croatian is a result of a conscious effort to ensure transparency. Bulgarian uses the Cyrillic alphabet and while the written language looks completely different compared to the other 15 languages we studied, it boasts the lowest entropy values. One notable exception in terms of transparency of Slavic languages is Polish. Ever since the Latin alphabet was adopted in Polish, it was clear that it could not accommodate palatal and retroflex consonants as well as the nasal vowels, typical of this language. Subsequent reforms did improve things to an extent, but due to numerous complex and often inconsistent rules, Polish orthography remains the least transparent one among the Slavic languages.

Interesting is also the very different degree of transparency of the Scandinavian orthographies that were included in our study, i.e. the orthographies of the two very closely related languages Danish and Swedish. The extremely differing entropy values can be explained by two processes, as summarised by Elbro (2005: 33): firstly, “Danish orthography was already old when a national norm was first established around the year 1200”, and secondly, “things have become worse since the 1200s” as “spoken Danish has changed more than most Germanic languages since the 1200s”. In other words, while the Danish spelling norm always has preserved a rather archaic pronunciation, the spoken language has developed even faster than in many other languages and the spelling has not been adjusted to these changes. In Danish, many sounds that have been lost in pronunciation are still preserved as silent letters, such as the spelling <mild> for /milʔ/, or the spelling of <tolv> for /tɔlʔ/, or the spelling <lærere> for /lɛ:ɔɔ/. In comparison, these words are spelled <mild>, <tolv> and <lärare> in Swedish, and pronounced /mild/, /tɔlv/ och /lærarə/. Also loan words have different appearances in the two languages: Danish has more foreign words than Swedish, as many loan words into Swedish are translated (rendering *calques*). An example for this is the Danish word *weekend* versus the Swedish calque *veckoslut* or the Old Norse word *helg* (from *helig*, Engl. ‘holy’). Furthermore, even for directly loaned words, Danish has a tendency to preserve the foreign spelling such as in Danish <niveau> versus Swedish <nivå> or Danish <restaurant> versus Swedish

<restaurang>.

Let us now return to the study conducted by Seymour, Aro & Erskine (2003), who measured literacy acquisition in 13 orthographies and interpreted their results on the basis of a “hypothetical classification of participating languages” (Seymour, Aro & Erskine 2003: 146). They classified the 13 orthographies on a five-point scale from *shallow* to *deep*. Importantly, they did not make a difference between grapheme-to-phoneme depth and phoneme-to-grapheme depth, as we did. Recalling that they elicited data from reading development only, we might assume that their classification is meant mainly or exclusively for *reading* opaqueness, and should be compared to our grapheme-to-phoneme measurements in the first place. Although their classification is very rough (with e.g. Italian, Spanish and German in one (nameless) category representing *semi-shallow*, and Portuguese, Dutch and Swedish in one (equally nameless) category between shallow and deep), their ranking is completely supported by our results.

This study represents a commensurate comparison of phoneme-to-grapheme correspondence and vice versa in 16 European languages. The findings can be taken as a predictor of reading and spelling difficulties in each of the languages, but can also serve as background information for psycholinguistics experiments. The fact that we modelled a beginning reader and writer with hardly any orthographical knowledge limits the validity of this study. We opted for this setup since it is vital for a cross-linguistic study to take the same criteria as a basis, and the criteria we used were easily applicable to the 16 languages. Modelling a slightly more advanced reader and writer, as Borgwaldt, Hellwig & de Groot (2005) did for Dutch, English, French, German, Hungarian, Italian and Portuguese, is a very useful additional approach. Another useful extension of the present study would be the inclusion of more languages.

## References

- Borgwaldt, S. R., F. Hellwig & A. M. B. de Groot. 2005. Onset entropy matters – letter-to-phoneme mappings in seven languages. *Reading and Writing* 18. 211–229.
- Chomsky, N. & M. Halle. 1968. *The sound patterns of English*. New York: Harper & Row.
- Elbro, C. 2005. Literacy acquisition in Danish: a deep orthography in cross-linguistic light. In R. Malatesha Joshi & P. G. Aaron (eds.), *Handbook of orthography and literacy*, 31–45.
- Frith, U., H. Wimmer & K. Landerl. 1998. Differences in phonological recoding in German- and English-speaking children. *Journal of the Society for the Scientific Study of Reading* 2. 31–54.
- Goswami, U. 2008. The development of reading across languages. *Annals of the New York Academy of Sciences* 1145. 1–12.
- Goswami, U., J. E. Gombert & L. F. de Barrera. 1998. Children’s orthographic representations and linguistic transparency: nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics* 19. 19–52.

- Goswami, U., C. Porpodas & S. Wheelwright. 1997. Children's orthographic representations in English and Greek. *European Journal of Psychology of Education* 12. 273–292.
- Katz, L. & R. Frost. 1992. The reading process is different for different orthographies: the orthographic depth hypothesis. In R. Frost & L. Katz (eds.), *Orthography, phonology, morphology, and meaning* (Advances in psychology 94), 67–84. Oxford: North-Holland.
- Leech, G., P. Rayson & A. Wilson. 2001. *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.
- Nicolai, G. & G. Kondrak. 2015. *English Spelling is not "close to optimal"*. Paper presented at the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), Denver, CO.
- Peereman, R. & A. Content. 1996. Orthographic and phonological neighbourhoods in naming: not all neighbors are as mighty in orthographic space. *Journal of Memory and Language* 37. 382–410.
- R Core Team. 2016. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Seymour, P. H. K., M. Aro & J. M. Erskine. 2003. Foundation literacy acquisition in European orthographies. *British Journal of Psychology* 94. 143–174.
- Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27 (3). 379–423.
- Treiman, R., J. Mullennix, R. Bijeljac-Babic & E. D. Richmond-Welty. 1995. The special role of rimes in the description, use and acquisition of English orthography. *Journal of Experimental Psychology: General* 127. 107–136.
- Wells, J. C. 1995. *Computer-coding the IPA: a proposed extension of SAMPA*. UCL Phonetics & Linguistics.
- Wimmer, H. & U. Goswami. 1994. The influence of orthographic consistency on reading development: word recognition in English and German children. *Cognition* 51. 91–103.
- Ziegler, J. C., D. Bertrand, D. Tóth, V. Csépe, A. Reis, L. Faisca, N. Saine, H. Lyytinen, A. Vaessen & L. Blomert. 2010. Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. *Psychological Science* 21. 551–559.
- Ziegler, J. C. & U. Goswami. 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin* 131. 3–29.
- Ziegler, J. C. & U. Goswami. 2006. Becoming literate in different languages: similar problems, different solutions. *Developmental Science* 9. 429–453.
- Ziegler, J. C., G. O. Stone & A. M. Jacobs. 1997. What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behaviour Research Methods, Instruments and Computers* 29. 600–618.



## Chapter 32

# Vowel co-occurrence restriction in Ainu

Hidetoshi Shiraishi

Sapporo Gakuin University

### 1 Introduction

Ainu (language isolate, Japan and Russia) exhibits a co-occurrence restriction on vowels in verbal and nominal derivations. This restriction was first reported in Chiri (1952) as a type of vowel harmony. Later, it was taken up by Ito (1984) and became one of the most discussed topics in Ainu phonology to date (e.g. Mester 1986; Ewen & van der Hulst 1988; Shibatani 1990; Krämer 1998; Sato 2010).

In this paper I will give an outline of the phenomenon and review the discussions made by these authors. While Ito (1984) assumes the co-occurrence restriction as being both assimilatory and dissimilatory (disharmonic) in nature, I will point out that there is a flaw in Ito's description of the phenomenon. This flaw invalidates most of her analysis and those defended in the above literature, which depend on her description. Instead, I will propose an alternative view in which the assimilatory nature of the co-occurrence restriction plays a central role, as originally suggested by Chiri in his first report of the phenomenon in 1952.

### 2 Co-occurrence restriction between vowels in disyllabic derivations

The description of the co-occurrence restriction given below is based on Chiri (1952). Ainu exhibits V1–V2 co-occurrence restriction in two derivational contexts: transitive verb formation and possessive noun formation. In transitive verb formation, the base to which a V-suffix attaches is either an intransitive verb (free morpheme) or a root (bound morpheme) (the distinction between bound and free morphemes is irrelevant to the discussion here).

	Intransitive verb or root	Transitive verb
a.	<i>kay</i> 'to be broken'	<i>kay-e</i> 'to break'
b.	<i>mos</i> 'be awake'	<i>mos-o</i> 'to wake up'
c.	<i>yak</i> 'to be crushed'	<i>yak-u</i> 'to crush'
(1) d.	<i>ran</i> 'to go down'	<i>ran-i</i> 'to lower'
e.	$\sqrt{\text{mak}}$ 'open'	<i>mak-a</i> 'to open'
f.	$\sqrt{\text{kom}}$ 'bent'	<i>kom-o</i> 'to bend'
g.	$\sqrt{\text{kar}}$ 'spinning'	<i>kar-i</i> 'to rotate'
h.	$\sqrt{\text{mes}}$ 'to come off'	<i>mes-u</i> 'to tear off'

In possessive noun formation, the V-suffix attaches to a base which is often referred to as the *conceptual form* in Ainu literature (Kindaichi & Chiri 1936; Tamura 2000). Through V-suffix affixation, the conceptual form changes to a possessive form in which the possessor is overtly expressed.

	Conceptual form	Possessive form
a.	<i>sa</i> 'sister'	<i>sa-ha</i> <sup>1</sup> 'one's sister'
b.	<i>re</i> 'name'	<i>re-he</i> 'one's name'
c.	<i>nan</i> 'face'	<i>nan-u</i> 'one's face'
(2) d.	<i>tek</i> 'hand'	<i>tek-e</i> 'one's hand'
e.	<i>haw</i> 'voice'	<i>haw-e</i> 'one's voice'
f.	<i>rek</i> 'beard'	<i>rek-i</i> 'one's beard'
g.	<i>hon</i> 'belly'	<i>hon-i</i> 'one's belly'
h.	<i>yup</i> 'brother'	<i>yup-i</i> 'one's brother'

The distribution of V1 and V2 is illustrated in the tables below. Observed frequencies (= word frequency) are from Chiri (1952). Expected frequencies (in brackets) and ratios (Tables 2 and 4) are added by the current author.

Table 1: Observed and expected frequencies (transitive verb).

V1 \ V2	i	u	e	a	o	
i	9 (4.0)	3 (4.7)	3 (4.0)	0 (1.4)	0 (1.7)	16
u	5 (4.8)	7 (5.6)	6 (4.8)	1 (1.7)	0 (2.0)	19
e	0 (4.5)	9 (5.3)	9 (4.5)	0 (1.6)	0 (1.9)	18
a	9 (10.0)	14 (11.8)	8 (10.0)	9 (3.6)	0 (4.3)	40
o	4 (4.8)	0 (5.6)	3 (4.8)	0 (1.7)	12 (2.0)	19
	28	33	28	10	12	112

<sup>1</sup> The [h] before V-suffix is epenthetic.

Table 2: Observed/Expected ratios (transitive verb).

V1 \ V2	i	u	e	a	o
i	2.25	0.64	0.75	0.00	0.00
u	1.04	1.25	1.25	0.59	0.00
e	0.00	1.70	2.00	0.00	0.00
a	0.90	1.19	0.80	2.50	0.00
o	0.83	0.00	0.63	0.00	6.00

Table 3: Observed and expected frequencies (possessive noun).

V1 \ V2	i	u	e	a	o	
i	22 (14.1)	1 (4.1)	1 (3.4)	0 (1.2)	0 (1.2)	24
u	28 (22.3)	3 (6.6)	7 (5.4)	0 (1.9)	0 (1.9)	38
e	14 (17.0)	7 (5.0)	8 (4.1)	0 (1.4)	0 (1.4)	29
a	12 (24.0)	17 (7.1)	3 (5.8)	8 (2.0)	1 (2.0)	41
o	19 (17.6)	0 (5.2)	4 (4.3)	0 (1.5)	7 (1.5)	30
	95	28	23	8	8	162

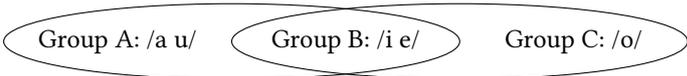
Table 4: Observed/Expected ratios (possessive noun).

V1 \ V2	i	u	e	a	o
i	1.56	0.24	0.29	0.00	0.00
u	1.26	0.45	1.30	0.00	0.00
e	0.82	1.40	1.95	0.00	0.00
a	0.50	2.39	0.52	4.00	0.50
o	1.08	0.00	0.93	0.00	4.67

### 3 Review of literature

#### 3.1 Chiri (1952)

Comparing Tables 1 and 3, Chiri (1952: 220) noticed that the distribution of V1–V2 is surprisingly similar in verbal and nominal derivations. Notably, he observed the following characteristics as being common to both nominal and verbal derivations. Firstly, /i u e/ outnumber /a o/ to a considerable degree in V2. Secondly, there is a preference for lining up identical vowels. This is especially the case with V2 /a/ and /o/, in which case V1 is nearly always /a/ and /o/, respectively. This preference is slightly weaker in V2 /e/, as non-identical V1–V2 sequences such as *u–e*, *a–e* or *o–e* are observed. In such cases, C2 is usually /y/ or /w/ (*ruy-e* ‘shake’, *haw-e* ‘one’s voice’, *poy-e* ‘to mix’), a regularity to which we return later. Finally, /a u/ and /o/ do not co-occur, whereas /i e/ can co-occur with any vowel. This final observation led Chiri to propose the following grouping of vowels.

- (3) 

In terms of vowel harmony, Group B is neutral since these vowels co-occur with any other vowel (Shibatani 1990: 14).

While (3) successfully accounts for the distribution of vowels illustrated in the tables above, it is not yet sufficient grounds to call the phenomenon vowel harmony, as Shibatani (1990) and Sato (2010) point out. Firstly, the classification in (3) cannot be characterized by a vowel feature with a phonetic basis known to trigger vowel harmony cross-linguistically, such as palatality (backness) or tongue-root position (Shibatani 1990: 15; Sato 2010: 164–165). Thus if this were a case of vowel harmony, it should have lost its transparency in the course of history. Nevertheless, such an assumption is missing in Chiri’s diachronic scenario, as we will see below. Secondly, this co-occurrence restriction applies to two morphological contexts only, namely transitive verb and possessive noun formation. It is not observed in other morphological contexts such as plural suffix (*kom-pa*, \**kom-po*) or personal prefix (*ku-komo*, \**ko-komo*) formation (Shibatani 1990: 15–16; Sato 2010: 163–166). There are no alternations which make vowel sequences in these morphological concatenations harmonic. This is to say that (3) is irrelevant to most morphological operations in Ainu.

Aside from the discussion of whether or not the restriction observed is vowel harmony, Chiri’s insightful observations hint at the phenomenon’s historical origin. He proposed the following diachronic scenario for possessive noun formation: at a previous historical stage, the V-suffix was /i/, a third person singular affix. Subsequently, this /i/ underwent a root-control type of harmony imposed by a preceding vowel. This led to the proliferation of V1–V2 forms with identical vowels.

According to Chiri, this scenario finds the following support: firstly, there is independent evidence for the use of /i/ as a third person singular affix. For instance in the Samani dialect, it surfaces as a prefix: *i-sapa* ‘his head’, *i-tanehe* ‘its seed’, *i-an* ‘it exists’, *i-pon* ‘it is small’. Secondly, Ainu exhibits cases of progressive assimilation, e.g.



Whatever Ito's contribution to the discussion of disharmony, her analysis cannot be readily accepted, as her description of the phenomenon has several flaws. Firstly, Ito declares at the outset that she will illustrate her case using data from transitive verb formation only, but this choice has seemingly led to the exclusion of possessive noun formation from consideration altogether. In fact, possessive nouns contain cases of *e-i*, which is predicted not to occur by MDR. Table 3 counts fourteen such cases, which is too many to eliminate as exceptions. This fact was either overlooked or neglected by Ito. It should therefore be pointed out that MDR is successful for only a small subset of verbs (thirteen in total). Secondly, Ito argues that there is independent support for MDR in the inventory of diphthongs. According to Ito (1984: 509–510), Ainu diphthongs consist of those which obey MDR (*iw*, *ew*, *uy*, *oy*, *aw*, *ay*) while those disobeying MDR do not exist (*ey*, *ow*). However, this observation is incorrect; Ainu has *ey* in underived forms: e.g. *sey* 'bivalve' or *teyne* 'to be wet'.

There are also shortcomings in Ito's proposed analysis. For instance, it is not clear how the three subcategorized groups interact with each other. Since Ito provides three different mechanisms to account for each group, her analysis gives the impression that there are three independent mechanisms involved (total assimilation, lexical encoding and MDR) although they apply to the same morphological context. Chiri's remarkable insight that both verbal and nominal derivations provide common grounds for V1–V2 distribution is thereby lost.

### 3.3 Sato (2010)

Sato (2010) faithfully follows Chiri's description of the phenomenon and focuses on the prevailing pattern of identical V1–V2.<sup>2</sup> Like Chiri, Sato (2010: 172) assumes it to occur as a consequence of root-control harmony, thus V1 affecting V2, and proposes to add the following step to Chiri's diachronic scenario. In a similar fashion to Chiri's reconstruction of possessive noun formation with a common V-suffix /i/, Sato proposes to reconstruct transitive verb formation with a common V-suffix /ə/. This /ə/ underwent root-control harmony (= total assimilation) (*kom-ə* > *komo*, *yas-ə* > *yasa*, *yup-ə* > *yupu*). Deviations from total assimilation were /ə/ > /e/ after /y/ (and possibly /w/) (*noy-ə* > *noye*, \**noyo*), and V1 /ə/ > /e/ and V2 /ə/ > /u/ after C2 /s/ (*məs-ə* > *mesə* > *mesu*).<sup>3</sup>

Sato's hypothesis reinforces Chiri's diachronic scenario by assuming a common vowel suffix /ə/ for transitive verb formation. He thereby succeeds in providing a common phonological context, root-control harmony, for both nominal and verbal derivations.

Nevertheless, Sato's hypothesis is not as convincing as Chiri's as he fails to provide a source of the reconstructed /ə/. Unlike Chiri's postulation of /i/ in possessive noun formation, /ə/ is not associated to any known morphological entity nor to any phonological process which might have produced such a vowel. It is this latter point to which we turn in the next section.

<sup>2</sup> I would like to thank Yasushige Takahashi (Hokkaido University) for bringing this work to my attention.

<sup>3</sup> Sato does not provide explanations of these irregular processes.

## 4 The nature of root-control harmony

An interesting case of V1–V2 distribution is found in Nivkh (language isolate, Russia), a geographic neighbor of Ainu. Although the vowel inventory is not entirely identical to that of Ainu (Nivkh has a high central vowel /i/), vowel distribution in Nivkh disyllabic roots exhibits some common characteristics. Table 5 illustrates the observed and expected frequencies of V1–V2 in Nivkh disyllabic roots, and Table 6 the observed/expected ratios (Shiraishi and Botma 2015, data from Pukhta 2002).

Table 5: Observed and expected frequencies (Nivkh).

V1 \ V2	i	ɨ	u	e	a	o	
i	12 (11.1)	14 (3.7)	5 (4.5)	0 (0.8)	1 (8.4)	0 (3.5)	32
ɨ	24 (17.0)	15 (5.7)	9 (6.9)	1 (1.1)	0 (12.8)	0 (5.4)	49
u	21 (14.6)	4 (4.9)	13 (5.9)	0 (1.0)	4 (11.0)	0 (4.6)	42
e	10 (10.4)	0 (3.5)	1 (4.2)	1 (0.7)	17 (7.8)	1 (3.3)	30
a	21 (28.1)	1 (9.5)	8 (11.4)	3 (1.9)	42 (21.1)	6 (8.9)	81
o	13 (19.8)	0 (6.7)	5 (8.0)	2 (1.4)	12 (14.9)	25 (6.3)	57
	101	34	41	7	76	32	291

Table 6: Observed and expected frequencies (Nivkh).

V1 \ V2	i	ɨ	u	e	a	o	
i	1.09	3.50	1.00	0.00	0.13	0.00	32
ɨ	1.41	2.50	1.29	1.00	0.00	0.00	49
u	1.40	0.80	2.17	0.00	0.36	0.00	42
e	1.00	0.00	0.25	1.00	2.13	0.33	30
a	0.75	0.11	0.73	1.50	2.00	0.67	81
o	0.65	0.00	0.63	2.00	0.80	4.17	57
	101	34	41	7	76	32	291

Like Ainu, the high O/E ratio in the diagonal cells from upper left to lower right (Table 6) shows the tendency to line up identical vowels. In Ainu, 41.1% of the verbs have identical V1–V2. In nouns it is 29.6%. In Nivkh, it is 37.1%.

In addition, there are the following asymmetries in the distribution of vowels. Firstly, a high vowel–/a/ sequence hardly ever occurs (*\*i–a*, *\*u–a*, *\*ɨ–a*). Secondly, V2 /o/ exhibits a strong tendency to line up with V1 /o/. In contrast, there seems to be no such restriction on V2 /i u/. Interestingly, these restrictions concern mainly V2 and resemble Chiri’s observation on Ainu seen earlier.

To sum up, both Nivkh and Ainu exhibit 1) a preference for lining up identical

vowels (the strongest being /o/-/o/), and 2) restriction on V2, namely, a high (front) vowel (*i*, *e*) is common whereas back non-high vowel (*o*) is not, unless preceded by an identical vowel. According to Shiraishi & Botma (2015), these two characteristics stem from a single phonological process in Nivkh: stress-dependent height harmony. The stress pattern of Nivkh is trochaic in polysyllabic roots – V1 stressed and V2 unstressed. Stress is realized as high pitch and increased duration. Shiraishi and Botma claim that this prosodic asymmetry affects the distribution of vowels. Being unstressed, V2 is a prosodically weak licenser and therefore able to host only two types of vowels:

1. Vowels which are supported sufficiently by the prosodically strong V1 by means of total or partial assimilation. Total assimilation is the strongest in /o/-/o/.<sup>4</sup> Partial assimilation comprises a case of height harmony, in which a non-high V2 is allowed only with a non-high V1 (*a-e*, *o-e*, *e-a*, *o-a*), but not with a high V1. High–non-high sequences (*\*i-a*, *\*u-a*, *\*i-a*) are excluded since a high V1 fails to license a non-high V2, having no height feature in common.
2. The intrinsically short vowels /i u/. Being *small*, these vowels do not interfere with the prosodic weakness of V2 and conform to the canonical trochaic stress pattern in Nivkh.<sup>5</sup> This is a case of unstressed vowel reduction, in which weak prosodic positions are “more susceptible to co-articulatory effects from neighboring strong vowels” (Barnes 2006: 193). It is this mechanism that Shiraishi and Botma assume to underlie the asymmetric distribution of vowels in Nivkh.

The question is whether a similar mechanism of stress-dependent harmony can be assumed in the asymmetric distribution of vowels in Ainu. One critical difference from Nivkh is that Ainu has iambic as a dominant stress pattern in underived contexts. In the derived disyllabic forms discussed above, stress depends on morphological boundaries: CVC-V is iambic (*ka'y-e*) while CV-CV is trochaic (*sa-ha*, *re-he*). The vast majority of cases are iambic as there are only 25 CV-CV forms. Thus if we were to assume a stress-dependent root-control type of harmony, we are forced to assume that Ainu originally had a trochaic stress pattern which later became iambic at some point in its history. Is such an assumption justifiable?

Possible support for this idea comes from a correlation between trochaic stress patterns and morphological complexity. Ainu is a typical iamb language; in simplex forms, stress falls on the second syllable unless the first is heavy (CVC): *sa'pa* ‘head’, *'sinrit* ‘ancestors’. The interesting case is the morphologically complex forms. According to Sato (2015; forthcoming), the stress pattern in morphologically complex forms is largely dependent on morphological boundaries. When the first element of a complex is CV, trochee is dominant. Sato counts 58 trochees and 6 iambs in Tamura’s Ainu dictionary (1996).

<sup>4</sup> This is possibly a remnant of labial harmony prevailing in the area, as Tungusic and Mongolic languages exhibit a similar pattern (Li 1996).

<sup>5</sup> Like /i u/, /i/ has a short duration. However, it is not small, as it is not a focalized vowel (Harris 2005). See Shiraishi & Botma (2015) for further discussion.

- (5) a. 're-kor  
name-have  
'to have a name'  
b. 'e-re  
eat-CAU  
'let someone eat'  
c. ni-'esisuye  
stick-swing  
'to swing a stick'

When the first element is CVC, the first segment in the second element plays a decisive role. If it is C, the complex is exclusively trochaic (*'tek-moymoke* 'to move hands'). If it is V, iambic is dominant but trochees are also observed. Sato counts 109 iambs and 29 trochees in Tamura's dictionary.

- (6) a. sik-'erayke  
eye-kill  
'to glare at'  
b. am-'us-pe  
claw-attach-thing  
'crab'  
c. 'cip-o  
boat-get\_on  
'to get on a boat'  
d. 'mim-us  
meat-stick  
'to be fat'

Thus while iamb is dominant in simplex forms, trochee is dominant in complex forms. Sato (forthcoming) infers from this fact that trochee is the basic and the archaic stress pattern in complex forms. Iambs (6a, b) are innovative, possibly developed under the influence of simplex forms.

Typically, an iamb in a complex form undergoes resyllabification from CVC.V to CV.CV (*si.ke.ray.ke*, *a.mus.pe*), thereby conforming to the canonical iambic stress pattern in simplex forms (Sato forthcoming). If the iamb-trochee asymmetry were associated with morphological complexity, then the hypothetical historical shift of stress from V1 to V2 that we are assuming above could be associated with a shift in morphological complexity – the shift signals a change from morphologically complex to simplex in the mind of language users. It is this shift that could have led to a mismatch between the stress pattern and the distribution of vowels that we are witnessing now.

## References

- Barnes, Jonathan. 2006. *Strength and weakness at the interface: positional neutralization in phonetics and phonology*. Berlin/New York: Mouton de Gruyter.
- Chiri, Mashiho. 1952. Ainugo ni okeru boinchowa [Vowel harmony in Ainu]. *Bulletin of the Hokkaido University Faculty of Arts* 1. 101–118.
- Ewen, Colin & Harry van der Hulst. 1988. [high], [low] and [back] or [I], [A] and [U]? In Peter Coopmans & Aafke Hulk (eds.), *Linguistics in the Netherlands 1988*, 49–57. Dordrecht: Foris.
- Harris, John. 2005. Vowel reduction as information loss. In Philip Carr, Jacques Durand & Colin Ewen (eds.), *Headhood, elements, specification and contrastivity*, 119–132. Amsterdam: John Benjamins.
- Ito, Junko. 1984. Melodic dissimilation in Ainu. *Linguistic Inquiry* 15 (3). 505–513.
- Kindaichi, Kyosuke. 1931. Ainu yūkara gohō tekiyō [An outline grammar of the Ainu epic poetry]. In *Ainu jojishi yukara no kenkyū*, 1–233. Tokyo: Tōyō Bunko.
- Kindaichi, Kyosuke & Mashiho Chiri. 1936. Ainu gohō gaisetu [An outline of Ainu grammar]. In *Chiri mashiho chosakushū [Writings of Chiri Mashiho]*, vol. 4, 3–197. Tokyo: Heibonsha.
- Krämer, Martin. 1998. A correspondence approach to vowel harmony and disharmony. *Working papers Theorie des Lexicons* 107.
- Li, Bing. 1996. *Tungusic vowel harmony: description and analysis*. The Hague: University of Amsterdam PhD thesis.
- Mester, Armin. 1986. *Studies in tier structure*. Amherst: University of Massachusetts PhD thesis.
- Pukhta, Maria. 2002. *Nivkh-Russian conversation and daily-life thesaurus* (Endangered languages of the Pacific Rim A2-012). Osaka Gakuin University.
- Sato, Tomomi. 2010. Chiri hakase to Ainugogaku [Doctor Chiri and Ainuology]. In Hironobu Kosaka (ed.), *Chiri Mashiho*, 159–173. Sapporo: Crews.
- Sato, Tomomi. 2015. Ainugo no goseigo no akusento kisoku to sono reigai nit suite [Compound accent rules and exceptions in Ainu]. In Anna Bugaeva & Iku Nagasaki (eds.), *Ainugo kenkyū no shomondai*, 1–13. Sapporo: Hokkaido shuppan kikaku center.
- Sato, Tomomi. Forthcoming. Ainugo no goseigo ni okeru reigaiteki akusento to sono rekishiteki kaishaku [Exceptions to compound accent rule and its historical implications in Ainu]. In Yeonju Lee (ed.), *Akusento ronshu*. Sapporo: Hokkaido University.
- Shibatani, Masayoshi. 1990. *The languages of Japan*. Cambridge: Cambridge University Press.
- Shiraishi, Hidetoshi. Forthcoming. Phonetics and phonology. In Anna Bugaeva (ed.), *The handbook of the Ainu language*. Berlin: Mouton.
- Shiraishi, Hidetoshi & Bert Botma. 2015. *Stress-dependent harmony in Nivkh*. Presentation at the 23rd Manchester Phonology Meeting.
- Tamura, Suzuko. 1996. *Ainu dictionary (Saru dialect)*. Tokyo: Sofukan.
- Tamura, Suzuko. 2000. *The Ainu language*. Tokyo: Sansendo.

## Chapter 33

# From dialectometry to semantics

Dirk Speelman

University of Leuven

Kris Heylen

University of Leuven

Aggregate-level studies of linguistic variation typically adopt an onomasiological perspective on linguistic variation. Recently, however, a number of corpus-based techniques have been developed in the distributional semantics framework to detect semantic shifts across large corpora (Sagi, Kaufmann & Clark 2011; Cook & Hirst 2011; Gulordava & Baroni 2011). In this chapter, we apply one such technique to the corpus-based, aggregate-level investigation of semasiological regional variation in Dutch. More specifically, we use token-based vector space models, in which (a random subset of) the tokens of a target word are represented as a ‘token cloud in vector space’. In order to compare the use of a target word in two regional varieties of Dutch, viz. Netherlandic Dutch and Belgian Dutch, we build a token cloud for each variety and superimpose both token clouds. Next, we apply measures, which we call *separation indices*, and which quantify to which extent the superimposed clouds exhibit non-overlapping areas (or areas with less overlap). Such areas are of interest because they signal possible differences in the (number of) senses or usage patterns of the word in both varieties. The purpose of these *separation indices* is to quantify semasiological distances between language varieties and by consequence to allow for a (dia)lectometric approach to the study of semasiological variation. This chapter reports on a methodological pilot study that investigates the merits of four candidate types of *separation indices*.

## 1 Introduction

The methods that were developed within the framework of dialectometry (Nerbonne & Kretzschmar 2003) have been a rich source of inspiration for many different types of studies into aggregate-level language variation, both within dialectometry *sensu stricto* and in lectometry more in general – we use ‘lectometry’ as an umbrella term for dialectometry, stylometry, sociolectometry, etc. In this chapter, we specifically zoom in on corpus-based studies of regional and register variation that adopt

a ‘lectometric approach’. A specific challenge in all corpus-based lectometric studies on lexical/grammatical variation is dealing with semantic differences. Whereas in survey data of the type often used in dialectometric studies, the context in which words/expressions are used by survey participants is kept constant, such consistency does not apply to corpus materials. The contexts in which words are used in corpora vary dramatically from one instance to another and issues related to e.g. polysemy and vagueness are hard to address, especially when the corpora are large and manual inspection of all usage instances is not an option.

Semantic vector space models, or simply vector space models or VSMs (Turney & Pantel 2010), a technique that is often used in natural language processing in tasks such as thesaurus extraction and word sense disambiguation, offer promising possibilities for semi-automatically accounting for lack of synonymy in corpus-based studies of lexical/grammatical variation. For instance, Ruetten et al. (2014) incorporate VSMs into a ‘lectometric framework’. In that study, in which an onomasiological perspective is adopted, a weighting mechanism is installed that penalizes, in the ‘lectometric calculations’, data points that occupy a peripheral position according to the semantic similarity scores that were derived from VSMs. This way the effect of potentially problematic data points on the lectometric results is reduced.

In this chapter, rather than attempting to neutralize potential noise coming from non-synonymy, we make non-synonymy, or rather, semasiological variability, the topic of (lectometric) investigation, thus switching to a semasiological perspective. VSMs play a crucial role in our approach. Using token-based VSMs, we build so-called token clouds, for a specific target word, for two regional varieties of Dutch, viz. Netherlandic Dutch and Belgian Dutch and we superimpose both token clouds.

Token clouds will be explained in more detail in Section 2. For the time being, Figure 1 informally illustrates the concept. All panels in Figure 1 show token clouds for the Dutch word *monitor*. In the middle panel we see a token cloud for Netherlandic Dutch (NL), with individual points representing individual tokens. Proximity between tokens is a proxy for semantic similarity of the usage contexts of the tokens. In the right panel we see a token cloud for Belgian Dutch (BE). In the left panel both clouds are superimposed. The example illustrates that there is an area where there are only Belgian Dutch tokens (roughly coinciding with the bottom right quadrant of the plots). Manual inspection of the tokens reveals that by and large the tokens in this area are tokens where *monitor* has the meaning YOUTH LEADER, whereas in the other parts of the plot, where the token clouds do overlap, most tokens have either the meaning COMPUTER SCREEN or the meaning SCREEN OF A MEDICAL DEVICE. As it turns out, in Netherlandic Dutch the word *monitor* lacks the meaning YOUTH LEADER, which is exactly what is reflected in the presence of the non-overlapping area.

Having built our token clouds, we then apply measures, which we call *separation indices*, and which quantify to which extent the superimposed clouds exhibit non-overlapping areas (or areas with less overlap). As illustrated in the example in Figure 1, such areas are of interest because they signal possible differences in the (number of) senses or usage patterns of the word in both varieties. The purpose of

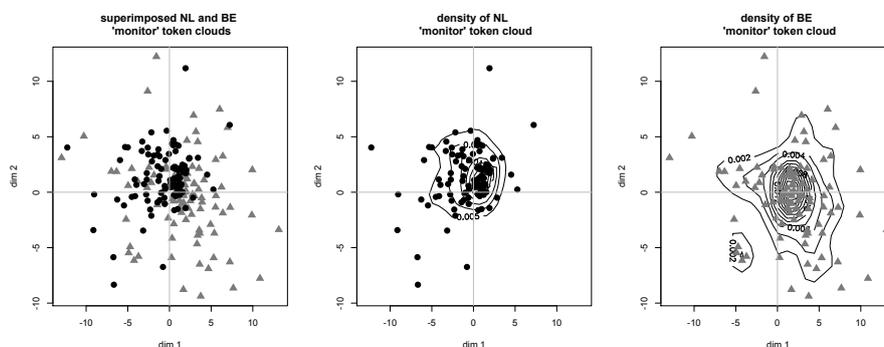


Figure 1: Example of two superimposed token clouds.

these *separation indices* is to quantify semasiological distances between language varieties and by consequence to allow for a (dia)lectometric approach to the study of semasiologic variation. This chapter reports on a pilot study that investigates the merits of four candidate types of *separation indices*.

The structure of the chapter is as follows. In Section 2, we explain how the token clouds are built and we describe four candidate types of *separation indices*. In Section 3, we present results from a case study in which we explore to which extent these *separation indices* yield sensible results. In Section 4, finally, we draw some conclusions.

## 2 The method: token clouds and separation indices

In this section, we first explain how we build token clouds in vector space. Next, we describe the different candidate types of *separation indices* that will be used in the case study in the next section.

### 2.1 Building VSMs

The most commonly used kind of VSMs are so-called type-based VSMs. These VSMs are matrices in which the usage, in the corpus, of each target word is summarized in a single row. Columns represent so-called features. In the VSMs used in this chapter, these features simply are words that occur in the vicinity of the target word. The cells in the matrix express the frequency with which each target word co-occurs with each feature, or rather, they contain so-called positive pointwise mutual information values (PPMI) that are derived from the raw co-occurrence frequencies. PPMI values express the ‘attraction’ between a target word and a feature.

The rationale then is that, given a large enough corpus, target words with similar meanings tend to have similar row vectors. Therefore, distances between row vectors (typically calculated as one minus the cosine of the angle between the two row vec-

tors) can be used as a proxy for differences in meaning/usage. Table 1 illustrates the structure of a type-based VSM (using English words). Only a few rows and columns are shown, and the values in the cells are not shown. The idea would be that in a good VSM the row vectors of *car* and *vehicle* would be much more similar to each other than to the row vector of *coffee*.

Table 1: Part of a type-based VSM.

	<i>home</i>	<i>drink</i>	<i>traffic</i>	<i>wheel</i>	...
<i>car</i>	...	...	...	...	...
<i>vehicle</i>	...	...	...	...	...
<i>coffee</i>	...	...	...	...	...
...	...	...	...	...	...

In token-based VSMs, each instance (=token) of a word in the corpus, or at least a representative number of such instances, has its own row. This is illustrated in Table 2, using, as an example, tokens of the English word *car* (please ignore, for the time being, that Table 2, contains tokens from two varieties). In token-based VSMs, it would not be a good idea to simply use the words in the vicinity of the target word as ‘atomic’ features (i.e. as columns). Since there are only a few context words in each token, this would lead to very sparse, very uninformative row vectors for the individual tokens. For instance, switching again to the *monitor* example from the introduction, in such an ill-chosen approach the fact that one *monitor* token has the word *kinderen* ‘children’ in its context and another *monitor* token has *jongeren* ‘youngsters’ in its context, would not lead to the desired effect of their two row vectors somehow resembling each other (since *kinderen* and *jongeren* would be treated as unrelated ‘atomic’ features).

What we do instead to build the row vector of a token in a token-based VSM, is for each of the context words in that token, we first retrieve its type vector, i.e. we retrieve its row vector from a type-based VSM. In order to clarify this, let us use the notation Cs for the context words that appear in a specific token. For instance, in the token *...I park my car in the second garage ...*, with *car* the target word, words such as *park*, *second*, and *garage* would be Cs. It is of these Cs that we retrieve the type vectors. Let us, in the present context, use the notation CCs for the features that are used in the type vectors of the Cs. In other words, the CCs are the (type-based) features of the (token-specific) Cs of the target word.

Having retrieved the type vectors of all Cs in a token, we add up these type vectors to build the token representation of the token (so that the CCs will become the features of the token-based VSM). For instance, the row representation of the example token just given would be the sum of the type vectors of *park*, *second*, *garage*, etc. It

should be added, though, that we use a weighted sum, in order to take into account that some Cs (e.g. *park, garage*) are more important than others (e.g. *second*). More specifically, the row vector of a token is the weighted sum of the type vectors of all its Cs, with the weights being the importance of the Cs, measured as the type-based ‘attraction’ (PPMI) between the target word and the C. A less concise description of this procedure, which is a slightly modified version of Schütze (1998), can be found in Heylen et al. (2015). Importantly, returning once again to the *monitor* example, in this approach, the row vector of a token with the C *kinderen* ‘children’ in it and the row vector of a token with the C *jongeren* ‘youngsters’ in it will tend to be similar, because the type vectors of these two important Cs can be expected to be very similar.

The above describes how token-based VSMs for a single variety can be built. Superimposed token-based VSMs for two varieties, as illustrated in Table 2, are a straightforward extension. This time we use one type-based VSM for each variety (both having the same features) and we use a sufficiently large sample of tokens from both varieties. With this, we create a matrix, as illustrated in Table 2, that has as its rows the tokens from both varieties. The row vectors are calculated as before, with this complication that for building the row for a token from variety A, information is retrieved from the type-based VSM for A, and for building the row for a token from variety B, information is retrieved from the type-based VSM for B.

Table 2: Part of a matrix with two superimposed token-based VSMs for the target word *car*.

	$CC_1$	$CC_2$	$CC_3$	$CC_4$	...
car 1 from US	...	...	...	...	...
car 2 from US	...	...	...	...	...
car 3 from US	<i>weighted sum of Cs of car 3 from US</i>				
...	...	...	...	...	...
car 1 from UK	...	...	...	...	...
car 2 from UK	...	...	...	...	...
...	...	...	...	...	...

## 2.2 Token clouds in original vector space and reduced vector space

We speak of token clouds, because you can think of the information in a token-based VSM as a cloud of points (the tokens) sitting in a high-dimensional space, in which each CC is a dimension and the PPMI values are coordinates. In spite of the high-dimensional nature of such a space, it is straightforward to calculate (cosine-based) distances between the points (=tokens). However, if we want to be able to visualize the tokens cloud, we need to reduce the number of dimensions. We solve this by

applying non-metric MDS to the original matrix (in which we used cosine-based distances), in order to derive from it a 2D-simplification (in which we use Euclidean distances). We call this two-dimensional space the ‘reduced vector space’, as opposed to the ‘original vector space’ we started out with. It is this ‘reduced vector space’ that is represented in plots such as the ones in Figure 1.

### 2.3 Separation indices

In the case study in the next section, we will test four types of *separation indices*, which, as explained before, are meant to quantify the degree to which there are non-overlapping areas in the superimposed token clouds. We call the first two measures ‘global’ indices, because they assess to which extent the clouds as a whole tend to consist of ‘larger areas’ that lack overlap. The final two measures, on the other hand, are ‘local’ indices; they assess to which extent, at a more fine-grained level, there are smaller areas that lack overlap.

The first ‘global’ index, DR, which stands for *distance ratio*, is a slightly modified version of the clustering index proposed in McClain & Rao (1975). For each item (=token), we calculate  $A$  the mean distance from the item to other items from the same class (=variety) and  $B$  the mean distance from the item to items from the other variety. The *separation index* for the item is  $B/A$ . The *separation index* for the complete token cloud is the mean *separation index* of the items.

The second ‘global’ index, SIL, stands for *silhouette width* (Rousseeuw 1987). For each item, we calculate  $A$  the mean distance from the item to items from its own class (=variety), and  $B$  the mean distance from the item to the other class (=variety). The *separation index* for the item is  $(B - A)/\max(A, B)$ . The *separation index* for the complete token cloud is the mean *separation index* of the items.

The first ‘local’ index, SCP, stands for (smallest) ‘*same class path*’. This index takes a parameter  $k$  by means of which you specify the level of granularity you want to inspect (with smaller  $k$  corresponding to more ‘local’ patterns). It expresses how easy it is to draw paths that connect  $k + 1$  same-variety tokens while encountering as few other-variety tokens as possible. The *separation index* for an item is calculated as the separation score of the shortest path of length  $k$  that connects that item to other items from the same class (=variety); the separation score of this path is the mean of the separation score of the steps it consists of; the separation score of any step from  $A$  to  $B$  is one divided by the rank of the distance of  $B$ . The *separation index* for the complete token cloud is the mean *separation index* of all items. The explanation of this index sounds complicated, and needs further explanation, but the idea is simple. In order to determine the *separation index* for a token, the procedure tries to build a path that connects this token to  $k$  other tokens of the same variety (in one chain, stepping from  $A$  to  $B$ , from  $B$  to  $C$ , etc.) and that is as small as possible. The smaller the path that is found, the higher the *separation index* for the token. A path is small if the average length of its individual steps is small. How small an individual step from  $A$  to  $B$  is, is determined by how many tokens from the other variety are closer to  $A$  than  $B$  is. The fewer such other-variety tokens there are, the smaller the step, and the higher its separation score. For instance, if there are no other-variety tokens

that are closer to *A* than *B* is, then the distance of *B* has rank 1 and therefore the separation score of the step going from *A* to *B* is  $1/1$ , which is 1, and which is the highest possible separation score. If there is one other-variety token closer to *A* than *B* is, then the distance of *B* has rank 2 and the separation score of the step from *A* to *B* is  $1/2$ . If there are two other-variety tokens closer to *A* than *B* is, then the rank is 3 and the separation score is  $1/3$ . Etc. In sum, and informally: a step is small (and therefore its separation score is high) if it doesn't cross much 'territory occupied by the other variety'. In a similar vein, the complete path of length *k* of a token is small (and therefore the token's *separation index* is high) if the path doesn't cross much 'territory occupied by the other variety'.

The second 'local' index is KNN, which stands for *k nearest neighbours*. This too is a measure that takes a parameter *k* by means of which one can specify a level of granularity. For each item (=token) we calculate the proportion of same-class items (=same-variety items) among its *k* nearest neighbours; that proportion is the *separation index* for the item. The *separation index* for the complete token cloud (i.e. all items) is the mean *separation index* for the items.

Although these four types of *separation indices* (DR, SIL, SCP, KNN) have different scales, they share a number of characteristics. Firstly, higher values indicate more presence of non-overlapping areas (so a higher degree to which the varieties occupy separate areas in the superimposed token clouds). Second, they take as their input the distances between the tokens. Since we have distances between the tokens both for the 'original vector space' and for the 'reduced vector space', we apply the *separation indices* to both.

As a result, we end up with eight sets of *separation index* calculations, since we have four types of *separation indices*, which we all apply to both the 'original vector space' and to the 'reduced vector space'. The empirical questions then are be to which extent the results will be similar across the eight sets, and, if not, how they differ.

### 3 Case study

We have built token clouds for 42 words, 21 of which are mentioned in several language advice resources (such as taaltelefoon.vlaanderen.be) as being used differently in Belgium and The Netherlands. For the other 21 words we found no such claims, nor did we have any other reason to expect semasiological regional differences. Of the former 21 words, 7 are claimed, in the language advice literature, to differ with respect to the (fixed) expressions or idioms that are often used in the two varieties, and 14 are claimed to have different possible/popular senses in the two varieties.

The following are the words that are included in the study:

- category no (no claims about differences found): *appel, auto, ballon, bos, broek, bureau, centrum, deur, dier, fruit, gebruiker, heling, kamer, kop, land, nacht, neus, school, steun, stoel, verlof*;

- category expr (expressions/idioms are claimed to differ): *biecht, boontje, geschenk, mosterd, mouw, straatje, vijg*;
- category sense (senses are claimed to differ): *academicus, bank, bolletje, kledje, kous, middag, monitor, pan, patat, poep, puntje, tas, vlieger, wagen*.

For each of these words, we randomly collected 300 tokens from a large Belgian newspaper corpus (LeNC=Leuven Nieuws Corpus; 1.2 billion words) and 300 tokens from a large Netherlandic newspaper corpus (TwNC=Twente Nieuws Corpus; 500 million words), and we merged both sets in one token cloud. We used about 5000 CCs (the intersection of the top 7000 high frequency words in LeNC and TwNC, minus the top 100 high frequency words); the context window used for the type-based VSMs was 4:4 (i.e. four words to the left and four to the right of the target word). The context window for determining the Cs was 10:10. Of the candidate Cs, we only kept those that were sufficiently important according to the type-based VSM of the variety the token came from ( $LLR > 1$  and  $PPMI > 1$ ) and that also occurred in the corpus for the other variety. We dropped tokens without suitable Cs (typically retaining about 500 tokens out of the original 600). Stress in the MDS solution that we used to build the ‘reduced vector spaces’ varied from .15 to .28.

We then calculated the four *separation indices* (DR, SIL, SCP, and KNN), both for the ‘original vector space’ and for the ‘reduced vector space’. In the ‘local’ *separation indices* SCP and KNN,  $k$  was set to 10 and an additional weighting procedure was used (that we will not go into). Finally, the resulting eight sets of *separation index* results were all standardized, in order to make it easier to compare them.

For all eight sets of *separation index* results, we ran a regression analysis with standardized *separation index* as response variable and with word category as predictor. Word category (cat) had the levels no, expr, and sense; we used dummy coding (=treatment coding), with cat=no as reference value. Figure 2 shows, for all eight regression models, the estimates for cat=expr and cat=sense (with 95% confidence intervals).

A few observations can be made. First, in all eight models the average *separation index* in the case of cat=expr is significantly higher than in the case of cat=no. Second, in only a few models the average *separation index* in the case of cat=sense is significantly higher than in the case of cat=no. More specifically, the latter effect is least present in models in which ‘global’ *separation indices* are applied to the ‘original vector space’, and is most clearly present in model in which ‘local’ *separation indices* are applied to the ‘reduced vector space’.

After obtaining these results, we replicated the case study four times, with the same corpora and the same words, but each time taking other random subsets of tokens. Each time, the results were very similar; the aforementioned observations were robust across all replications.

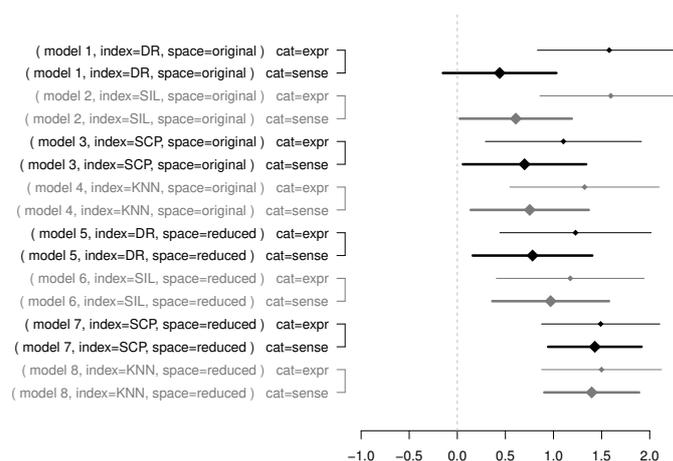


Figure 2: Estimates for cat=expr and cat=sense (with 95% confidence intervals) in 8 regression models, with reference level cat=no.

## 4 Conclusions

In this chapter, we explored the possibility of applying (dia)lectometric techniques to the investigation of (aggregate-level) semasiological variation. For such a thing to be possible, it is necessary to be able to quantify semasiological differences across language varieties. In a methodological pilot study, we tested eight different ways of quantifying semasiological differences. All approaches that were tested produced sensible results with respect to the detection of regional differences at the level of ‘different (fixed) expressions or idioms’. Regional difference at the level of ‘different (number of) senses’, on the other hand, proved harder to detect, with the approach that applied ‘local’ *separation indices* to the ‘reduced vector space’ outperforming the other approaches. The results suggest that the dimension reduction can be instrumental in the quantitative identification of semasiological patterns in the data.

## References

- Cook, Paul & Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 265–274.
- Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of*

Dirk Speelman & Kris Heylen

*the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 67–71.

- Heylen, Kris, Thomas Wielfaert, Dirk Speelman & Dirk Geeraerts. 2015. Monitoring polysemy. Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172.
- McClain, John O. & Vithala R. Rao. 1975. Clustisz: a program to test for the quality of clustering of a set of objects. *Journal of Marketing Research* 12(4). 456–460.
- Nerbonne, John & William Kretzschmar. 2003. Introducing computational techniques in dialectometry. *Language Resources and Evaluation* 3. 245–255.
- Rousseeuw, Peter F. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* 20. 53–65.
- Ruette, Tom, Dirk Geeraerts, Yves Peirsman & Dirk Speelman. 2014. Semantic weighting mechanisms in scalable lexical sociolectometry. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: linguistic variation in text and speech*, 205–230. Berlin/New York: De Gruyter Mouton.
- Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 161–183. Berlin/New York: De Gruyter Mouton.
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.
- Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.

## Chapter 34

# On blended selfies and tainted smoothies

Oscar Strik

European Research Council; University of Antwerp

Muriel Norde

Humboldt University Berlin

Karin Beijering

Research Foundation - Flanders (FWO); University of Antwerp

### 1 Introduction

There's a particular kind of word that acquires such a cultural momentum that it is able to skyrocket from non-existence to ubiquity within the span of a dozen years. The word *selfie* is a textbook example. Due to the rise of both mobile phone cameras and social media, the selfie now has a wide cultural distribution, and as such the word has become commonplace. Such is its popularity that it was elected Word of the Year in The Netherlands (2013), Belgium (2013), Great Britain (2013), France (2014), Spain (2014, as *selfi*), and as *Jugendwort des Jahres* in Austria (2014).<sup>1</sup> All this is hardly news, but a more interesting phenomenon is that the word *selfie* has been spawning a plethora of variants, such as *groupfie* and *shoefie*, and not just in English, but in several of the languages that have borrowed the word.

In this article, we present an impressionistic inventory of new words based on *selfie* in Swedish and Dutch. Adopting a Construction Morphology approach (Booij 2010; Norde & van Goethem 2015), we will examine which word formation processes are used to create these new forms. We also compare what we will call the *selfie* 'family' to the *smoothie* 'family', which shows the potential for similar morpho-phonological

<sup>1</sup> <https://onzetaal.nl/nieuws-en-dossiers/dossiers/woorden-van-het-jaar;> [https://de.wikipedia.org/wiki/%C3%96sterreichisches\\_Wort\\_des\\_Jahres.](https://de.wikipedia.org/wiki/%C3%96sterreichisches_Wort_des_Jahres)

processes in word formation, but which has a much smaller level of actual productivity.

## 2 Formal and semantic properties of *selfie* and *smoothie*

The suffix *-y/-ie* has diminutive meaning, both when used with proper names (*Charlie, Davey*) and when used with other nouns (*baby, ducky*) (Shields 2001). The coinage of the word *selfie*, first attested in Australian English in 2002 according to the *Oxford English Dictionary*,<sup>2</sup> makes use of the nominalisation function of the diminutive suffix, a process that is used extensively in Dutch (Booij 2010: 54), but also in other English words such as *quicke*, *cutie*, and *smoothie*.

The meaning of *selfie* is non-compositional – after all, it does not mean ‘small self’ – and seems to be subject to change, since the OED definition “a photographic self-portrait; esp. one taken with a smartphone or webcam and shared via social media” need not strictly apply in all cases anymore. This was recently illustrated by the case of British national Ben Innes posing with Seif Eldin Mustafa, the Egyptian hijacker of EgyptAir Flight 181, in a photo that went viral on the Internet and was often described as a *selfie* of him and the hijacker.<sup>3</sup> Despite the label, the photo was not taken by Innes himself, but by a stewardess. It is, strictly speaking, not a *self*-portrait, but a regular one. All the same, the usage indicates that the meaning of *selfie* is expanding to include the meaning ‘photograph of self with someone notable’, with the meaning of the ‘self’ element shifting from who actually holds the camera to who is the ‘topic’ of the photograph. This broadening in meaning is illustrative of the flexibility in semantics that we also encounter in the derivatives of *selfie*, as we will see.

The term *smoothie*, in turn, has been attested in its currently relevant meaning since 1977.<sup>4</sup> The *smoothie* itself is a blend – physical, not linguistic – of fruit/vegetable juice/pulp and dairy (e.g. (soy) milk or yoghurt). The mouth texture of the drink appears to be the reason to use the suffix *-ie* to transform adjective *smooth* into a count noun. Again, the semantics are non-compositional, since a *smoothie* is not just anything that is smooth (to the taste), but a specific type of drink.

The construction of *selfie* can be represented in the following way, using conventions of Construction Morphology:

- (1)  $[[self]_i -ie]_j \leftrightarrow [photograph\ that\ features\ and\ is\ taken\ by\ SEM_i]_j$

Similarly, *smoothie* is constructed as follows:

- (2)  $[[smooth]_i -ie]_j \leftrightarrow [drink\ that\ has\ a\ SEM_i\ mouth\ texture]_j$

<sup>2</sup> <http://www.oed.com/view/Entry/390063?redirectedFrom=selfie#eid>.

<sup>3</sup> <http://www.theguardian.com/uk-news/2016/mar/30/briton-ben-innes-posed-selfie-egyptair-hijacker-praised-by-relatives>.

<sup>4</sup> See <http://www.oed.com/view/Entry/182768?redirectedFrom=smoothie#eid>. The word also has an earlier meaning ‘smooth talker’, ‘suave man’, but this sense is not likely to be relevant to the *smoothie* that is our current subject.

Both the formal and semantic make-up of these constructions is flexible when it comes to the coinage of variants of these original words, as will become clear in the treatment of these new forms in Swedish and Dutch.

Finally, we may observe that in both Swedish and Dutch the original words are treated as a unit with respect to syllable boundaries, rather than keeping the suffix as a separate syllable; that is to say, for the purposes of expanding the pattern, *self-ie* and *smooth-ie* are re-segmented as *sel-fie* and *smoo-thie*, respectively. Thus, *-ie* qualifies as a cohering suffix (Booij 2005: 115; 2010: 8–9). As we will see, this re-segmentation is reflected in the usage of *-fie* and *-thie* as word creation suffixes.

### 3 Data

#### 3.1 The *selfie* group

##### 3.1.1 Swedish

To collect our Swedish data, we used the Twittermix-subcorpus of Språkbanken (324,570,469 tokens), which was queried on September 8 and 9, 2016.<sup>5</sup>

According to the Swedish *Institutet för Språk- och Folkminnen* (ISF) the word *selfie* appeared in the Swedish language in 2013.<sup>6</sup> In that same year, it was chosen as the most popular new word by readers of the popular scientific journal *Språktidningen*. In line with the Swedish tradition to find Swedish equivalents for foreign words, ISF recommended to use *egobild* ‘ego picture’, or *självis* (formed from the pronoun *själv* ‘self’ + the nominalizing suffix *-is*) instead. However, these alternatives did not really catch on. In the *Twittermix* corpus, the singular form *selfie* occurs 4,939 times, against 43 for *självis* and 27 for *egobild*. Attempts to at least establish native inflection of the loanword have not been very successful either. According to *Svenska Akademiens Ordlista*, the plural form of *selfie* ought to be *selfier*.<sup>7</sup> However, this form is only attested once in the corpus. The most frequent plural form is *selfies* (1,060 tokens), according to what in Swedish grammars is known as the seventh declension (Teleman, Hellberg & Andersson 1999: 79ff). This declension features a plural in *-s* and is largely restricted to loans, particularly those from English (Mickwitz 2010).

So let us now turn to the neologisms based on *selfie*. Using the *Korp* interface of *Språkbanken* we searched for all words ending in *-fie*, excluding personal names – to avoid the Swedish fondness of compound first names with Sofie as the second member – and irrelevant nouns such as (misspelled) *filosofie* or *coffie*. This left us with 6,907 tokens, many of which were unanalysable and/or impossible to check as the original tweet was long gone. We therefore had to refrain from providing a detailed quantitative analysis, but a survey of the available data suggests some interesting

<sup>5</sup> <https://spraakbanken.gu.se/korp/>.

<sup>6</sup> <http://www.sprakochfolkminnen.se/download/18.42699e142b734b551101b3/1398151044710/Nyordslistan-2013.pdf>.

<sup>7</sup> <http://www.sprakochfolkminnen.se/sprak/sprakradgivning/frage/1adan.html?url=-27634753%2Fcgi-bin%2Fsrfl%2Fvisasvar.py%3Fsok%3Dselfie%26svar%3D78404&sv.url=12.c17e514db30bb2a810ea>.

patterns nevertheless. First of all, most new-formations appear to be compounds,<sup>8</sup> which fall into various semantic categories:

- (a) **pictures of self with a specific group or person:**  
*gruppselfie* ‘group selfie’, *familjeselfie* ‘family selfie’, *påve-selfie* ‘selfie with pope’;
- (b) **pictures of self with some being or object:**  
*ko-selfie* ‘selfie with cows’, *kyrkselfie* ‘selfie with church’;
- (c) **pictures of self in a specific state:**  
*nakenselfie* ‘picture of naked self’, *sjuksselfie* ‘picture of sick self’, *svettselfie* ‘picture of sweating self’;
- (d) **pictures of self in a specific location:**  
*tunnelbaneselfie* ‘picture of self on the tube’, *toaselfie* ‘loo selfie’, *badrumselfie* ‘bathroom selfie’;
- (e) **pictures of self at a specific time:**  
*morgonsselfie* ‘morning selfie’, *julsselfie* ‘Christmas selfie’;
- (f) **pictures of self performing a specific activity:**  
*valsedelutdelningsselfie* ‘ballot-paper-handing-out-selfie’, *skiselfie* ‘skiing selfie’;
- (g) **pictures of object belonging/related to self:**  
*skuggselfie* ‘picture of own shadow’.

Quite frequent, but less interesting, are *-fie* formations and blends borrowed directly from English: *wefie* ‘picture of us’, *dogfie*, *felfie* ‘farmer selfie’, *shelfie* ‘picture of bookshelf’, *smellfie* ‘picture of self in a badly smelling situation’,<sup>9</sup> *trelfie* ‘selfie with two others’.

Finally, there are some *-fie* formations involving native roots, but these are rare, presumably because Swedish lacks a diminutive suffix and hence English *-ie* (and its extended variant *-fie*) were not always recognized as a potential suffix (also because English *-ie* is not particularly productive anyway). Interestingly however, some of the *-fie* formations below have a *-selfie* equivalent (*familjeselfie* – *familjifie* ‘family selfie’, *skiselfie* – *skifie* ‘skiing selfie’):

<i>bomfie</i> ‘barrier selfie’	<i>badfie</i> ‘bathroom selfie’
<i>hissfie</i> ‘elevator selfie’,	<i>barnfie</i> ‘selfie with kids’
<i>famil(j)fie</i> ‘family selfie’,	<i>skifie</i> ‘skiing selfie’
<i>tomtjie</i> ‘picture with Christmas Gnome’	<i>chantarellfie</i> ‘picture of Chanterelles in a frying pan’
<i>valfie</i> ‘picture of self voting’	

<sup>8</sup> Note that there are bound to be many more *selfie* compounds in *Twittermix*, given the Swedish tendency to spell compounds as two words (*Språkriktighetsboken* 2005: 43), but we did not look at these.

<sup>9</sup> Mostly falling into two categories: people smelling their own armpits, and (primarily) men changing a baby’s nappy with nose and mouth covered.

Also worth noting is *ny frisyrfie* ‘new haircut selfie’, where *-fie* takes scope over a noun phrase.

### 3.1.2 Dutch

The inventory of Dutch forms is based on the *Groningen Twitter Corpus*, which was queried on August 4 and September 25, 2016.<sup>10</sup> To take stock of Dutch coinages based on *selfie*, we looked at the most frequent words starting with a hash (#) and ending in *-fie*. This allowed us to find forms that were often used along with a photo, so we could look at it and determine the semantics of the form in question in cases where there was any doubt.<sup>11</sup>

Similarly to the Swedish case, the introduction of *selfie* in Dutch met with some resistance. Notably, upon *selfie* becoming the Dutch Word of the Year 2013, writer and comedian Kees van Kooten suggested that it be replaced with *otofoto*, a form derived from *auto-* ‘self’ and *foto* ‘photograph’.<sup>12</sup> While the suggestion garnered some positive responses, the form appears not to have caught on, appearing only 333 times in the Groningen Twitter Corpus, as compared to 182,423 for *selfie*.

While Dutch also uses compound variants on *selfie* (e.g. *sambaselfie*, *nomakeupselfie*), blended and suffixed forms appeared to be more common than in Swedish. In terms of semantic variation, the Dutch forms cover a range similar to that of the Swedish forms:<sup>13</sup>

- (a) **pictures of self with a specific group or person:**  
*solfie* ‘picture of self with a soldier’, *polfie* ‘picture of self with a police officer’;
- (b) **pictures of self with some being or object:**  
*brilfie* ‘picture of self wearing glasses’;
- (c) **pictures of self in a specific state:**  
*pietfie* ‘picture of self in blackface as Zwarte Piet’;
- (d) **pictures of self in a specific location:**  
*stufie* ‘picture of self in a recording studio’, *kanselfie* ‘picture of self preaching from a pulpit’;
- (e) **pictures of self at a specific time:**  
*ontbijtselkie* ‘picture of self at breakfast’, *oranjeselfie* ‘picture of self at Koningsdag [King’s Day]’;
- (f) **pictures of self performing a specific activity:**  
*stemfie* / *stemsselfie* ‘picture of self voting’, *stufie* ‘picture of self studying’, *doehetselfie* ‘picture of self making DIY renovations to the house’;

<sup>10</sup> <http://www.let.rug.nl/gosse/Ngrams/>.

<sup>11</sup> Interestingly, this indicates that in the case of some tweets, not just the linguistic but also the photographic context is needed to determine the meaning of the *selfie* coinage.

<sup>12</sup> <http://nos.nl/artikel/587976-otofoto-ja-zeggen-twitteraars.html>.

<sup>13</sup> Note that some of these forms may have additional meanings.

(g) **pictures of object belonging/related to self:**

*zwerfie* ‘picture of collected litter’;

(h) **pictures of self reflected in object:**

*spiegelfie*, ‘picture of self in mirror’, *balfie* ‘picture of self reflected in spherical Christmas ornament’.

### 3.2 The *smoothie* group

#### 3.2.1 Swedish

In the *Twittermix* corpus, the loan *smoothie* occurs 1096 times in the indefinite singular, and 448 times in the indefinite plural, but we found only two spin-off formations, to wit *goothie* (< *god* ‘seed’ + *-thie*), and *semloothie* (< *sembla* ‘sweet bun’ + *-thie*). Both of these cases appear to be one-off coinages by individuals.

#### 3.2.2 Dutch

In Dutch the *smoothie* family appears to be similarly limited. While we had initially hoped to find more forms because of the presence of the word *oathie* in the Dutch linguistic landscape,<sup>14</sup> no analogous forms were found in the Groningen Twitter Corpus. The only reasonably frequent forms were regular compounds such as *aardbeiensmoothie* ‘strawberry smoothie’ and *spinaziesmoothie* ‘spinach smoothie’.

Interestingly, the *Oathie* is *not* a type of drink; instead, it is the brand name of a type of wheat-free bread that is marketed primarily on health grounds,<sup>15</sup> which might explain the association with a *-thie* suffix. The presence of the suffix is indicated by the *-h-*, since the base word is *oat*, on account of the oat flakes that are a key ingredient. It would have been possible to simply have called the bread *Oatie*, using the classic *-ie* suffix; the fact that it was not indicates that an association with *smoothie* was likely intended. In the case of Dutch, the combination of a word like *oat* with *-thie* may be made easier due to the frequent pronunciation of /ð/ and /θ/ as /t/ in English loans. This is affirmed by the *Oathie* brand itself, who offer a pronunciation guide “ootie” (i.e. /oti/) on their website.<sup>16</sup>

While strictly falling outside of the scope of this section, we may mention the existence of *Fruithie* brand names in both Czechia<sup>17</sup> and Taiwan,<sup>18</sup> which indeed refer to types of fruit smoothie. This further suggests that in the case of the *smoothie* family, the creation of original brand names is a main contributor to its admittedly limited productivity.

<sup>14</sup> Among other places on billboards of the supermarket chain AH to go.

<sup>15</sup> Tapping into the recent consumer trend of treating wheat and gluten as unhealthy, even for people who have no gluten allergy.

<sup>16</sup> <http://www.oathie.com/wat-is-het.html>.

<sup>17</sup> <https://vimeo.com/103649373>; <https://www.facebook.com/cafelagarto/photos/a.167918056565094.36662.114764291880471/1129909533699270/?type=3&theater>.

<sup>18</sup> [https://www.facebook.com/fruithie/?hc\\_location=ufi](https://www.facebook.com/fruithie/?hc_location=ufi).

## 4 Blends and taints

Before we move on, some brief words on the terminology of different types of morphological word formation processes will be helpful.

Some of the forms we've described appear to have a lot in common with blends. Blends (e.g. *smog*, *motel*) are often defined as "a word constructed from the beginning of one word and the end of another" (Bauer 2004: 22).<sup>19</sup> Analysis as a blend is possible if the form in question is composed of a first part of word X and a latter part of *selfie* or *smoothie*. This is particularly plausible in cases where the overlap is greater than just the *-fie/-thie* element, such as *helfie* 'picture of one's hair', *belfie* 'picture of one's posterior', *welfie* 'workout selfie', *drelfie* 'picture of self when drunk', and *felfie* 'farmer selfie'. The aforementioned *groothie* 'seed smoothie' and *semloothie* 'sweet bun smoothie' are examples from the *smoothie* family.

The other cases, such as *groupfie* and *oathie*, are composed of one lexical morpheme and a generalized suffix-like element *-fie/thie*. These cases are slightly different from so-called *-gates* (i.e. variants on *Watergate*, such as *nipplegate* and *Monicagate*) where the conjoined element is recognizable as a full lexeme in its own right. (Hüning 2000).<sup>20</sup> Rather, what we are dealing with is an original suffix that acquires part of the phonological material of the word to which it is attached, a process described by Jespersen (1922: 386–388) as *extension of suffixes*. Haspelmath (1995: 8–9; 2002: 56) calls it *secretion*, using *-aholic* and French *-tier* (as in *bijou-tier* < *fruit-ier*) as examples.<sup>21</sup> More specifically, *-fie* and *-thie* represent a sub-type of such a process of extension, namely *tainting* (Jespersen 1922: 386–388), in which the suffix not only acquires phonological material from the attached word, but also a more specific meaning than the original suffix.

## 5 The formal and semantic properties of the *selfie* and *smoothie* families

As we have seen in our inventory, the neologisms based on *selfie* and *smoothie* differ in terms of both the formal and semantic relationship between the head word and the postposed element. In the case of compounds such as *stemsselfie* 'vote selfie' or *morgonsselfie* 'morning selfie', the meaning is compositional and transparent. However, for words made with a tainted suffix and blends the meaning isn't always obvious due to missing phonological material of the head word and/or the non-compositionality of the semantics.

<sup>19</sup> For an elaborate exploration of blends and related phenomena, see Fertig (2013: 66–67) and Lepic (2016).

<sup>20</sup> Such forms are called "libfixes" by Zwicky: <https://arnoldzwicky.org/2010/01/23/libfixes/>.

<sup>21</sup> For Jespersen, *secretion* is a slightly different phenomenon where a part of a word is reanalysed as having grammatical meaning, e.g. the reanalysis of *-n* as a plural suffix between Old and Middle English (Jespersen 1922: 384–385).

### 5.1 The *selfie* family

As we've established, *-fie* has a distinct meaning from its origins in *-ie*, but it also differs from *selfie*. When we analyse the neologisms ending in *-fie*, it becomes challenging to assign a very specific semantic value to the suffix. It would be safe to define *-fie* as 'a photograph', but the relationship between the suffix and the head word may be different in each case.

The starting point is the fully substantive micro-construction mentioned in section 2:

- (3)  $[[\text{self}]_i\text{-ie}]_j \leftrightarrow [\text{photograph that is taken by and features SEM}_i]_j$

However, in the neologisms, we encounter a range of related but subtly different constructions. In these cases 'self' is always the person *presenting* the photo, but not necessarily the one who *took* it:

- (4)  $[[\text{a}]_{\text{Ni}}\text{-fie}]_j \leftrightarrow [\text{photograph that is taken by and features self plus SEM}_i]_j$   
(e.g. *groupfie*)
- (5)  $[[\text{a}]_{\text{Xi}}\text{-fie}]_j \leftrightarrow [\text{photograph that features self plus SEM}_i]_j$   
(This construction applies to various usages of *selfie* neologisms where self is not the photographer.)
- (6)  $[[\text{a}]_{\text{Ni}}\text{-fie}]_j \leftrightarrow [\text{photograph that is taken by and features self at SEM}_i]_j$   
(e.g. *badfie* 'bathroom selfie')
- (7)  $[[\text{a}]_{\text{Xi}}\text{-fie}]_j \leftrightarrow [\text{photograph that is taken by and features self while performing SEM}_i]_j$   
(e.g. *stufie* 'study selfie')
- (8)  $[[\text{a}]_{\text{Ni}}\text{-fie}]_j \leftrightarrow [\text{photograph that is taken by self and features SEM}_i \text{ belonging to self}]_j$   
(e.g. *chantarellfie* 'picture of my Chanterelles')
- (9)  $[[\text{a}]_{\text{Ni}}\text{-fie}]_j \leftrightarrow [\text{photograph that is taken by self and features self reflected in SEM}_i]_j$   
(e.g. *balfie* 'Christmas ornament selfie')

### 5.2 The *smoothie* family

For *smoothie*, we encounter a similar situation, albeit one with less diversity in terms of derived constructions. Again, the basic construction of *smoothie* is as follows:

- (10)  $[[\text{smooth}]_i\text{-ie}]_j \leftrightarrow [\text{drink that has a SEM}_i \text{ mouth texture}]_j$

For the Swedish *-thie* coinages *groothie* and *semloothie*, as well as *Fruithie*, the construction is different:

- (11)  $[[\text{a}]_{\text{Ni}}\text{-thie}]_j \leftrightarrow [\text{smoothie that contains or tastes like SEM}_i]_j$

Finally, we may analyse the construction underlying *Oathie* as follows, where the semantic broadening from *-thie* ‘a type of smoothie, a healthy drink’ to ‘healthy food or drink’ is expressed formally:

$$(12) \quad [[a]_{Ni}\text{-thie}]_j \leftrightarrow [\text{healthy nutritional item that contains SEM}_i]_j$$

## 6 Conclusion

While differing in terms of actual productivity, we have shown that *selfie* and *smoothie* (in both Swedish and Dutch) have the potential to spawn neologisms based on a spectrum of word creation methods: compounding, blending, and (tainted) suffixation.

The fertility of *selfie* is primarily analogically driven by its form, and by the core meaning ‘photograph’. In other words, the tail ends of various blends and the suffix *-fie* indicate a type of photograph, but the exact relationship between them and the word they attach to depends on the context. In many cases, the full meaning of the *-fie* word is partly determined by the picture for which the *-fie* word serves as a caption. This symbiosis between linguistic meaning and graphic context adds an extra dimension to these particular word formations. In addition, in all cases the presence of a ‘self’ featuring in or related to the photo is heavily implied, even if the morpheme *self* is not used anymore.

In the case of the *smoothie* family, productivity is mostly driven by the need for creating original brand names. The lack of wider productivity may be explained by the smaller cultural relevance of the smoothie as opposed to the selfie. In both Swedish and Dutch, *selfie* was at least twice as frequent as *smoothie*.

Finally, we may note that the repeated coinage of words using the word creation processes described here has a jocular and playful character as well, as also pointed out by Lepic (2016: 104). Many forms appear to be created mainly or in part to be funny and/or original, rather than to fill serious gaps in the lexicon. While this kind of word creation rarely leads to particularly frequent or long-lasting neologisms, it reminds us that the role of creativity and playfulness in intentional language change needs to be considered. It is not unthinkable that by now commonplace words such as *selfie* and *smoothie* began their lives in such a humble fashion.

## References

- Bauer, Laurie. 2004. *A glossary of morphology*. Washington: Georgetown University Press.
- Booij, Geert. 2005. *The grammar of words. An introduction to linguistic morphology*. Oxford: Oxford University Press.
- Booij, Geert. 2010. *Construction morphology*. Oxford: Oxford University Press.
- Fertig, David. 2013. *Analogy and morphological change*. Edinburgh: Edinburgh University Press.

- Haspelmath, Martin. 1995. The growth of affixes in morphological reanalysis. *Yearbook of morphology 1994*. 1–29.
- Haspelmath, Martin. 2002. *Understanding morphology*. London: Arnold.
- Hüning, Matthias. 2000. Monica en andere gates. Het ontstaan van een morfologisch procédé. *Nederlandse Taalkunde* 5(2). 121–132.
- Jespersen, Otto. 1922. *Language. Its nature, development, and origin*. London: George Allen & Unwin.
- Lepic, Ryan. 2016. Lexical blends and lexical patterns in English and in American Sign Language. *Online Proceedings of the tenth Mediterranean Morphology Meeting MMM10 (Haifa) 7-10 September 2015*. 98–111.
- Mickwitz, Åsa. 2010. *Anpassning i språkkontakt. Morfologisk och ortografisk anpassning av engelska lånord i svenskan*. PhD Thesis, University of Helsingfors = *Nordica Helsingiensia* 21.
- Norde, Muriel & Kristel van Goethem. 2015. Emancipatie van affixen en affixöiden. Degrammaticalisatie of lexicalisatie? *Nederlandse Taalkunde* 20(1). 109–148.
- Shields Jr., Kenneth. 2001. On the origins of the English diminutive suffix *-y, -ie*. *Studia Anglica Posnaniensia* 36. 141–144.
- Språkkriktighetsboken*. 2005. Stockholm: Norstedts.
- Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik 2: Ord*. Stockholm: Norstedts.

## Chapter 35

# “Featurometry”

Benedikt Szmrecsanyi

University of Leuven

The paper applies a method to determine feature similarities based on an aggregate analysis of co-occurrence patterns to a dataset that catalogues the presence or absence of 76 morphosyntactic features from all domains of grammar in 46 international varieties of English. The analysis uncovers distributional dimensions that are largely language-external in nature; features do not appear to pattern in terms of structural properties.

### 1 Introduction

Classical dialectometry in the Séguy-Goebel-Nerbonne tradition calculates dialect distances based on a joint analysis of a large number of features; subsequently, dialect distances – or similarities, for that matter – are explored as a function of geographic space. The analysis of feature aggregates is instructive, but classical dialectometry has been criticized for not paying enough attention to how individual linguistic features are implicated in the large-scale relationships that dialectometry seeks to investigate. As a consequence, the recent literature has been exploring methods that focus not only on the big picture but that also highlight how individual features contribute to that picture. For example, Wieling & Nerbonne (2011) use bipartite spectral graph partitioning to simultaneously identify similarities between dialect varieties as well as their most distinctive linguistic features; Ruetten, Ehret & Szmrecsanyi (2016) utilize individual differences scaling for demonstrating how individual lexical features contribute to large-scale lexical relationships.

In this contribution I explore a more radical way to bring features to the fore in (dia)lectometry, drawing inspiration from a recent proposal coming forward from the formal-generative community for

“modeling and understanding the differences and similarities between linguistic constructions [...] based on their geographical distribution. In a nutshell, cluster orders [the constructions under study in the paper, BS] that occur in the same locations will be assumed to be more alike than those that have a different geographical distribution. Note that in this setup it is largely irrelevant whether

or not those dialect locations form a contiguous region. The locational data are merely used as binary variables that sketch a detailed empirical picture [...]” (Van Craenenbroeck 2014: 7)

In other words, the method calculates feature similarities based on an aggregate analysis of co-occurrence patterns in each of the dialect locations under analysis. The aim is to identify parameters that are assumed to fuel such co-occurrence patterns; geography is really only relevant to the extent that one needs to know which features co-occur in the same locations. Van Craenenbroeck (2014) refers to this perspective as “reverse dialectometry”, but the present author feels that this label may be a bit misleading: given that classical dialectometry is defined as investigating dialect relationships as a function of geographic space, “reverse dialectometry” would seem to promise to investigate geographic space as a function of dialect relationships. But as was explained above, geography really takes a back seat in Van Craenenbroeck (2014)’s proposal, which is why this contribution uses the label “featurometry” instead.

Van Craenenbroeck (2014) applies the method to a fairly specialized, syntax-oriented dataset covering verb clusters in Dutch dialects, and finds that it yields plausible results. My aim in this contribution is to evaluate the method from a more explicitly dialectologically and sociolinguistically oriented point of view. By way of a case study, I will analyze co-occurrence patterns in the morphosyntax survey coming with the *Handbook of Varieties of English* (Kortmann & Szmrecsanyi 2004), which catalogues the presence or absence of 76 morphosyntactic features from all domains of grammar in 46 international varieties of English. The question is this: can we identify salient dimensions (“parameters”) of variation using the method?

## 2 The dataset

I analyze the morphosyntax survey (<http://www.varieties.mouton-content.com/>) that accompanies the *Handbook of Varieties of English* (Kortmann & Szmrecsanyi 2004). This survey of non-standard English morphosyntax was conducted as follows: we compiled a catalogue of 76 features – essentially, the usual suspects in previous dialectological, variationist, and creolist research – and sent out this catalogue to the authors of the chapters in the morphosyntax volume of the *Handbook*. For each of these 76 features, the contributors were asked to specify into which of the following three categories the relevant feature falls in the relevant variety, or set of closely related varieties:

- A pervasive (possibly obligatory) or at least very frequent;
- B exists but a (possibly receding) feature used only rarely, at least not frequently;
- C does not exist or is not documented.

For the purposes of the present study, I lump the A and B ratings into an ‘attested’ category, while C counts as “not attested”. Kortmann & Szmrecsanyi (2004: 1142–1144) discuss the survey procedure, as well as the advantages and drawbacks of

the method, in considerable detail. Suffice it to say here that the survey covers 46 non-standard varieties of English. All seven anglophone world regions (British Isles, America, Caribbean, Australia, Pacific, Asia, Africa), as well as a fair mix of L1 varieties (such as e.g. Appalachian English), indigenized L2 varieties (such as e.g. Indian English), and pidgins/creoles (such as e.g. Tok Pisin), are included.

### 3 Statistical analysis

I will apply two analysis techniques to the dataset: multiple correspondence analysis (MCA), which is the technique used in Van Craenenbroeck (2014), and multidimensional scaling (MDS), which is widely used in the dialectometry literature. In each case, for the sake of simplicity attention will be restricted to the first two dimensions yielded by statistical analysis.

#### 3.1 Multidimensional scaling (MDS)

MDS (see Kruskal & Wish 1978) is a well-known dimension reduction technique that translates distances between objects (in our case, features) in high-dimensional space into a lower-dimensional representation. To determine distances, I use the well-known squared Euclidean distance measure, which calculates the distance between any two features as the number of varieties in which the two features do not co-occur. Based on these distances I perform classical MDS utilizing R’s `cmdscale()` function. The resulting MDS map is displayed in Figure 1.

In dimension 1, features with the highest negative scores, in the left half of the plot, include feature [74] (lack of inversion in main clause *yes/no* questions), [49] (*never* as preverbal past tense negator), and [10] (*me* instead of *I* in coordinate subjects). These features are known to be recurrent in varieties of English wherever they are spoken (see Kortmann & Szmrecsanyi 2004: Table 3). At the right end of Dimension 1, the top runners with positive scores include [33] (*after-Perfect*), [64] (relative particle *at*), and [63] (relative particle *as*). These features are very rare in varieties of English (see Kortmann & Szmrecsanyi 2004: Table 2). Dimension 1, therefore, sorts features according to how widespread/rare they are.

As for dimension 2, the features with the highest negative scores (lower half of plot) are [55] (existential/presentational *there’s*, *there is*, *there was* with plural subjects), [71] (*as what/than what* in comparative clauses), and [70] (unsplit *for to* in infinitival purpose clauses). We know that these features are fairly typical of British varieties of English (Kortmann & Szmrecsanyi 2004: Table 8); [55] and [71] are also known to be characteristic of L1 varieties more generally (Kortmann & Szmrecsanyi 2004: Table 23). The set of features with the highest positive scores in dimension 2 (top half of plot) includes [50] (*no* as preverbal negator), [40] (zero past tense forms of regular verbs), and [72] (serial verbs). These features, which tend to be reductive in nature, set apart English-based pidgin and creole languages from other varieties (Kortmann & Szmrecsanyi 2004: Table 25). Dimension 2 thus arranges features in

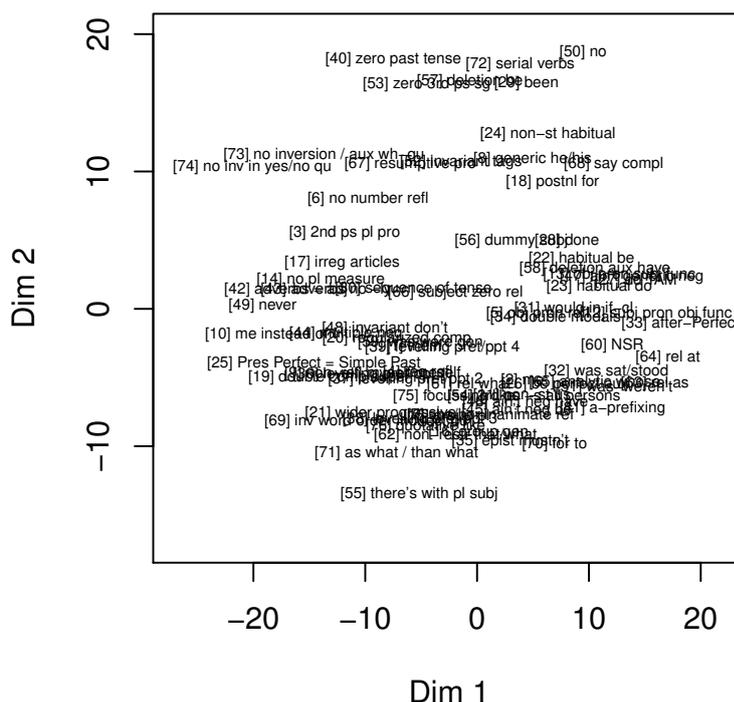


Figure 1: Multidimensional scaling map.

terms of how characteristic they are of pidgin and creole languages, as opposed to British-type L1 varieties of English.

By way of an interim summary, MDS picks up two largely language-externally defined dimensions of variation: widespreadness and pidgin-/creoleness (versus Britishness and/or L1-ness). The plot indicates that we find most variance in widespreadness among those features that are neither particularly characteristic of pidgins/creoles, nor of British and/or L1 varieties of English.

### 3.2 Multiple correspondence analysis (MCA)

MCA is a technique, similar in spirit to factor analysis, to examine how categorical variables (in our case: features) are associated with each other and to establish the extent to which they can be organized to yield common dimensions of variation. Unlike

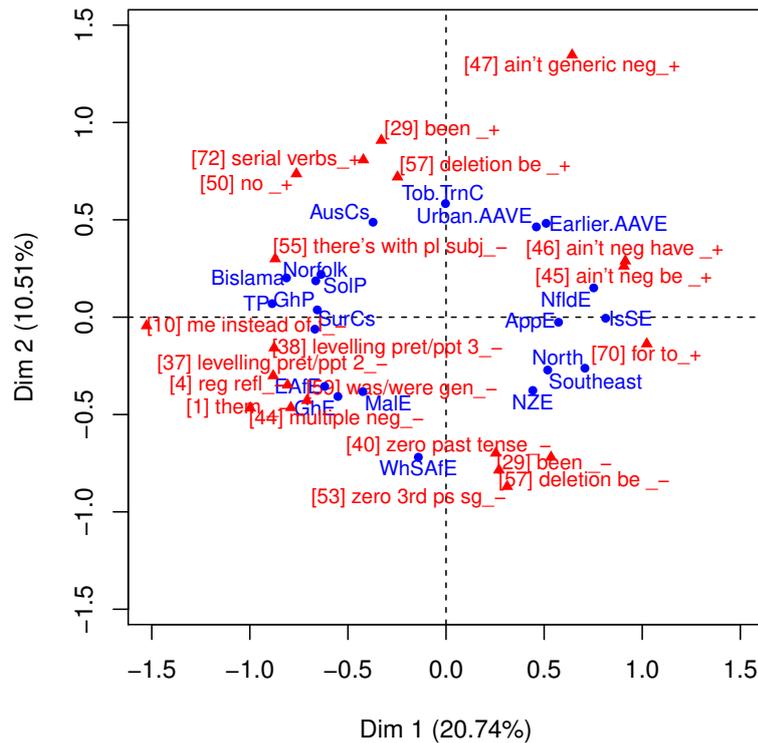


Figure 2: Multiple Correspondence Analysis. Display is limited to the 20 features and varieties that have the highest contribution on the dimensions. ‘+’ suffixed to a feature’s label indicates presence of the feature, ‘-’ indicates absence.

MDS, MCA also provides information about the behavior of individual observations (in our case, varieties): a particular variety will appear in the same part of the plot as the values of the features by which the variety is characterized (Levshina 2015: 375–376). The distance measure used in MCA is the chi-square distance measure. The following analysis was carried out using the `mca()` function in the `FactoMineR` package (Lê, Josse & Husson 2008). The resulting plot is displayed in Figure 2; note that for the sake of readability, the plot only displays the 20 features and 20 varieties that have the highest contribution on the dimensions (arguments `select = "contrib 20"` and `selectMod = "contrib 20"`).

Let us discuss Figure 2 by quadrant. In the upper left hand quadrant, we find features such as [72] (serial verbs) and [50] (*no* as preverbal negator), which as we saw

are characteristic of English-based pidgin and creole languages – and indeed, the varieties that the MCA plot identifies as particularly attracted to these features (e.g. Australian Creoles) are all pidgins and creoles. In the upper right hand quadrant three features are identified as particularly diagnostic: [47] (*ain't* as generic negator before a main verb), [46] (*ain't* as the negated form of *have*), and [45] (*ain't* as the negated form of *be*). The varieties located in this quadrant are all North American, and indeed we know that *ain't* is particularly characteristic of North American English (Kortmann & Szmrecsanyi 2004: Table 10). In the lower right hand quadrant MCA locates some British varieties (dialects in the North and in the Southeast of England) as well as New Zealand English, a variety that is known to be fairly close to British English, at least in terms of grammar. The features that MCA identifies as being distinctive for this quadrant include [70] (unsplit *for to* in infinitival purpose clauses); features [29] (past tense/anterior marker *been*) and [57] (deletion of *be*) are typically absent. This distributional pattern is typical of British varieties of English, according to the literature (Kortmann & Szmrecsanyi 2004: 1162–1165). In the lower left-hand quadrant, finally, we find primarily indigenized L2 varieties such as Malaysian English and Ghanaian English. The MCA plot suggests that these varieties are primarily characterized by the absence of features such as [1] (*them* instead of demonstrative *those*) and [44] (multiple negation).

So in the big picture, the MCA plot appears to identify as the most important dimension of variation (Dim 1) the contrast between native varieties (right) against pidgins/creoles and L2-varieties (left). The vertical dimension (Dim 2) is harder to interpret, but some exceptions notwithstanding seems to be capturing a language-externally defined contrast between orientation toward North American English (top) versus orientation toward British English (bottom).

## 4 Discussion and concluding remarks

This contribution has applied two “featurometrical” analysis techniques to a dataset documenting the distribution of 76 non-standard grammatical features in dozens of spoken varieties of English around the world. The analysis has revealed a number of dimensions that appear to fuel the distribution of non-standard grammatical features. According to multidimensional scaling (MDS), what counts is how widespread features are, and the extent to which they tend to be attested in pidgin and creole languages or not. Multiple correspondence analysis suggests that the distribution of features is primarily sensitive to the contrast between native varieties vis-à-vis pidgins/creoles and L2-varieties, and secondarily to the contrast between varieties that orient toward North American English as opposed to British English.

The dataset under study in this contribution is an extremely well-studied one, and it is fair to say that these contrasts and dimensions have not escaped notice in the literature (see e.g. Kortmann & Szmrecsanyi 2004; Szmrecsanyi & Kortmann 2009a,b). For example, the top biconditional implication identified by Szmrecsanyi & Kortmann (2009a: 223) is that the occurrence/non-occurrence of feature [45] (*ain't* as the negated form of *be*) is conditioned on the occurrence/nonoccurrence of feature [46]

(*ain*’t as the negated form of *have*), and vice versa; and this pattern comes through very clearly in the MCA plot in Figure 2. It is, of course, good to see that a new perspective yields results that do not contradict what we know to be true. But there should also be some sort of added bonus. In the case of the dataset analyzed here, the bonus primarily consists of the nice visualizations generated by MDS and, in particular, MCA: these are clearly superior to the somewhat dreary feature lists that e.g. Kortmann & Szmrecsanyi (2004) present.

As for more substantial insights, I note that the dimensions uncovered in this study are largely language-external in nature; features do not appear to pattern extensively in terms of their structural properties (an exception is MDS dimension 2, which can be said to arrange features in terms of how reductive they are). It is of course true that we know that e.g. creole languages are structurally different from non-creole languages (see, for example, McWhorter 2001), and so to the extent that features are diagnostic of creoles we may be *indirectly* dealing with structural issues, language-internally conditioned grammaticalization processes and such after all. But my point is that the patterns we are seeing are not *primarily* driven by inherent properties of the features under study, but – externally – by properties of the varieties in which they occur. Thus what Van Craenenbroeck (2014) calls “noise” (i.e. extra-grammatical aspects, as opposed to “the signal”, which is about structure) seems to be what really structures the dataset under study here. Observe also that the MCA plots in Van Craenenbroeck (2014) reveal more structure and are less cloud-like than the diagrams presented above. So is there really no structural signal in the survey under study in this paper? I caution in this connection that unlike Van Craenenbroeck (2014), the present study does not include any theory-inspired supplementary variables, which is why the analysis presented here is rather exploratory. Also, attention was restricted, for the sake of clarity, to the first two dimensions in MDS and MCA, but more structure may be hiding in higher-numbered dimensions. Let us also keep in mind that the features included in the morphosyntax survey are more “surfacy” and more of a mixed bag than the genuinely syntactic features studied in Van Craenenbroeck (2014). Finally, the fact that the morphosyntax survey covers a range of variety types – native vernaculars, indigenized L2 varieties, pidgin and creole languages – may overwhelm any structural signal potentially present in corners of the dataset, which is why future research is encouraged to conduct separate analyses for each variety type.

## Acknowledgments

I am grateful to Cora Pots and Jeroen van Craenenbroeck for helpful comments on an earlier draft. The usual disclaimers apply.

## Appendix: features covered in the survey

1. *them* instead of demonstrative *those* (e.g. *in them days ... one of them things ...*)
2. *me* instead of possessive *my* (e.g. *He’s me brother, I’ve lost me bike*)
3. special forms or phrases for the second person plural pronoun (e.g. *youse, y’all*)

Benedikt Szmrecsanyi

4. regularized reflexives-paradigm (e.g. *hisself, theirselves/theirself*)
5. object pronoun forms saving as base for reflexives (e.g. *meself*)
6. lack of number distinction in reflexives (e.g. plural *-self*)
7. *she/her* used for inanimate referents (e.g. *She was burning good* [about a house])
8. generic *he/his* for all genders (e.g. *My car, he's broken*)
9. *myself/meself* in a non-reflexive function (e.g. *my/me husband and myself*)
10. *me* instead of *I* in coordinate subjects (e.g. *Me and my brother/My brother and me were late for school*)
11. non-standard use of *us* (e.g. *Us George was a nice one*).
12. non-coordinated subject pronoun forms in object function (e.g. *You did get he out of bed*)
13. non-coordinated object pronoun forms in subject function (e.g. *Us say 'er's dry*)
14. absence of plural marking after measure nouns (e.g. *four pound, five year*)
15. group plurals (e.g. *That President has two Secretary of States*)
16. group genitives (e.g. *The man I met's girlfriend is a real beauty*)
17. irregular use of articles (e.g. *I had nice garden*)
18. postnominal *for*-phrases to express possession (e.g. *The house for me*)
19. double comparatives and superlatives (e.g. *That is so much more easier to follow*)
20. regularized comparison strategies (e.g. *He is the regularest guy*)
21. wider range of uses of the Progressive (e.g. *I'm liking this, What are you wanting*)
22. habitual *be* (e.g. *He be sick*)
23. habitual *do* (e.g. *He does catch fish pretty*)
24. non-standard habitual markers other than *be* and *do*
25. levelling of difference between Present Perfect and Simple Past (e.g. *Were you ever in London?*)
26. *be* as perfect auxiliary (e.g. *They're not left school yet*)
27. *do* as a tense and aspect marker (e.g. *This man what do own this*)
28. completive/perfect *done* (e.g. *He done go fishing, You donate what I has sent you?*)
29. past tense/anterior marker *been* (e.g. *I been cut the bread*)
30. loosening of sequence of tense rule (e.g. *I noticed the van I came in*)
31. would in *if*-clauses (e.g. *If I'd be you, ...*)
32. *was sat/stood* with progressive meaning (e.g. *when you're stood there*)
33. *after*-Perfect (e.g. *She's after selling the boat*)
34. double modals (e.g. *I tell you what we might should do*)
35. epistemic *mustn't* ('can't, it is concluded that ...not'; e.g. *This mustn't be true*)
36. levelling of preterite and past participle verb forms: regularization of irregular verb paradigms (e.g. *catch-catched-catched*)
37. levelling of preterite and past participle verb forms: unmarked forms (frequent with e.g. *give* and *run*)
38. levelling of preterite and past participle verb forms: past form replacing the participle (e.g. *He had went*)
39. levelling of preterite and past participle verb forms: participle replacing the past form (e.g. *He gone to Mary*)
40. zero past tense forms of regular verbs (e.g. *I walk for I walked*)
41. *a*-prefixing on *ing*-forms (e.g. *They wasn't a-doin' nothin' wrong*)
42. adverbs (other than degree modifiers) derived from adjectives lack *-ly* (e.g. *He treated her wrong*)
43. degree modifier adverbs lack *-ly* (e.g. *That's real good*)
44. multiple negation / negative concord (e.g. *He won't do no harm*)
45. *ain't* as the negated form of *be* (e.g. *They're all in there, ain't they?*)
46. *ain't* as the negated form of *have* (e.g. *I ain't had a look at them yet*)
47. *ain't* as generic negator before a main verb (e.g. *Something I ain't know about*)
48. invariant *don't* for all persons in the present tense (e.g. *He don't like me*)
49. *never* as preverbal past tense negator (e.g. *He never came* [= he didn't come])
50. *no* as preverbal negator (e.g. *me no iit brekfus*)
51. *was-weren't* split (e.g. *The boys was interested, but Mary weren't*)
52. invariant non-concord tags (e.g. *innit/in't it/isn't* in *They had them in their hair, innit?*)
53. invariant present tense forms due to zero marking for the third person singular (e.g. *So he show up and say, What's up?*)
54. invariant present tense forms due to generalization of third person *-s* to all persons (e.g. *I sees the house*)
55. existential / presentational *there's, there is, there was* with plural subjects (e.g. *There's two men waiting in the hall*)
56. variant forms of dummy subjects in existential clauses (e.g. *they, it*)
57. deletion of *be* (e.g. *She \_\_\_ smart*)
58. deletion of auxiliary *have* (e.g. *I \_\_\_ eaten my lunch*)
59. *was/were* generalization (e.g. *You were hungry but he were thirsty*)
60. Northern Subject Rule (e.g. *I sing vs. \*I sings, Birds sings, I sing and dances*)

61. relative particle *what* (e.g. *This is the man what painted my house*)
62. relative particle *that* or *what* in non-restrictive contexts (e.g. *My daughter, that/what lives in London,...*)
63. relative particle *as* (e.g. *He was a chap as got a living anyhow*)
64. relative particle *at* (e.g. *This is the man at painted my house*)
65. use of analytic *that his/thats, what his/whats, at's, as'* instead of *whose* (e.g. *The man what's wife has died*)
66. gapping or zero-relativization in subject position (e.g. *The man \_\_\_ lives there is a nice chap*)
67. resumptive / shadow pronouns (e.g. *This is the house which I painted it yesterday*)
68. *say*-based complementizers
69. inverted word order in indirect questions (e.g. *I'm wondering what are you gonna do?*)
70. unsplit *for to* in infinitival purpose clauses (e.g. *gutters for to drain the water away*)
71. *as what / than what* in comparative clauses (e.g. *It's harder than what you think it is*)
72. serial verbs (e.g. *give* meaning 'to, for', as in *Karibuk giv mi, 'Give the book to me'*)
73. lack of inversion / lack of auxiliaries in *wh*-questions (e.g. *What you doing?*)
74. lack of inversion in main clause *yes/no* questions (e.g. *You get the point?*)
75. *like* as a focussing device (e.g. *How did you get away with that like?*)
76. *like* as a quotative particle (e.g. *And she was like, what do you mean?*)

## References

- Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: morphological and syntactic variation in English. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 1142–1202. Berlin/New York: Mouton de Gruyter.
- Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional scaling*. Newbury Park, London, New Delhi: Sage Publications.
- Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 25(1). 1–18.
- Levshina, Natalia. 2015. *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic typology* 5(2/3). 125–166.
- Ruette, Tom, Katharina Ehret & Benedikt Szmrecsanyi. 2016. A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics* 21(1). 48–79. (13 April, 2016).
- Szmrecsanyi, Benedikt & Bernd Kortmann. 2009a. The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119(11). 1643–1663. (27 May, 2011).
- Szmrecsanyi, Benedikt & Bernd Kortmann. 2009b. Vernacular universals and angloversals in a typological perspective. In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, 33–53. London, New York: Routledge.
- van Craenenbroeck, Jeroen. 2014. The signal and the noise in Dutch verb clusters. A quantitative search for parameters. KU Leuven.
- Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language* 25(3). 700–715. (18 August, 2016).



## Chapter 36

# Bootstrapping a dependency parser for Maltese – a real-world test case

Jörg Tiedemann

University of Helsinki

Lonneke van der Plas

University of Malta

This paper evaluates the practicality of methods intended to bootstrap dependency parsers for new languages on a real-world test case: Maltese. Previous work has evaluated cross-lingual methods, such as annotation projection and model transfer, by proxy, i.e., by selecting target languages from the set of languages available in multilingual treebanks, because truly under-resourced languages do not have test sets available. As a result, experiments in previous work are often limited to closely related Indo-European languages, or lack real-world scenarios. At this exact point in time, Maltese is an excellent candidate to evaluate the usefulness of the cross-lingual methods proposed in previous work: treebank development is in progress, no syntactic parsers are available, but certain NLP tools and corpora have recently become available. Maltese belongs to the branch of Semitic languages that is different from the languages for which NLP resources are most widely available. However, it has been under the influence of several Indo-European languages due to its turbulent history. It is therefore an even more interesting test case for exploring multi-source projection and the contribution of various languages with respect to their linguistic influence on the Maltese language.

## 1 Introduction

State-of-the-art methods for inducing NLP tools, such as part-of-speech (PoS) taggers and dependency parsers, rely on large quantities of hand-annotated data. For most of the world's languages these annotations are not available, because their creation is a costly and time-consuming enterprise. Cross-lingual learning methods try to bootstrap NLP tools for low-resource languages despite the lack of annotated resources in those languages. Early work focused on annotation projection or *data transfer* (Hwa et al. 2005; Yarowsky, Ngai & Wicentowski 2001). In that approach, annotations are

projected from the well-resourced source languages to the low-resource target language using parallel corpora. Secondly, in *model transfer*, models are trained on the annotated source language and applied to the target language. Without adaptation, this only works reasonably well for closely related languages (Agić et al. 2014). Typically, models need to be delexicalised unless there is a substantial lexical overlap between source and target language. Due to the availability of harmonised PoS annotations (Petrov, Das & McDonald 2012), shared PoS features across languages can be used by delexicalised models (McDonald et al. 2013), which can be further improved using cross-lingual word clusters or target language adaptation. Guo et al. (2015) extend delexicalized transfer models with cross-lingual distributed representations to include lexical knowledge. They apply alignment-based projection and canonical correlation analysis (CCA) to map monolingual distributed word representations. Lastly, translating treebanks is a cross-lingual method proposed by Tiedemann, Agić & Nivre (2014). A given source language treebank is translated by an existing MT system, for example a statistical MT model trained on parallel data. Annotations are projected from source to synthetic target language sentences with the same techniques as before with the main benefit that they come from manually verified annotations instead of automatically parsed data sets.

An overview of common techniques of cross-lingual methods is presented in (Tiedemann & Agić 2016). We base our work on a similar setup but move away from proxy-based evaluation to a real-world scenario. To the best of our knowledge, no other publications focus their work on truly under-resourced languages. Agić, Hovy & Søgaard (2015) come closest to this idea by focusing on languages with small treebanks using resources such as the Bible and the Watchtower corpus that cover many low-density languages. In contrast to previous work, we basically start from scratch, having a small treebank for development purposes with the goal of making the best use out of the tools and resources available for our target language, Maltese.

## 1.1 Maltese

Maltese is a language spoken by the people of the Maltese archipelago that lie in the Mediterranean Sea some 80 kilometres south of Sicily as well as several Maltese communities abroad, totalling around 450,000 speakers. The Maltese language had a very interesting development in that it was under the influence of many languages from different language families in the course of Malta's history and has therefore been classified as a mixed language (Aquilina 1958). Maltese is assumed to originate from an Arabic dialect, brought to Malta by the Arabic conquerors in 870, that is close to the dialects spoken by the inhabitants of Tunisia. Malta had very strong links with Sicily and Italy and the Christian world in general and lost contact with the Arabic community in the 19th century. The language shows significant signs of Romancisation. Not only the lexicon, which has over 55% elements of Italo-Romance origin, but also morphology, syntax, and semantics are influenced by Italo-Romance languages (Stolz 2011). English started to have an impact on the language during the 19th century and more notably during the 20th century, largely due to the bilingual situation. After Malta's independence in 1964, Maltese became an official language

and since its membership in the EU in 2006 it is also an official language of the union.

The previous paragraph underpins the statement that bootstrapping a dependency parser for Maltese is an excellent real-world test case for cross-lingual methods. The development of Maltese NLP tools and corpora happens to be at the exact point in time, where cross-lingual methods would be very useful. The first steps towards treebanking have been made: a few hundred sentences have been annotated. Just enough to evaluate the crosslingual methods, but not enough to train a good lexicalised parser. The Maltese NLP community is joining forces to develop NLP tools for Maltese but is behind the other European languages. As a result, some NLP tools (a PoS tagger and a chunker) have become available over the past couple of years and a large monolingual corpus has been released in 2016. An electronic dictionary with morphological information and translations into English has been released very recently as well. Fortunately for us, Malta’s membership in the EU has resulted in the availability of reasonably large quantities of parallel data. Furthermore, the fact that Maltese is strongly influenced by a variety of languages and cultures makes it a fascinating case for multi-source transfer methods.

## 2 Cross-lingual parsing

In order to answer the question concerning the practicality of cross-lingual methods in a real-world scenario, we experimented with several flavours of cross-lingual parsing approaches that have proven successful in previous work. We will briefly introduce the approaches below.

### 2.1 Annotation projection

In this approach, we apply the heuristics proposed by Hwa et al. (2005) that follow the direct correspondence assumption to make it possible to map annotations from one language to another through automatic word alignment in bitexts. The projection rules ensure the creation of valid tree structures in the target, which is necessary for the training procedures we apply. To reduce the negative influence of dummy nodes, we add the improvements proposed by Tiedemann & Agić (2016). In particular, we rely on the removal of dummy leaf nodes and the conflation of unary productions in the parse tree that involve dummy nodes. Furthermore, we remove all sentences that include any dummy node or dummy relation in their parse tree after projection and the modifications mentioned above.

For each source language we then use 40,000 sentences with projected annotations to train target language parsers. We also introduce additional features to the projected data by looking up morphological features in a monolingual dictionary. PoS information is used for some basic disambiguation.

## 2.2 Model transfer

The simplest transfer approach is to use delexicalised models that rely entirely on universal PoS tags. These baseline models work surprisingly well on closely-related languages but have a lot of short-comings due to the over-simplification of their feature models and their strong structural correspondence assumptions. An interesting option for target language adaptation is relexicalization by means of self-training. The simple idea is to apply the delexicalized model to monolingual target language data to create automatically annotated data for training fully lexicalized parsing models.

## 2.3 Translating treebanks

A third cross-lingual parsing technique proposed by Tiedemann, Agić & Nivre (2014) applies machine translation to create synthetic training data with projected annotation from the original treebank. This approach has been shown to be quite successful and often better than annotation projection. One of the main advantages is the use of manually verified source language annotation instead of noisy parsed out-of-domain data.

The biggest disadvantage is, of course, the lack of translation quality. Another critical problem is the requirement of sufficient training data for creating MT models. Typically, we cannot expect to have sufficient parallel data available to train statistical MT models for low-resource languages. However, the situation is different for Maltese due to its status in the EU.

For the translation approach, we follow the basic setup of our previous work and train standard phrase-based SMT models for all language pairs using a standard pipeline that includes word alignment, phrase extraction and phrase scoring.<sup>1</sup> Word alignments are already available from the annotation projection experiments and use the same output of the symmetrised alignments produced by efmara and its fertility-enhanced HMM model. Phrase extraction and phrase scoring use standard settings of the Moses pipeline (Koehn et al. 2007). The language model is estimated using kenlm (Heafield et al. 2013) with an order of 5 and modified Kneyser-Ney smoothing from the entire Maltese corpus described below.

## 3 Tools and resources

The **treebank** that we use as our test set was created by Slavomír Čěplö from Charles University in Prague. This corpus is still under heavy development. At the time of writing, 371 sentences have been tokenised and manually annotated with PoS tags and Universal Dependency (UD) relations by one annotator only. It contains sentences from the following domains: journalistic (239 sentences), short stories (14 sentences), encyclopedic and instructional (118 sentences).

<sup>1</sup> We skip tuning in our current experiments because there are no suitable development sets for Maltese. Nevertheless, standard weights are usually a good initial guess and the scores in our experiments suggest that the models produce reasonable output for the task at hand.

The **Korpus Malti v3.0**<sup>2</sup> is the latest achievement in an ongoing project of gathering digital resources for Maltese (Gatt & Ċéplö 2013). It contains 250 million tokens in various genres. It is tagged with the Maltese tagger described below, with an accuracy of around 97%. In addition, it includes partial lemma information and morphological analysis. We use the corpus for our relexicalisation experiments and for language modeling in the translation approach.

**Ġabra** is a free, open lexicon for Maltese, built by collecting various different lexical resources into one common database. Ġabra was originally built in 2013 as part of a master’s thesis project by Camilleri (2013) While it is not yet a complete dictionary for Maltese, recent efforts resulted in an on-line dictionary with 15,259 entries and 4,514,367 inflectional word forms. Many of these are linked by root, include translations in English, and are marked for various morphological features.

**Tagging Maltese** was carried out using the SVMTool (Giménez & Màrquez 2004), trained on a manually tagged corpus of ca. 28,000 tokens. The Maltese Tagset v3.0<sup>3</sup> was developed by Slavomír Ċéplö and Albert Gatt.

The **parser** we apply in our experiments is a graph-based model implemented in the mate tools (Bohnet 2010) and we use version 3.61.<sup>4</sup> For training the source language models we used the treebanks provided by the Universal Dependencies (UD) project version 1.3.<sup>5</sup> Note that we reduce dependency relations to universal categories to make it possible to transfer labels across languages.

### 3.1 Parallel data sets

Our two main approaches, annotation projection and treebank translation, heavily rely on parallel data sets. The DGT translation memory (Steinberger et al. 2012) can be used to produce an aligned multilingual corpus of the European Union’s legislative documents (Acquis Communautaire) in 24 EU languages. This memory contains most, although not all, of the documents which make up the Acquis Communautaire, as well as some other documents which are not part of the Acquis.

In our experiments we apply a subset of 19 languages for which we have sufficient data for all languages in parallel including Maltese. The final corpus we use for our experiments comprises in the end about 1.27 million truly parallel sentences with roughly 19 to 26 million tokens per language.

We tokenized Maltese using an in-house tokenizer that is compatible with the treebank test set and applied UDPipe<sup>6</sup> (Straka, Hajič & Straková 2016) for tokenising the other 18 languages according to the standards of the universal dependency treebanks. Thereafter, we aligned all languages with Maltese on the word level using efmara<sup>7</sup>, an efficient implementation of fertility-based alignment models based on Gibbs sam-

<sup>2</sup> The corpus along with many other electronic resources for Maltese can be retrieved from <http://mlrs.research.um.edu.mt>.

<sup>3</sup> <http://mlrs.research.um.edu.mt/resources/malti03/tagset30.html>.

<sup>4</sup> <https://code.google.com/archive/p/mate-tools/downloads>.

<sup>5</sup> <http://universaldependencies.org/introduction.html>.

<sup>6</sup> <http://ufal.mff.cuni.cz/udpipe>.

<sup>7</sup> <https://github.com/robertostling/efmaral>.

pling with a Bayesian extension (Östling & Tiedemann 2016). The word alignments are then symmetrised using the intersection and the popular grow-diag-final-and heuristics. Both of these symmetrized alignments are facilitated in our annotation projection experiments according to the procedures proposed by Tiedemann (2015).

## 4 Results and discussions

In Table 1 we see the results of the three main methods presented in this paper. Scores are given for the Maltese test set described above with predicted PoS labels for a realistic estimation of performance. We can see that models trained on the translated treebanks are comparable but often better than the annotation projection results. Using re-lexicalisation of delexicalized models works surprisingly well for source languages like Spanish and Italian with attachment scores that are quite close to the corresponding annotation projection experiments. But in general, the scores are significantly lower. Note that the purely delexicalized models perform even worse.

Table 1: Results for the three methods and the 18 source languages.

src	Projection		Relexicalisation		Translation	
	LAS	UAS	LAS	UAS	LAS	UAS
bg	54.94	66.24	46.56	59.52	54.86	66.25
cs	53.54	64.44	44.58	56.04	53.73	65.21
da	52.87	63.77	41.43	51.56	51.54	62.31
de	54.33	65.05	40.30	49.94	55.58	65.45
el	36.79	53.02	34.22	44.01	37.87	56.27
en	59.39	<b>69.53</b>	51.11	62.14	59.62	68.88
es	<b>59.78</b>	69.41	55.54	<b>65.88</b>	<b>60.50</b>	<b>70.32</b>
et	35.02	48.42	28.48	44.97	35.82	52.53
fi	37.14	50.49	27.61	41.73	37.09	53.02
fr	57.73	67.64	53.09	63.46	58.70	68.65
hu	49.30	59.36	28.70	40.85	41.10	52.15
it	57.70	66.74	<b>56.04</b>	65.11	60.35	68.80
nl	51.23	59.71	41.84	49.89	51.38	60.48
pl	51.78	63.33	44.01	53.98	52.09	62.58
pt	54.04	63.74	50.34	58.85	55.20	65.17
ro	55.80	65.72	50.56	61.29	56.75	67.43
sl	58.46	67.59	45.39	55.34	57.90	67.58
sv	53.19	62.87	40.77	53.41	51.62	61.70

Is the fact that Maltese is a mixed language reflected in higher performances when transferring from languages that are known to have had an effect on the Maltese language? We can only give tentative answers due to the lack of in-depth analyses and comparable parallel data for Arabic. Nevertheless, the best languages for cross-

lingual learning are Italian and Spanish, two Romance languages, which is expected due to strong Romancisation of Maltese. A second place is taken by English. We obtain very good scores for projection and translation, but the re-lexicalised models lags behind. Linguistic literature acknowledges the influence of English, but states that it is more recent, and therefore less embedded in the language. This could explain why the re-lexicalised model, that relies for the larger part on the syntactic structure of the source language, falls behind. Looking at the results it seems possible to conduct comparative linguistic studies based on the success of cross-lingual learning. This would be an interesting avenue to explore in future work.

Another interesting question is whether multi-source models can be used to overcome individual weaknesses of the projected data sets. The mixed nature of Maltese may support a combination of source languages in particular. Table 2 shows that a simple concatenation of data from the several languages works surprisingly well. Unexpectedly, the linguistically biased combination of Romance languages and English does not lead to any gains over the model that uses all data sets. Another unbiased approach of selecting all source languages for which the supervised source language parsers reach a level of at least 80% LAS leads to only modest improvements over the combination of all languages, which further demonstrates the robustness of the simple multi-source approach. As a final step, we also tested the inclusion of inflectional features and lemmas coming from the lexical database of Maltese. To our disappointment, this leads to only minor improvements in LAS and even a slight drop in UAS. This suggests that adding lexical information without contextual disambiguation provides only little help but coverage issues may also be good reason for the failure of this approach.

Table 2: Multi-source projection models and combinations of methods.

Method	languages	LAS	UAS
Projection	all languages	62.51	71.54
Projection	en es fr it pt ro	62.52	71.28
Projection	bg cs en es it sl	62.77	<b>71.80</b>
Projection + inflectional info	bg cs en es it sl	<b>63.03</b>	71.54

#### 4.1 Measuring the practicality of cross-lingual parsing

The numbers in the previous paragraphs, albeit interesting from a research perspective, have little practical value. In order to compare the merits of cross-lingual methods, we determined the amount of manually annotated data needed to reach levels of performance that are equal to those stemming from cross-lingual methods. We, therefore, split our small test set into tiny training data of various sizes while using the remaining examples for testing.<sup>8</sup> Figure 1 shows the learning curve in our

<sup>8</sup> In this setup, test data is always changing depending on how much of the data we reserve for training. Hence, these scores are not completely comparable but the general trends should still be trustworthy

procedure.

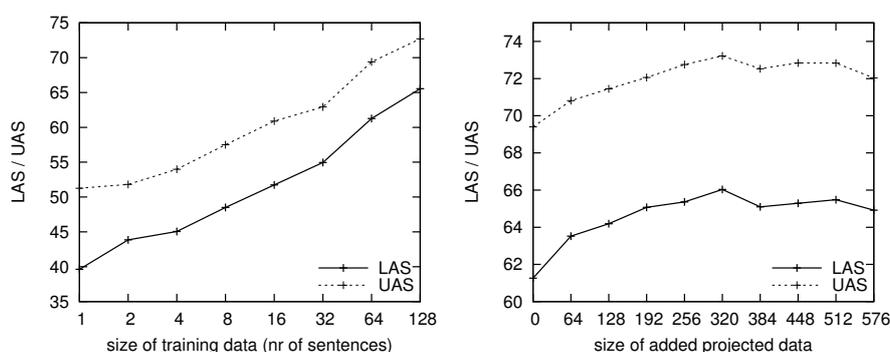


Figure 1: Figures showing a) the learning curve when training on manual data and b) the effect of adding projected data to a small amount of manually annotated data (64 sentences).

We can see that with as little as 64 training examples, we achieve a performance that is on par or above the best cross-lingual models discussed earlier. This comparison puts work on cross-lingual parsing into perspective. Most of the achievements presented in previous research are comparable to the results presented here but similar investigations on learning curves for tiny training data sets is usually not presented. The result above seems to strongly suggest that investing in annotation is a much wiser decision than spending time on tweaking a transfer model. A caveat may be that no expert can easily be found for many languages of the world and transfer models are still a valid choice for quickly building systems for a large number of languages. But the question remains whether this is really useful or not.

Another relevant question is whether hand-annotated data can successfully be combined with projected data to bootstrap models with very scarce resources. To study this, we ran another experiment in which we added small numbers of projected parse trees to a tiny treebank of 64 sentences for training parsing models that we then tested on the remaining test cases (with predicted PoS labels). The projected trees come from a multi-source model that we have re-trained on automatically parsed monolingual data from the Korpus Malti.

In this setup we use a simple concatenation strategy again and multiply the original treebank to match the size of the added noisy projected data. We can see in Figure 1 that the model trained on this augmented data indeed increases until a certain level of noise is reached that causes degradation of the parser. The improvements are still modest but it shows that there is some potential for such combined models. And given the small amount of projected trees necessary, there is room for filtering approaches to select high-quality projections.

---

to some extent.

## 5 Conclusions

We conducted a large number of experiments on dependency parsing to evaluate the practicality of cross-lingual learning for the real-world test case of Maltese, one of the official European languages, whose NLP tools and resources are under development. We leveraged all available resources in this realistic scenario and came to the following conclusions: firstly, the practicality of cross-lingual parsing is very much dependent on the current situation of the resources in the target language. Despite encouraging results with model and data transfer, cross-lingual parsing still lags far behind fully supervised models. For the scenario under discussion, small amounts of target language training data are available and these outperform state-of-the-art cross-lingual models. In such a setting, a possible use of data transfer may be the combination with scarce manually annotated data sets to bootstrap treebanks and parsers. A scenario in which cross-lingual methods could be of practical use is that of a large number of languages for which no linguistic resources are available at all. However, the practical use of such rough models still remains to be proven. In addition to these observations regarding the practicality of cross-lingual parsing, we reflected on the possibility of using cross-lingual learning for comparative explorations regarding the similarity between languages on specific linguistic levels such as the syntactic level. We leave the large-scale, detailed analyses needed for such an endeavor to future work.

## Acknowledgements

We would like to thank Slavomír Čéplö for making the manually annotated Maltese data available to us, as well as Albert Gatt for giving us access to the latest NLP tools and resources for Maltese.

## References

- Agić, Željko, Dirk Hovy & Anders Søgaard. 2015. If all you have is a bit of the Bible: learning POS taggers for truly low-resource languages. In *Proceedings of ACL*, 268–272.
- Agić, Željko, Jörg Tiedemann, Danijela Merkle, Simon Krek, Kaja Dobrovoljc & Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of LT4CloseLang*, 13–24.
- Aquilina, Joseph. 1958. Maltese, a mixed language. *Journal of Semitic Studies* 3. 58–79.
- Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*, 89–97.
- Camilleri, John. 2013. *A computational grammar and lexicon for Maltese*. Sweden: Chalmers University of Technology, Gothenburg MA thesis.
- Gatt, Albert & Slavomír Čéplö. 2013. Digital corpora and other electronic resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*. Lancaster, UK.

- Giménez, Jesús & Lluís Màrquez. 2004. SVMTool: a general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th LREC*. Lisbon, Portugal.
- Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang & Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *In proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark & Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, 690–696.
- Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas & Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering* 11(3). 311–325.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL*, 177–180.
- McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, 92–97.
- Östling, Robert & Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics (PBML)* (106). 125–146. <http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf>.
- Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, 2089–2096.
- Steinberger, Ralf, Andreas Eisele, Szymon Kłoczek, Spyridon Pilos & Patrick Schlüter. 2012. DGT-TM: a freely available translation memory in 22 languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Stolz, Thomas. 2011. Maltese. In *The languages and linguistics of Europe: a comprehensive guide*, 241–256. De Gruyter Mouton.
- Straka, Milan, Jan Hajič & Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).
- Tiedemann, Jörg. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of NoDaLiDa*.
- Tiedemann, Jörg & Željko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* 55. 209–248.

- Tiedemann, Jörg, Željko Agić & Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of CoNLL*, 130–140.
- Yarowsky, David, Grace Ngai & Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, 1–8.



## Chapter 37

# Identifying dialect regions from syntactic data

Erik Tjong Kim Sang

Meertens Institute Amsterdam

The *Syntactic atlas of Dutch dialects* (SAND) is a database of syntactic features observed in the language spoken by people from different dialect regions in The Netherlands and Flanders. We would like to know how specific syntactic features are for the different dialects. For this purpose we try to generate dialect boundaries from the syntactic data only. We show that a plausible binary division of the dialect can successfully be derived but that is more difficult to divide the data in three or more regions. We build on earlier work by Nerbonne, Heeringa & Kleiweg (1999), who performed this task for phonetic data, and on work by Spruit (2008), who also attempted to identify dialect regions from syntactic data.

## 1 Introduction

One of my early memories of the research work of John Nerbonne goes back to 1999, to the yearly meeting of the European Chapter of the Association of Computational Linguistics, held in Bergen, Norway. At that meeting John presented a paper, joint work with colleagues Wilbert Heeringa and Peter Kleiweg, in which they outlined how gradual differences between dialects could be derived and visualized from dialect data using computational methods (Nerbonne, Heeringa & Kleiweg 1999). Their visualization, a beautiful color map displaying gradual transitions between dialects as diverse as West Low German, Limburgish and West-Flemish, was the first color page ever in a proceedings of a major computational linguistics conference.

The paper and its followup work has inspired much other work, such as the Gabmap website in Groningen (Nerbonne et al. 2011), which enables researchers to visualize their own dialect data, and the project Maps and Grammar in Amsterdam (Barbiers et al. 2008), which examines the relation between linguistic properties and their geographic distribution. It is in the latter project that the current paper is based.

One of the research questions of the Maps and Grammar project is central to this paper. It is inspired by the data set we have available in the project: the *Syntac-*

*Erik Tjong Kim Sang*

*tic atlas of the Dutch dialects.* Unlike John Nerbonne and his colleagues, who used pronunciation data, we exclusively want to use this syntactic data set for deriving dialect boundaries. Therefore our research question is: *can reasonable dialect boundaries automatically be derived from syntactic linguistic data only?* In order to answer this question, we analyze our research data, compute linguistic distances between geographic locations and divide the locations in groups based on the distances.

After this introduction, we will discuss related work in section two. In the third section we will present our data while section four outlines our methods for analyzing and clustering the data. In the same section we present and discuss some results. We conclude in section five.

## 2 Related work

Nerbonne, Heeringa & Kleiweg (1999) showed how mathematical methods such as the Levenshtein distance, clustering and multidimensional scaling can be used for comparing word pronunciations, computing Dutch dialect distances and visualizing the dialects on a map. Similar work has been done by Goebel (2010) for French dialects. Spruit (2008) was the first to apply clustering and multidimensional scaling to syntactic dialect data, from the Dutch SAND database (Barbiers et al. 2008).

Like with data sets from other fields, a challenge in working with linguistic data is how to deal with missing data. Spruit (2008) replaced all missing data of the SAND with the zero values but as a result of this underspecified dialects may appear more similar than they really are. van Craenenbroeck (2014) adopted a similar strategy for locations visited by the SAND team and used mathematical methods for estimating data values in locations skipped by the team. Tjong Kim Sang (2015) proposed to keep all unknown values in the data and to use a new distance function for comparing the values. We will adopt this method when dealing with missing data (see section 4).

## 3 Data

We use linguistic data from the Syntactic Atlas for Dutch Dialects (SAND) (Barbiers et al. 2008). This data is based on interviews with people from 267 locations in The Netherlands and Flanders, the Dutch-speaking part of Belgium. The interviews have been transcribed and have been checked for the presence of predefined syntactic variables. We use the subset of 220 syntactic variables which are available from the digital version of the atlas: DynaSAND (Barbiers 2006).

A problem of the data set is the limited availability of negative values. The data was developed as a source for dialect maps, which means that the focus in the collection process was on finding locations where syntactic features were present and not on marking locations where they were absent. Therefore only 17% of the maps in the atlas contain explicit negative data (Tjong Kim Sang 2014).

For data modeling, different data values are required in order to be able to distinguish separate groups. For this purpose, earlier studies that used the SAND data

have used the evidence of absence assumption: if a syntactic feature or value was not observed at a location then it was assumed that the variable did not occur in the local dialect (Spruit 2008; van Craenenbroeck 2014). While this assumption might be perfectly valid in other contexts, its application here can be challenged. The interviews were held by different interviewers which asked different questions. So data feature absence might also be caused by the interviewers rather than only by the local dialect. Therefore we did not use the evidence of absence assumption and kept all unknown feature values in our data (marked as ?).

In the SAND data, most (81%) of the syntactic features are not binary but contain many values (Tjong Kim Sang 2014). An example of this is verb order in phrases with three verbs (A, B, C), which may contain up to six different values (ABC, ACB, BAC, BCA, CAB, CBA). Previous studies that used the data have represented such features as sets of binary values (Spruit 2008; van Craenenbroeck 2014). For example, six word order patterns are represented as six binary values. This causes such features to be represented in the model several times which increases their importance to the model. While the best weight of each feature is unknown, we believe it is unfair to give certain syntactic features more weight because of their shape. Therefore we use each feature only once in the model and use combined feature values when local dialects allow more a feature to have more than one value.

## 4 Method and results

We used *k*-means clustering (Manning & Schütze 1999) for dividing the 267 locations in separate regions. For this purpose, it was necessary to define what the distance between two feature values was. This was nontrivial for our data, since they contain basic feature values, sets of basic feature values and unknown values. We define the distance between two basic values as 0 when they are equal and 1 when they are different. The distance between sets of values is 0 when the sets share a common value and 1 otherwise. The distance between an unknown value and any other value (including unknown) is defined as 0.5.

Next, we define the distance between two locations as the sum of the distances between their syntactic feature values. We used these settings for the *k*-means algorithm to compute the best two regions starting from two randomly chosen region centers. This experiment was repeated 100 times and each time the algorithm generated the same two regions with the locations Erica in Drenthe and Bever in East Flanders as centers. The resulting map can be found in Figure 1 (Left).

The clustering algorithm has divided the Dutch-speaking community along the national border: The Netherlands vs. Flanders (left part of Figure 1). Only in the southwest and in south-east the linguistic boundary did not follow the national border. In the southwest, three Dutch locations were classified as Flemish: Oostburg, Hoek and Hulst (see the bottom right of Figure 1). All three are part of the region Zeelandic Flanders which is separated by the rest of The Netherlands by water. Until 2003, when a tunnel connection was opened, the only way to reach the region was either by boat or by a detour over land via Belgium. For this reason, the region has strong

ties with the neighboring Flanders. Dialect researchers Daan and Blok, classified Oostburg and Hoek among the Dutch dialects and Hulst among the Flemish dialects (Daan & Blok 1969)

In the southeast, 13 Flemish cities Bree, Eigenbilzen, Genk, Grote-Spouwen, Hamont, Hasselt, Houthalen, Jeuk, Lauw, Maaseik, Opglabbeek, Stokkem and Tongeren were classified as Dutch. This is not surprising: the dialect spoken in these cities is more common to the dialect of the neighboring Dutch province than to the neighboring Flemish region. The dialect map of Daan & Blok (1969) classified all of Limburgish dialects as one big group, including four locations which assigned to the Flemish group by  $k$ -means: Borgloon, Eksel, Lummen and Sint-Truiden (see the top right part of Figure 1).

Next we applied the  $k$ -means algorithm for dividing the Dutch language in three regions. We ran the algorithm 100 times from sets of three randomly chosen initial locations as centers of the regions, but this time 21 different region divisions were suggested. Most of them had one (43%) or two (54%) of the region centers in common with the two-region division of Figure 1. However, this time the proposed regions seemed less sensible. For example, the most frequent set (19%) with region centers Bevere, Boskoop and Erica, split Flanders in two parts but allocated part of The Netherlands to one of the parts (see Figure 2, left part). Only three of the 21 suggestions corresponded with known dialect borders to some extent: one which identified Frisian, be it with some unrelated locations (Figure 2, center), one which recognized French Flanders, with one additional location (Figure 2, right) and one in which East Flanders was separated from West Flanders.

$k$ -means did not work well for dividing the linguistic space in three regions. The algorithm did not converge to a single configuration and few of the proposed sets were reasonable. This did not improve when we attempted to divide the space in four regions. In 100 runs, the algorithm generated 56 different region sets. The most frequent of these (11%) was an extension of the most frequent set of three regions (Figure 2, left). None of the regions were complete and Belgian Limburg contained a section of four adjacent nodes which belonged to four different regions. None of the four-region sets that we inspected was very good and we conclude from this that for our data set  $k$ -means does not work very well in generating more than two regions.

## 5 Concluding remarks

We have applied  $k$ -means clustering to a set of syntactic dialect data in order to find out if it was possible to derive reasonable dialect boundaries. For the case of dividing the data set in two regions, this worked well. In 100 runs, the clustering algorithm always converged to the same two regions without any gaps or separate islands. This is impressive because no geometric data was used during the classification process. The syntactic space which was used for classification, is different from geometric space. The location of the region centers is evidence of this: they are by no means the geographic mean points of their region (Figure 1, left, the blank locations Bevere for Flanders in the green area and Erica for The Netherlands in the red area).

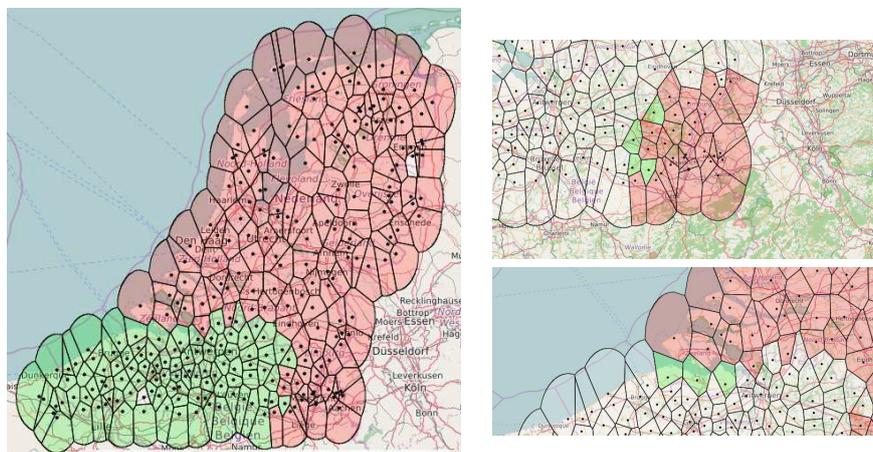


Figure 1: **Left:** a map with the two main Dutch-speaking regions identified by the  $k$ -means algorithm from the syntactic SAND data. **Right top:** the transnational area where Limburgish is spoken according to Daan & Blok (1969). **Right bottom:** the locations Oostburg, Hoek and Hulst in Zeelandic Flanders (green) are the only three Dutch locations classified as Flemish by  $k$ -means.

However,  $k$ -means performed less well for generating more than two regions. In the three-region case, it suggested 21 different region configurations, a few of which made some sense but these were not the most frequently proposed ones. None of the proposed sets of four regions that we checked, was reasonable. However, the few plausible cases for the three-region case (Figure 2, center and right) show that the data set contains useful information for automatic dialect boundary detection. We just might not have found the best algorithms to use this information.

There are different directions for future study. First, we would like to explore an idea of Heeringa (2004) for dealing with missing data: to use average scores as distances so that missing data values can safely be ignored. Next we would like to experiment with alternative hard clustering algorithms. We have already performed initial experiments with Expectation Maximization (Fraley et al. 2012), which generate similar results for the two-region case and suggest that there is even more useful information for this task in the data than we found with  $k$ -means clustering. We also would like to go back from the detected regions to the syntactic variables to check if certain dialect boundaries are predicted by individual syntactic variables. Some dialects may not be separated by a sharp boundary but by a fuzzy boundary, like on the map of Daan & Blok (1969). It would be interesting to see if we can identify such transition zones with computational measures. And finally, it would be interesting to apply these clustering methods to mixed data, for example a data set which included both syntactic and phonological data.

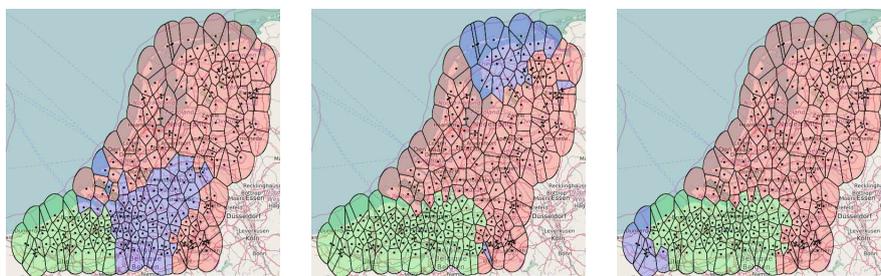


Figure 2: *k*-means suggested 21 different sets of three regions in 100 runs. **Left:** the most frequent set (19%) which split Flanders and assigned several locations from the Dutch South to one of the new regions. **Center:** a division in which the province Friesland has been identified, plus six unrelated locations (set frequency: 6%). **Right:** a configuration in which French Flanders, with one extra location, has been separated from the rest of Flanders (set frequency: 2%).

The work described in this paper was inspired by earlier work by John Nerbonne, in particular Nerbonne, Heeringa & Kleiweg (1999). Like John Nerbonne, we aim at creating models and visualizations for large sets of dialect data. Even though John is leaving his group in Groningen, his work will keep to have an influence in the field of dialectology in the coming decade. We wish him well in his future activities.

## References

- Barbiers, Sjf. 2006. *Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND)*. <http://www.meertens.knaw.nl/sand/> Accessed 26 February 2015.
- Barbiers, Sjf, Johan van de Auwera, Hans Bennis, Eefje Boef, Gunther Vogelaer & Margreet van der Ham. 2008. *Syntactic atlas of the Dutch dialects*. Amsterdam University Press.
- Daan, Jo & D. P. Blok. 1969. *Van Randstad tot Landrand*. Noord-Hollandse Uitgevers Maatschappij, Amsterdam, The Netherlands.
- Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy & Luca Scrucca. 2012. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*.
- Goebel, Hans. 2010. Dialectometry: theoretical prerequisites, practical problems and concrete applications. *Dialectologia* Special Issue(1). 63–77.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University of Groningen.
- Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations statistical natural language processing*. MIT Press.

- Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia: revista electrónica* (Special Issue II).
- Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Comparison and classification of dialects. In *Proceedings of EACL99*. ACL, Bergen, Norway.
- Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD thesis, LOT, Utrecht, The Netherlands.
- Tjong Kim Sang, Erik. 2014. *SAND: relation between the database and printed maps*. Tech. rep. Meertens Institute, Amsterdam, The Netherlands.
- Tjong Kim Sang, Erik. 2015. *Discovering Dialect Regions in Syntactic Dialect Data*.
- van Craenenbroeck, Jeroen. 2014. *The signal and noise in Dutch verb clusters – A quantitative search for parameters*. Manuscript, [http://jeroenvancraenenbroeck.net/s/paper\\_signal\\_noise.pdf](http://jeroenvancraenenbroeck.net/s/paper_signal_noise.pdf), version 18 December 2014, Retrieved 26 February 2015.



## Chapter 38

# Morphology changes faster than phonology<sup>1</sup>

Esteve Valls

University of Barcelona

In Valls, Wieling & Nerbonne (2013), we proved that the north-western Catalan varieties of Catalonia and Andorra (in contrast to those located in Aragon) have undergone a process of dedialectalization due to linguistic advergence to standard Catalan. In this paper, we intend to analyze whether this process affects the morphological and phonological components of language differently. To do so, we have performed a generative analysis of the corpus that has allowed us to discriminate the unpredictable (i.e. morphophonological) elements from the predictable (i.e. strictly phonological) components of language.<sup>2</sup> Therefore, we have been able to calculate the morphological and phonological distance among the north-western varieties separately and to visualize (by means of multidimensional analysis) the different evolutions of the morphological and the phonological components of these varieties throughout four generations. We attempt to shed light on the debate on which language components are more likely to the influence exerted by standard languages. We conclude that, at least in north-western Catalan, the hierarchy of instability of linguistic elements is clear: morphology has leveled faster than phonology. As a result, most of these varieties have undergone a process of accentualization.

## 1 Introduction

In Valls, Wieling & Nerbonne (2013) we took advantage of a range of dialectometric methods that allowed us to calculate and analyse the linguistic distance between 41 varieties in apparent time from an aggregate perspective. Specifically, we paid attention to the process of structural dialect loss due to linguistic advergence to standard Catalan undergone by most north-western Catalan dialects located in Catalonia

---

<sup>1</sup> This work is part of the project FFI2013-46987-C3-1-P ([www.ub.edu/GEVAD](http://www.ub.edu/GEVAD)), funded by the Spanish Ministerio de Economía y Competitividad.

<sup>2</sup> In this paper, we will refer to *morphophonological components* simply as *morphological components* in order to clearly contrast the differences between morphophonological underlying elements and strictly phonological elements (i.e. the phonological rules that produce the phonetic outputs).

(Spain) and Andorra. We also provided evidence that the dialect levelling that takes place in these two areas strongly contrasts with the relative stability of the Catalan dialects on the other side of the Catalan-Aragonese border in Spain, due to the fact that Catalan is not an official language in the Autonomous Community of Aragon. These opposite sociolinguistic situations (i.e. Catalonia and Andorra have strong language policies to support Catalan, whereas Aragon does not) have triggered a twofold process of vertical advergence between the Catalan spoken in Catalonia and Andorra towards the prestigious variety, on the one hand, and of horizontal divergence between these dialects and those located in Aragon, on the other hand. This situation has notably strengthened the border differences between Aragon and Catalonia during the last 80 years.

This paper is a step forward in the analysis of the abovementioned process of dedialectalization, because we intend to study the different *speeds* of change depending on the *nature* of language components from an aggregate perspective. Our hypothesis is that morphological aspects of language tend to change faster than phonological ones, a situation that would lead to *accentualization* according to Chambers & Trudgill (1998: 5). Because of space limitations, we will only mention five hierarchies of instability of linguistic components to show that there is no consensus about which of these two categories tend to change faster due to prestige reasons in situations of language or dialect contact:

- a) According to Viaplana (1999), who also focussed on north-western Catalan: morphology > phonology.
- b) According to Sankoff (2002), who focussed on several language contact situations: lexicon > phonology > morphology / syntax.
- c) According to Berruto (2005), who focussed on the influence of standard Italian on Italo Romance languages: lexicon > phonetics / phonology > morphology / syntax.
- d) According to van Coetsem (1988), who focussed on several language contact situations: lexicon > syntax and phonology > morphology.
- e) According to van Bree (1985), who focussed on the influence of Netherlandic Dutch on two Netherlandic dialects: lexicon > morphology > phonology > syntax.

## 2 Corpus

The dataset used in this paper was conceived as a corpus of contemporary north-western Catalan and covers the whole area where this dialect is spoken: all of Andorra and two dialect areas within Spain, specifically the western half of the Autonomous Community of Catalonia (with the exception of the Occitan-speaking Val d'Aran) and the eastern counties of the Autonomous Community of Aragon. Fieldwork was carried out in forty villages (two in Andorra, eight in Aragon, and thirty in

Catalonia) located in twenty counties. We added an artificial variety, standard Catalan, to these forty localities; hence, on the whole, we examine forty-one varieties. We interviewed 320 informants, 8 per locality, divided into four age groups. We selected a subset of 363 glosses per informant from the general questionnaire. These are distributed in eight (regular) morphological categories: articles, locative adverbs, verbs, and clitic, demonstrative, neuter, possessive and personal pronouns. The final corpus contains 113,749 items and 680,639 sound segments. More details about it can be found in Valls, Wieling & Nerbonne (2013).

### 3 Methodology

To examine the different speed of linguistic change on morphology and phonology, we first needed to discriminate, by means of a generative analysis, the unpredictable components of language (i.e. the underlying morphophonological differences) from its predictable elements (i.e. the regular phonological phenomena that generate the phonetic outputs). The aim was to calculate the linguistic distance for both categories separately in order to analyse how they evolved from the older speakers (F4) to the younger speakers (F1).

We illustrate this point by looking at the complete paradigm of the first person singular pronominal clitic (in bold in the examples below) in Cervera, Catalonia (see Table 1). The first two forms are proclitics followed by a consonant (a) and a vowel (b); the last two forms are enclitics preceded by a consonant (c) and a vowel (d).

In the examples below, the older speakers (F4) from Cervera use the form [me] to refer to the first person singular pronominal clitic before a verb starting with a consonant: [me'rento] 'I wash myself'. Younger speakers (F1), instead, use a more standardized form in the same context: [em]. Therefore, we might think that the underlying morphological forms of this clitic are /me/, /em/ and /m/. The fact is, however, that this clitic only appears with a vowel when the addition of the clitic to the verb creates a sequence that cannot be properly syllabified, that is: preceding a consonant ( $\_ \#C$ : [me]/[em] *rento*) or preceded by a consonant ( $C\# \_$ : *irritant*[me]). This is crucial to come to the conclusion that the underlying morphological form of the clitic is /m/ for both F4 and F1 speakers and that, being [e] the unmarked epenthetic vowel of north-western Catalan, in clitic clusters it is inserted to satisfy syllabic requirements.

Under this view, the difference between F4 and F1 is not morphological but merely phonological, as it lies in the position of the epenthesis. In F4, the epenthetic vowel is always placed to the right of the clitic, i.e. [me] in (a) and (c). In F1, instead, it always appears at the periphery of verb-clitic sequences, i.e. at the beginning in (a) but at the end in (c). That indicates that there has been a phonological change between F4 and F1 due to the different strategies that speakers undergo to repair syllabification. As a consequence, during the generative analysis we decided to assign a *Phonological Rule 1: epenthesis to the right* to the F4 speakers and a *Phonological Rule 2: peripheral epenthesis* to the F1 speakers. Thus, the morphological distance between F4 and F1 is 0, but the phonological distance between these two age cohorts is 1.

Table 1: Complete paradigm of the first person singular pronominal clitic in Catalan. It may appear in four contexts:  $\_ \#C$  (a),  $\_ \#V$  (b),  $C\# \_$  (c) and  $V\# \_$  (d).

	Cervera (F4)	Cervera (F1)
a. <i>em rento</i> ‘I wash myself’	[mɛˈrɛ̃nto]	[ɛmˈrɛ̃nto]
b. <i>m’irrita</i> ‘he irritates me’	[miˈrita]	[miˈrita]
c. <i>irritant-me</i> ‘irritating me’	[iriˈtamme]	[iriˈtamme]
d. <i>renta’m</i> ‘wash me’	[rɛ̃ntam]	[rɛ̃ntam]

This is the method we followed to discriminate the unpredictable components of language (i.e. the underlying morphological differences) from its predictable elements (i.e. the regular phonological rules that produce the phonetic outputs). As a result, we obtained two different databases: one containing the underlying morphological forms and another one comprising the phonological phenomena involved. To simplify the analysis, each phenomenon has been assigned a number. In addition, all the inputs (both the underlying forms and the processes) have been manually aligned. The measure of distance we used is the following:

$$\text{dist}(i, j) = \frac{\sum_{k=1}^{\text{long}} \text{dif}_k(i, j)}{\text{long}} \times 100 \quad (1)$$

That is, the linguistic distance between two varieties ( $i, j$ ) is the result of the summation of their differences (each having a value of 1) with regard to a linguistic variable  $k$  and dividing them by *long*, which is the length of each item compared. This dialectometric method was developed by Viaplana (1999) at the University of Barcelona. More detailed information about it can be found in Clua & Lloret (2006) and Valls et al. (2011).

## 4 Results

In this section we present six plots which show the results of applying multidimensional scaling to three distance matrices: Figs. 1 and 2 show the linguistic distance among varieties on the basis of phonetic data; Figs. 3 and 4 show the results of calculating the linguistic distance using morphological data; finally, Figs. 5 and 6 show the linguistic distance among varieties taking into account phonological data only.

To investigate dialect change, we will contrast the oldest and youngest age groups (i.e. F4 and F1). If we look at the MDS plots dealing with these age groups in Figs. 1 and 2, we can observe two noteworthy facts. First, the varieties that, based on the data of older speakers’ pronunciation (F4), were regularly spread on the left side of the plot have undergone a process of homogenization in the younger (F1) speakers, i.e. a gradual reduction of their original differences. However, this new, more homogeneous grouping is still located in Fig. 2 *at some distance* from the standard. Second, the va-



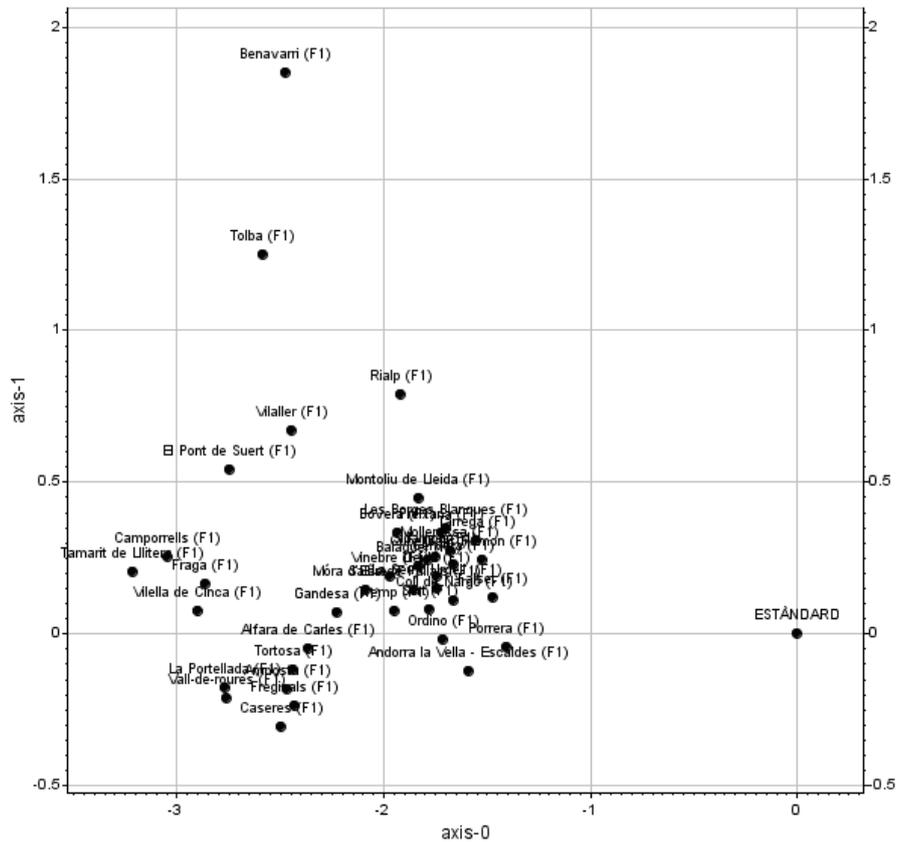


Figure 2: Multidimensional analysis based on the pronunciation of the younger speakers (F1). The plot visualizes 58% of variance.

counteract the strong morphological advergence to the standard undergone by these dialects.

Figs. 3 and 4 should help us prove this hypothesis. These figures display the linguistic distance among varieties on the basis of morphological data. The results are revealing: first, we can see that the morphological distance to the standard is smaller than the phonetic distance both among the F4 and the F1. As a consequence, it is likely to assume that the weight of the phonological rules is more important than the weight of the morphological components in the linguistic distance of Figs. 1 and 2. In fact, the process of *structural dialect loss* (Hinskens, Auer & Kerswill 2005: 11) undergone by most north-western dialects is clearly visible among the F1 speakers in Fig. 4. Their position in the plot indicates that they have massively adopted many standard morphological features. Second, we can see that *tortosí* and *ribagorçà* are

38 Morphology changes faster than phonology

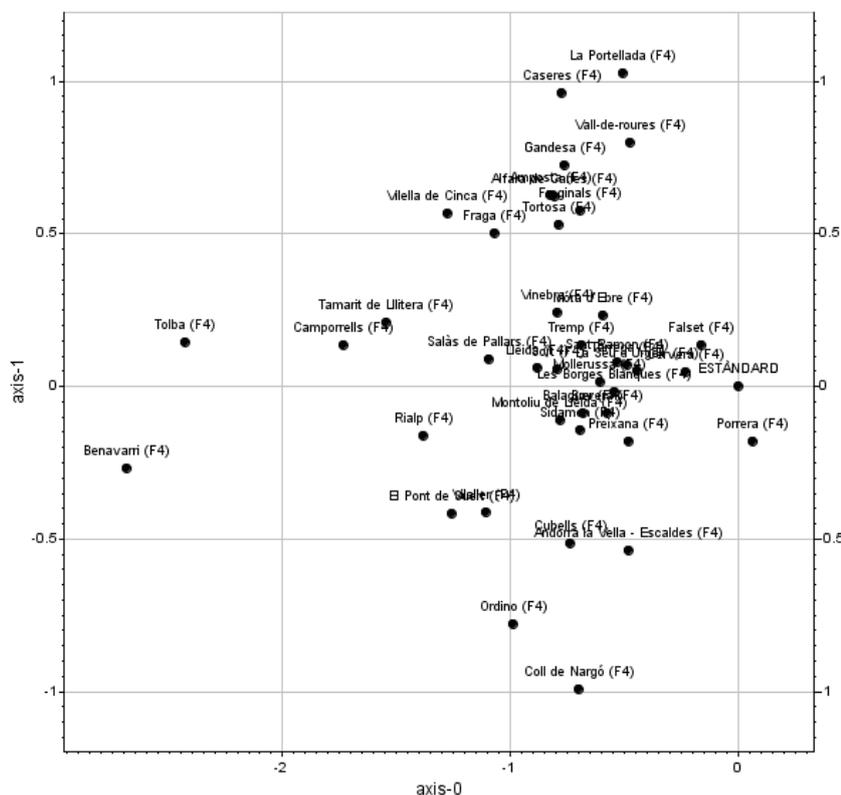


Figure 3: Multidimensional analysis based on the morphological components of the older speakers (F4). The plot visualizes 48% of variance.

the most conservative areas in Catalonia both among the F4 and the F1 speakers; the younger speakers of *pallarsès*, instead, have abandoned most of their morphological peculiarities. Third, the comparison of these two figures shows again the impact of the border effect between Catalonia and Aragon, especially in the central area, where the Catalan varieties have converged more with the standard.

If we agree with the fact that the morphological levelling of the majority of the north-western varieties with the standard has been very important, we will have to agree that the linguistic distance with respect to the standard that could be observed in Figs. 1 and 2 has to be attributed to the maintenance of the phonological rules that characterize these varieties. This hypothesis is confirmed by Figs. 5 and 6, which show the linguistic distance among varieties taking into account only phonological data.

A quick glimpse to the results shows that the levelling undergone by most north-

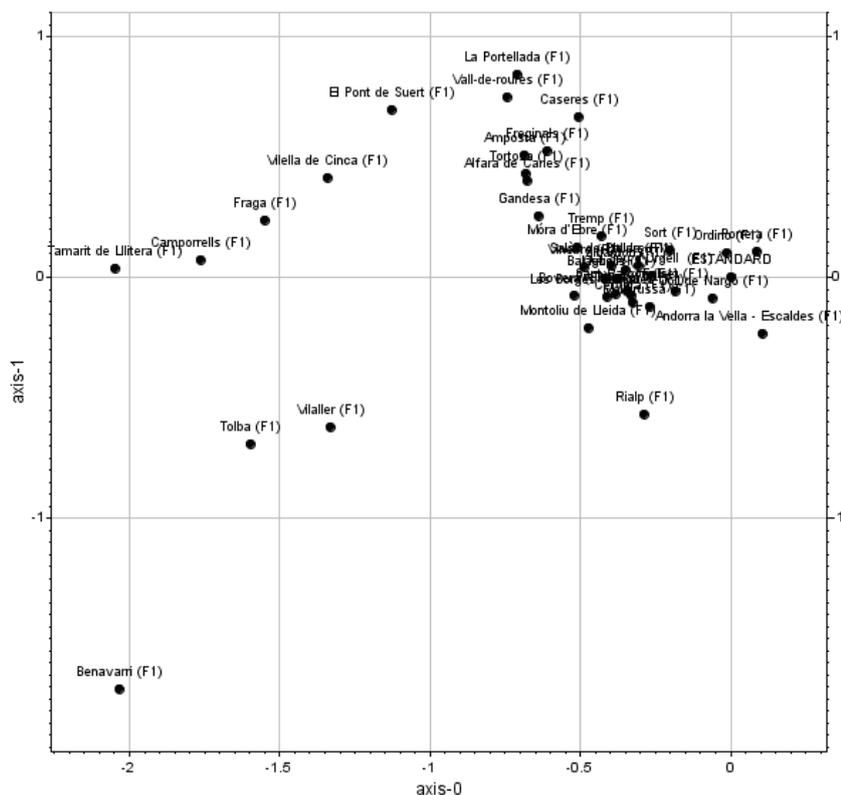


Figure 4: Multidimensional analysis based on the morphological components of the younger speakers (F1). The plot visualizes 58% of variance.

western varieties (with the exception of the most conservative dialects: *ribagorçà*, *tortosí* and the Aragonese varieties) is almost complete. There is an astonishing difference, however, between these plots and those based on morphological data: here the distance between north-western varieties and standard is very important. This fact confirms that nowadays most of these varieties have undergone a process of *accentualization*, as they are currently characterized by a small bunch of phonological rules.

## 5 Conclusions and prospects

In this paper, we have presented a way to calculate linguistic distance in the morphological and the phonological components of language separately. To do so, we used a dialectometric approach developed at the University of Barcelona. This ap-

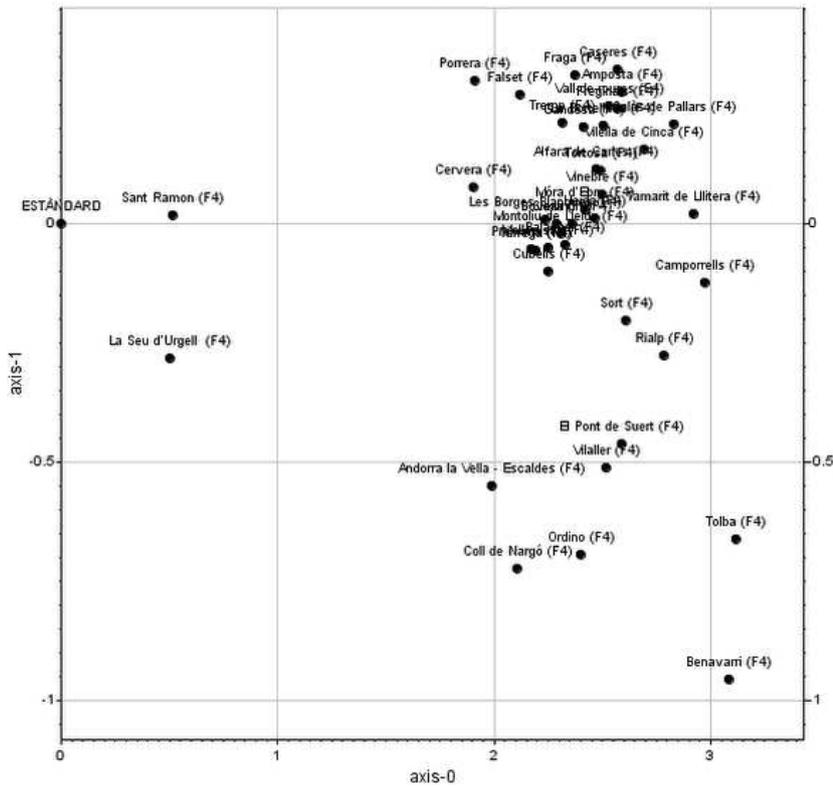


Figure 5: Multidimensional analysis based on the phonological rules of the older speakers (F4). The graph visualizes 55% of variance.

proach claims that it is possible to increase the accuracy of the final groupings by capturing the differences among varieties not only quantitatively but also qualitatively, by means of analysing the underlying differences that remain invisible in the phonetic data. Our purpose, however, has been more theoretical than methodological: we have tried to demonstrate from an aggregate perspective that the process of standardization undergone by most north-western varieties of the Catalan language affects the morphological aspects and the phonological aspects in a different way. We have actually given evidence that, at least in these varieties, morphology tends to level with the standard faster than phonology. This process has resulted in an accentualization of the north-western varieties of Catalonia (with the exception of two conservative dialects), and has allowed us to support the hierarchy of instability of linguistic elements proposed in Viaplana (1999): morphology > phonology. Next step should be to analyse whether this hierarchy is valid in other Catalan varieties and in

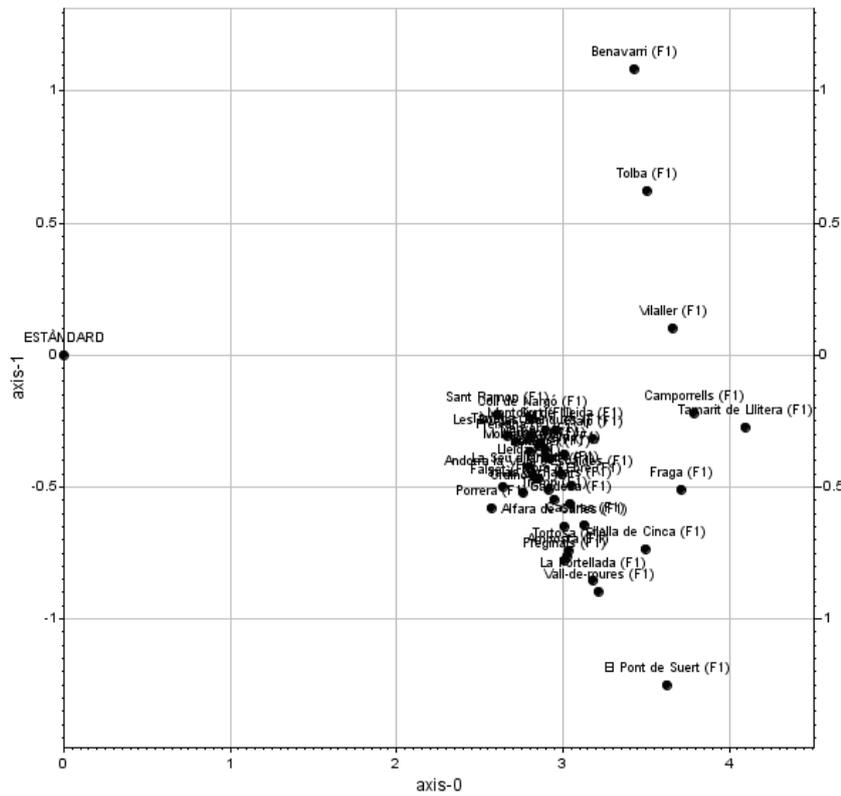


Figure 6: Multidimensional analysis based on the phonological rules of the younger speakers (F1). The graph visualizes 56% of variance.

other languages, but this is something that goes beyond the aims of this contribution.

## References

- Berruto, Gaetano. 2005. Dialect/standard convergence, mixing, and models of language contact: the case of Italy. In Peter Auer et al. (eds.), 81–95.
- Chambers, Jack & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.
- Clua, Esteve & Maria-Rosa Lloret. 2006. New tendencies in geographical dialectology: the Catalan Corpus Oral Dialectal (COD). In Jean-Pierre Y. Montreuil (ed.), *New perspectives on Romance linguistics*. Vol. 2: *Phonetics, phonology, and dialectology*. Amsterdam / Philadelphia: John Benjamins.

- Hinskens, Frans, Peter Auer & Paul Kerswill. 2005. The study of dialect convergence and divergence: conceptual and methodological considerations. In Peter Auer et al. (eds.), 1–48.
- Sankoff, Gillian. 2002. Linguistic outcomes of language contact. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 638–668. Malden / Oxford / Victoria: Blackwell.
- Valls, Esteve, John Nerbonne, Jelena Prokić, Martijn Wieling, Esteve Clua & Maria-Rosa Lloret. 2011. Applying the Levenshtein distance to Catalan dialects: a brief comparison of two dialectometric approaches. *Verba. Anuario Galego de Filoloxía* 39. 35–61.
- Valls, Esteve, Martijn Wieling & John Nerbonne. 2013. Linguistic advergence and divergence in north-western Catalan: a dialectometric investigation of dialect leveling and border-effects. *Literary and Linguistic Computing* 28.
- van Bree, Cor. 1985. Structuurverlies en structuurbehoud in het dialect van Haaksbergen en Enschede, een onderzoek naar verschillen in resistentie. *Leuvense Bijdragen* 74. 1–35.
- van Coetsem, Frans. 1988. *Loan phonology and the two transfer types in language contact*. Dordrecht: Foris.
- Viaplana, Joaquim. 1999. *Entre la dialectologia i la lingüística. La distància lingüística entre les varietats del català nord-occidental*. Barcelona: Publicacions de l'Abadia de Montserrat.



## Chapter 39

# Effective communication. A Platonic case study

Gerry C. Wakker  
University of Groningen

As a contribution to the valedictory volume for John Nerbonne I present a case study of two thematically comparable passages in Plato for which I will make use of an analysis based on discourse cohesion and above all on the rhetorical and manipulative use of particles to show how a single discourse participant may convey two completely different views using the same linguistic means. The basic issue in both passages is the same: is the specialist the best and most convincing speaker on the specialism in question or is this the orator or, in modern terminology, the communication expert?

### 1 Introduction

There is nowadays growing awareness that specialists, however good they may be in their profession, are not always and in all situations the best communicators. This applies not only to communication to non-specialists, but also to mutual communication between specialists in various work-related situations.

This awareness plays an important role in health care. In Groningen this led to a close cooperation between the Faculty of Arts and the Faculty of Medical Sciences/University Medical Center Groningen: the creation of the Health Communication Platform<sup>1</sup> which develops many initiatives in health communication research and education. This research in the Faculty of Arts is inherently linguistic and is, therefore, embedded in the Centre of Language and Cognition Groningen, the research institute successfully directed by John Nerbonne as scientific director for many years (1999–2012).

Health communication may well be a popular theme nowadays, but it is in no way new. I'd like to take the reader back to classical antiquity, where the question whether communication about health issues can best be done by the skilled physician or by a

---

<sup>1</sup> Kennisplatform Gezondheidscommunicatie, see <http://www.rug.nl/research/platform-gezondheidscommunicatie/>.

*Gerry C. Wakker*

communication expert, an expert orator or reciter, is already discussed. Let us have a closer look at two passages of Plato, the philosopher, who lived from 427–347 BC. In many of his dialogues, Plato presents us the philosopher Socrates in discussion with the so-called Sophists, traveling scholars who offered their expertise to the public against payment of a tuition fee. In these two passages I will study by which linguistic means the moderator of the discussion, Socrates, moderates, and even manipulates, the debate and brings his interlocutor to accept a conclusion that he cannot deny, though it is not at all the conclusion he was dreaming of but rather the contrary (sections 3 and 4). As a general introduction, I'll first give a more general description of discourse cohesion and the framework used to describe ancient Greek particles (section 2).

## 2 Discourse cohesion in general<sup>2</sup>

The Platonic passages will be studied as discourse, i.e. as texts functioning within a specific communicative situation.<sup>3</sup> Discourse coherence and cohesion are pivotal terms in discourse analysis. In ordinary language use we assume discourse to be coherent. The term coherence refers to the appearance of a discourse as a unified whole, instead of as a set of unrelated utterances simply placed together at random.<sup>4</sup> The presence of coherence is subjective (both for the speaker and for the addressee), since it is the result of an interpretation process and it depends on evaluation. However, the presence of certain linguistic elements may help the addressee to arrive at the interpretation intended by the speaker. These linguistic or cohesive means explicitly mark the coherence of a discourse. It is generally agreed that cohesion consists of grammatical and lexical elements that mark connections between parts of a discourse.<sup>5</sup>

Although coherent discourse usually makes use of different cohesion devices to explicitly mark the coherence relations, it is possible that discourse displays coherence without cohesion devices, or vice versa, cf. e.g. (ex. taken from Tanskanen 2006: 16-7):

- (1) A: That's the telephone.  
B: I'm in the bath.  
A: O.K.

Although linguistic means to indicate the specific relation between the statements in (1) are absent, the addressee A can still infer a (subjective) interpretation about this relation in its context (probably concluding that B cannot answer the telephone).

---

<sup>2</sup> This section is an updated version of S. J. Bakker & Wakker (2009: xi-xiv) and Wakker (2009: 65-66).

<sup>3</sup> By discourse I refer to the dynamic process of speech and writing in its situational context. Cf. Brown & Yule (1983: 23-5); Kroon (1995: 30 n. 50); Alfonso (2014: 35-38, 41-43); Christiansen (2014: 34).

<sup>4</sup> See e.g. Taboada (2004: 1-4); Tanskanen (2006: 1, 7, 20); Alfonso (2014: 11-22); Gruber & Redeker (2014: 2-5).

<sup>5</sup> The distinction coherence vs. cohesion is based on the fundamental work of Halliday & Hasan (1976). See Alfonso (2014) and Christiansen (2014) for a recent overview of the literature.

Thus, coherence is felt to be present.<sup>6</sup> In many cases, however, the relation is made explicit, either by intonation or by the use of cohesive devices such as discourse markers or particles.<sup>7</sup> Thus speaker B may say *But I'm in the bath*, indicating by the discourse marker *but* that there is some adversativity between the two utterances. Exactly which element is contradicted or corrected is inferable from and therefore dependent on the context: in this case it is not the fact that the phone rang or needs to be picked up, but the presupposed element of A's implied request that speaker B is to pick up the phone. This implication is corrected by (the implications of) B's utterance. When *but* or some similar discourse marker is present, the adversative relation is immediately clear to the hearer. In this way, a discourse marker can ease the cognitive effort of a hearer. The discourse marker helps to block unwanted alternative interpretations. In other words, in normal language use discourse markers make relations explicit, rather than creating them.

In (rather rare and marked) cases, we see the contrary: despite the presence of clear cohesive means, the coherence of the discourse in question is difficult, if not impossible, to establish, cf. (2):

- (2) Courses ended last week. Each *week* has seven days. Each *day* I feed my cat. *It* has four legs, *and it* is in the garden. *The latter* has six letters.  
(example from Tanskanen 2006: 16, slightly modified)

Cohesive elements (anaphoric reference, repetition, coordination) are in italics. Despite their presence, this piece of discourse does not form a coherent unified whole.

Though particles belong to the most important cohesion devices they are very difficult to describe: they usually fall outside the syntactic structure of the clause in which they occur, and their semantics is elusive. A pragmatic discourse approach is more rewarding.<sup>8</sup> Instead of a *referential* meaning (contributing to the representation of an event, a situation, an action etc), particles have a *functional* meaning, which has to do with the placing of the described state of affairs in the communicative (textual and non-textual) context.<sup>9</sup> From the point of view of the addressee, particles may be considered a kind of road signs in the text which help him keep track of the structure of the text or find out the communicative purpose or expectations of the speaker. From the perspective of the speaker, particles may be described as a means of placing the unit they have in their scope into a wider perspective, which may be the surrounding context (and its implications) or the interactional situation the text

<sup>6</sup> Cf. the cooperative principle of Grice (1975: 45-58), stating that discourse participants should make their contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the discourse in question. The fact that the addressee comes to this conclusion may be seen as a conventional implicature, i.e. as an inference the addressee can make from the presumption that speakers are seeking to provide useful information.

<sup>7</sup> Kroon (1995: 36) explains that particles belong to the larger category of discourse markers, which indicate how a text unit is integrated into the discourse context. See also Blakemore (1992: 136), Christiansen (2014: 161), Gruber & Redeker (2014: 6).

<sup>8</sup> See e.g. Levinson (1983: 100), Abraham (1986: 87-100), Kroon (1992: 53-8, 114; 1995: 34-57). As to ancient Greek, see e.g. Rijksbaron (1997), S. J. Bakker & Wakker (2009), Koier (2013), Drummen (2015).

<sup>9</sup> See, for instance, Kroon (1992: 55-6; 1995: 41, 61-2), Wakker (2009).

forms part of. The speaker may also use particles as a means of trying to influence the interpretation by the addressee.

Central to a pragmatic approach are three further observations: first, every discourse can be analysed at at least three levels: the representational, the presentational or text structuring, and the *interactional* level.<sup>10</sup> At each level particles may occur. Most particles are primarily linked to one of the levels, but may also function secondarily at another level, as the Greek examples will show.

Second, discourse is usually structured hierarchically. One may discern various layers in discourse, for instance, *embedding* and *embedded* sequences. The starting point of an embedded sequence (also called PUSH) and the point where a speaker returns to the embedding sequence (POP) are very often marked by particles and other relators.

Third, particles are often used in an ironic or rhetorical way, rather than in a literal way. Reinterpretation (often in a conventionalized way) by the addressee is necessary in such cases, if the literal interpretation clearly cannot be meant by the speaker in the given context. Often a speaker uses a particle to influence the addressee's interpretation. Thus *it seems* or *maybe* (or *pou* in ancient Greek) may be used as a disclaimer, as an indication that the speaker is not sure about his statement, and thus gives the addressee the possibility to disagree with the speaker's statement (without disturbing the conversation or their mutual relationship), but when added to something evidently true *maybe* may be used ironically as underlining the truth of the statement (because it is evident that the addressee cannot but agree). As a consequence it also functions in this way if the statement is not self evident; it would, however, be a rough interruption of the argumentation if the addressee were to question the truth of the statement (rhetorical or manipulative use).

I will show how the combination of the above three observations makes the process of using and interpreting particles in their context subtle and complicated, but also essential for effective communication.

Within the above framework the main particles used in these two Platonic passages may be described as in Table 1.<sup>11</sup>

### 3 Plato Ion

The first passage belongs to a brief dialogue<sup>12</sup> called *Ion*. Ion is a rhapsode, a professional singer and reciter of epic poetry who also lectures on Homer. He is the proud winner of a recent rhapsode contest. This leads to a discussion about the question whether the rhapsode gives his performance thanks to his skill and knowledge (his *technè*) or by virtue of divine possession or inspiration. Ion considers himself a skillful professional. During the discussion, Ion has first to admit that if he is an expert, he

<sup>10</sup> These distinctions are based upon Kroon (1995).

<sup>11</sup> For a justification of these values see E. J. Bakker (1993), Sicking & van Ophuijsen (1993), Sicking (1997), Wakker (1997; 2009), Koier (2013: 271-328) and Drummen (2015), who each mention further relevant literature.

<sup>12</sup> Plato, *Ion* 536e1-538b6. There is discussion about the date of composition and the authenticity of the *Ion*. For a clear overview of the discussion, see Rijksbaron (2007: 1-14).

Table 1: Function of cohesive devices (particles).

Cohesive device (particle)	Discourse level at which the particle primarily functions	Basic semantic value of the connection	Explicit relation with previous text	Relevance discourse unit	Commitment avowed by S	Commitment presupposed in A (by S)
<i>gár</i>	presentational	explanation	yes	PUSH (explanation, elaboration)	neutral	neutral
<i>oûn</i>	presentational	next important point	yes	POP-particle: important point	neutral	neutral
<i>oukoûn</i> (introducing a rhetorical – question)	presentational + interactional (i.e. here: attitudinal)	next important question	yes	POP-particle: important question	suggesting positive answer	positive
<i>toînun</i>	presentational + attitudinal	next highly important point; you must note	yes	POP-particle: highly important point	high: 'you take it from me that'	low
<i>ára</i>	attitudinal	given the preceding: we cannot but accept (even if it is surprising)	yes	on the basis of the preceding (concluding)	low	low
<i>pou</i>	attitudinal	perhaps/it seems	neutral	no	(feigned) uncertainty about truth	neutral
<i>dèpou</i>	attitudinal	as we both see ( <i>dè</i> = δῆ) it seems ( <i>pou</i> ) > evidently	yes	neutral	high (evidently it seems)	high (it seems)
<i>dé</i> (= δέ)	presentational	next new item	yes and no; discontinuity within larger text unit	neutral	neutral	neutral

will be an expert not only in Homer but also in other poets and in poetry as a whole. Ion wonders, however, why he has this great ability, acknowledged by everyone, concerning Homer only. Socrates now explains that his ability does not depend on skill

or knowledge, but on inspiration, ultimately inspired by the Muses themselves. As a magnet exerts power over a chain of iron rings, so the Muses' inspiration extends from poet to rhapsode to audience. Next (536e1-539d52), Ion is forced to admit that judgements about chariot driving, medical and other specialized issues mentioned by Homer are better left to the respective specialists than to the rhapsode. This brings Socrates to his concluding question: what is the object of the art of a rhapsode. After several unfelicitous answers of Ion, Socrates finally asks (540c): "Well, will the rhapsode know better than the doctor what sort of thing to tell a statesman when he is ill?" Ion cannot but reply that this is not the case and that clearly a doctor knows best. At the end, the only thing left for Ion is the following: instead of a proud expert Ion is to be characterized as (only) a divinely inspired reciter. Let us have a closer look at the argumentation and linguistic means used to lead the interlocutor to the desired result in 536e1-538b6.<sup>13</sup> The argumentation develops via the so-called Socratic method of argumentation, which consists of disciplined and systematic questioning, by which Socrates follows out the logical implications of thought. Before going to a next step of the questioning/argumentation Socrates always checks whether his interlocutor agrees. Particles are playing an essential role, and, as we will see, not every new step takes the same place in the argumentation, both objectively and subjectively in Socrates' view.

- (1) [536e] *Socr*: ..., but first answer me this: on what thing in Homer's story do you speak well? For (*ou gar dèpou*) not on all of them you do, I presume. *Ion*: I assure you, Socrates, on all without a single exception. *Socr*: (*ou dèpou*) Not, I presume, on those things of which you have in fact no knowledge, but which Homer tells. *Ion*: And what sort of things are they, which Homer tells, but of which I have no knowledge? [537a] *Socr*: (*ou*) Does Homer not speak a good deal about arts (*méntoi*), in a good many places? For instance, about chariot-driving: if I can recall the lines, I will quote them to you. *Ion*: No, I will recite them, for I can remember. (537a8-b5: recital of *Hom. Il.* 23.335ff) *Socr*: Enough. As to those lines then, Ion, will a doctor or a charioteer be the better judge [537c] whether Homer speaks them correctly or not? *Ion*: A charioteer, of course/I presume (*dèpou*). *Socr*: Because he has this art, or for some other reason? *Ion*: No, because it is his art. *Socr*: [537d1] (*oukoûn*) To every art then has been apportioned by the god a power of knowing a particular business, isn't it? (*ou gar pou*) For, I think, what we know by the art of piloting we cannot also know by that of medicine. *Ion*: (*ou dèta*) No, indeed. *Socr*: And what we know by medicine, we cannot by carpentry also? *Ion*: (*ou dèta*) No, indeed. *Socr*: (*oukoûn*) And this rule then holds for all the arts, that what we know by one of them we cannot know by another, isn't it? But (*dé*) before you answer that, just tell me this: do you agree that one art is of one sort, and another of another? *Ion*: Yes. *Socr*: (*âra*) Do you argue this as I do, and call one

<sup>13</sup> My translations are an adapted version of W. R. M. Lamb (1925), *Plato. Plato in Twelve Volumes*. Cambridge, MA: Harvard University Press/London: William Heinemann Ltd., see <http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman>. Especially the translations of the cohesion devices are mine.

art different from another when one is a knowledge of one kind of thing, and another a knowledge of another kind? [537e] *Ion*: Yes. *Socr.* (**gar pou**) Since, I suppose, if it were a knowledge of the same things—how could we say that one was different from another, when both could give us the same knowledge? Just as I know that there are five of these fingers, and you equally know the same fact about them; and if I should ask you whether both you and I know this same fact by the same art of numeration, or by different arts, you would reply, I presume (**dèpou**), that it was by the same? *Ion*: Yes. [538a] *Socr.* (**toinun**) Then tell me now, what I was just going to ask you, whether you think this rule holds for all the arts—that by the same art we must know the same things, and by a different art things that are not the same; but if the art is other, the things we know by it must be different also. *Ion*: I think it is so, Socrates. *Socr.* (**oukoûn**) Then he who has not a particular art will be incapable of knowing aright the words or works of that art, isn't it? [538b] *Ion*: True. *Socr.* (**oun**) Then will you or a charioteer be the better judge of whether Homer speaks well or not in the lines that you quoted? *Ion*: A charioteer. *Socr.* (**gar pou**) Because, I suppose, you are a rhapsode and not a charioteer. *Ion*: Yes. *Socr.* And (**dé**) the rhapsode's art is different from the charioteer's? *Ion*: Yes. *Socr.* (**ára**) Then if it is different, it is also a knowledge of different things. [538b6] *Ion*: Yes.

This passage starts with an open question explained by a second question introduced by the particle combination *ou gar dèpou*. Questions introduced by the negation *ou* are questions eliciting a positive answer (if the interlocutor intends to be cooperative). This steering question is marked by *gar*, a push particle, here introducing an explanation why the previous question is a relevant one. *Dèpou* is an attitudinal particle marking that the speaker presents his utterance as evident (*dè*), but at the same time leaves room for doubt of the addressee (*pou*). This combination can be used in a neutral or literal sense (as here), but when added to a really evident statement it emphasizes the truth and the *pou* part is only seemingly expressing doubt (ironic use). After *Ion*'s proud claim that he can speak well about everything, Socrates uses, again, *ou dèpou*, now in a more suggestive way in view of the previous part of the dialogue. *Ion* replies (non preferred reaction) with a counter question, thus showing that he is not ready for the next step. Socrates answers by a suggestive question introduced by *ou*<sup>14</sup> and giving an example. *Ion* eagerly offers to cite the lines in Homer referred to by Socrates. In 537c Socrates resumes his argumentation with a seemingly harmless and obvious question. *Ion* underlines his answer with the 'ironic' use of *dèpou*. They agree about the explanation (*technè*). After the example Socrates generalizes, in his characteristic way, by an *oukoûn*-question. *Oukoûn* is the combination of a yes-eliciting question (*ouk*) and a next step to a now important question (*oun*). It is, hence, a POP particle, and, as often, after an example, it introduces the general rule. With *ou gar pou* Socrates offers another example as explanation. *Pou* is added to this evident statement to express, here only feigned, uncertainty. *Ion* fully agrees, marking the answer as self evident (*dèta*). The same holds for the second question.

<sup>14</sup> After *about arts méntoi* is added, by which Socrates reinforces the truth value of the assertion implied by his question, cf. Rijksbaron (2007: 193-4).

With *oukoûn* (537d1) Socrates introduces the general rule. Before Ion is able to react, Socrates introduces a sidetrack or second line of argument, marked by *dé*, the most common Greek particle that marks a next new item (presentational particle), here the new question whether the *technai* differ and imply knowledge of different things, with which Ion agrees. Socrates continues with a question introduced by *âra*, the neutral marker of a yes/no-question. Ion gives an affirmative answer. With *gâr pou* (537e1) Socrates introduces an elaborating rhetorical question with ironical *pou* and sketches Ion's supposed reaction to an example, the self-evidentiality of which is marked by ironical *dèpou*. Ion agrees. In 538a we return to the main line of the argumentation (left in 537d2), which is at the same time the generalization of the example in 537d. Here Socrates marks with *toínun* that he will put forward a highly important point, a point that Ion must take notice of ("you take it from me that"). Ion consents. Socrates continues to his next step with an *oukoûn*-rhetorical question and after Ion's consent Socrates next (*oûn*) applies this to the Homeric lines cited before, explaining (*gâr pou*) this statement they agree upon with two questions, linked by *dé*. Socrates concludes this argumentation with an *âra*-statement: if it is different, it is also knowledge of different things. Given the previous argumentation, this conclusion must be accepted. In this way Ion is forced (after some more examples) that his statement that the rhapsode knows everything (536e1) is false.

The conclusion of this passage (536e1-538b6) is clear: only the expert is able to talk with his expertise and knowledge about the area of his expertise. In short, in our terms, the doctor can communicate on health issues in a more expert way than the rhetor, rhapsode or, to use modern terminology, the communication expert.

## 4 Plato Gorgias

In his dialogue *Gorgias*, Socrates enters into discussion with the famous orator and sophist Gorgias about the nature of rhetoric. In one passage (459a1-c5) the theme is the issue whether it is true that in a public of non-experts the orator persuades his audience with more ease than the expert. Let us have a closer look at this passage, which shows us the same technique of how Socrates leads and even manipulates the argumentation.

- (2) [459a] *Socr*: You were saying just now, I tell you (*toi*), that even in the matter of health the orator will be more convincing than the doctor. *Gorgias*: Yes, indeed, (*kai gâr*) I was, at least to the crowd. *Socr*: (*oukoûn*) And "to the crowd" means "to the ignorant"? For surely (*ou gâr dèpou*), to those who know, he will not be more convincing than the doctor. *Gorgias*: You are right. *Socr*: (*oukoûn*) And if he is to be more convincing than the doctor, he thus becomes more convincing than he who knows, isn't it? *Gorgias*: Certainly. *Socr*: Though not himself a doctor, you agree (*è gar*)? [459b] *Gorgias*: Yes. *Socr*: But he who is not a doctor is surely (*dèpou*) without knowledge of that whereof the doctor has knowledge. *Gorgias*: Clearly. *Socr*: He who does not know will, so it appears (*âra*), be more convincing to those who do not know than he who

knows, supposing the orator to be more convincing than the doctor. Is that, or something else, the consequence? *Gorgias*: In this case it does follow. *Socr*: (*oukoûn*) Then the case is the same in all the other arts for the orator and his rhetoric: there is no need to know [459c] the truth of the actual matters, but one merely needs to have discovered some device of persuasion which will make one appear to those who do not know to know better than those who know. *Gorgias*: (*oukoûn*) Is it not then a great convenience, Socrates, to make oneself a match for the professionals by learning just this single art and omitting all the others? *Socr*: We shall look into presently, if our argument so requires: (*dé*) for the moment let us consider first...

Socrates starts this part of the dialogue by picking up a previous statement (456b6-c2), by which Gorgias imagined the orator and physician as rival candidates for a medical appointment. Socrates marks his statement as particularly interesting for Gorgias by *toi*. Gorgias agrees, explaining and elaborating his statement (*kai gâr*). After Socrates' check whether he understands it correctly (with *oukoûn*, explained by the self evident *dèpou*), he proceeds to the next step, introduced by *oukoûn* and agreed upon by Gorgias. Socrates then adds two further questions, marked by *è gâr*<sup>15</sup> and emphasizing/ironical *dèpou*. Socrates next proceeds to the conclusion based on the previous statements (*ára*). After Gorgias' agreement Socrates proceeds to his more general conclusion, characteristically introduced by *oukoûn*. Gorgias remarkably (non preferred reaction) replies by a rhetorical question introduced by *oukoûn*, but implying he agrees and evaluating it as a great asset. Socrates puts this theme aside and switches with *dé* to the theme he is interested in now, thus taking again the lead in the discussion.

In an earlier passage (456b1-5) Gorgias already illustrated this conclusion on the basis of his own experience: "many and many a time have I gone with my brother or other doctors to visit one of their patients, and found him unwilling either to take medicine or submit to the surgeon's knife or cautery; and when the doctor failed to persuade him I succeeded, by no other art than that of rhetoric."

In our terminology, the conclusion in this dialogue (which suits its theme) is that the communication expert is more convincing and persuasive among non-experts than the doctor/physician.

## 5 Concluding remarks

We discussed two dialogues with a comparable topic, but leading to two opposite conclusions. In both cases it is Socrates who takes the initiative and leads his interlocutor by subtle use of particles. The argumentation proceeds step by step. Particles often mark the new step and indicate its status in the argumentation after agreement

<sup>15</sup> The particle *è* added to a statement indicates that the speaker considers the statement as inevitably true, see Wakker (1997). Added to a question the speaker asks the addressee, whether he confirms the truth. See Drummen (2015: 133-7).

Gerry C. Wakker

has been assured concerning the previous step. The results in both passages are internally conflicting, but suit the topic, the structure of the argumentation and the characters of the interlocutors.

Another conclusion may be drawn and applied to current issues: specialist and communication expert should work together, the one possessing professional knowledge (see *Ion*), the other knowledge about effective communication (see *Gorgias*). In both cases, however, great attention must be paid to the way particles help to structure messages and lead the interlocutor(s) or audience to the result aimed at in the given conversation.

## References

- Abraham, W. (ed.). 1986. *Tijdschrift voor Tekst- en Taalwetenschap (TTT) 6.2: Special issue on particles*.
- Alfonso, Pilar. 2014. *A multi-dimensional approach to discourse coherence: from standardness to creativity* (Linguistic Insights: Studies in Language and Communication v. 180). Bern: Peter Lang.
- Bakker, E. J. 1993. Boundaries, topics, and the structure of discourse: an investigation of the ancient Greek particle *dé*. *Studies in Language* 17. 275–311.
- Bakker, S. J. & Gerry C. Wakker (eds.). 2009. *Discourse cohesion in ancient Greek*. Leiden/New York: Brill.
- Blakemore, D. 1992. *Understanding utterances*. Oxford.
- Brown, G. & G. Yule. 1983. *Discourse analysis*. Cambridge.
- Christiansen, T. 2014. *Cohesion: a discourse perspective* (Linguistic Insights: Studies in Language and Communication v. 133). Bern: Peter Lang.
- Dodds, E. R. 1959. *Plato Gorgias. A revised text with introduction and commentary*. Oxford: Clarendon Press.
- Drummen, A. 2015. *Dramatic pragmatics. A discourse approach to particle use in ancient Greek tragedy and comedy*. Heidelberg PhD thesis.
- Grice, H. P. 1975. Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and semantics*. Vol. 3: *Speech acts*, 41–58. New York.
- Gruber, H. & G. Redeker. 2014. *The pragmatics of discourse coherence: theories and applications* (Pragmatics & Beyond New Series v. 254). Amsterdam/Philadelphia: John Benjamins.
- Halliday, M. A. K. & R. Hasan. 1976. *Cohesion in English*. London.
- Koier, E. 2013. *Interpreting particles in dead and living languages. A construction grammar approach to the semantics of Dutch ergens and ancient Greek pou*. Leiden PhD thesis.
- Kroon, C. H. M. 1992. Particula perplexa. In G. Bakkum et al. (eds.), *Pentecostalia*, 53–64. Amsterdam: UvA.
- Kroon, C. H. M. 1995. *Discourse particles in Latin. A study of nam, enim, autem, vero and at*. Amsterdam: Gieben.
- Levinson, S. C. 1983. *Pragmatics*. Cambridge: CUP.
- Rijksbaron, A. (ed.). 1997. *New approaches to Greek particles*. Amsterdam: Gieben.

- Rijksbaron, A. 2007. *Plato. Ion or: on the Iliad. Edited with introduction and commentary*. Leiden/Boston: Brill.
- Sicking, C. M. J. 1997. Particles in questions in Plato. In A. Rijksbaron (ed.), 157–174.
- Sicking, C. M. J. & J. M. van Ophuijsen. 1993. *Two studies in Attic particle usage: Lysias & Plato*. Leiden: Brill.
- Slings, S. R. 1997. Adversative relators between PUSH and POP. In A. Rijksbaron (ed.), 102–130.
- Taboada, M. T. 2004. *Building coherence and cohesion*. Amsterdam/Philadelphia.
- Tanskanen, S. K. 2006. *Collaborating towards coherence: lexical cohesion in English discourse*. Amsterdam.
- Wakker, Gerry C. 1997. Emphasis and affirmation: some aspects of μήν in tragedy. In A. Rijksbaron (ed.), 102–130.
- Wakker, Gerry C. 2009. ‘Well I will now present my arguments’. Discourse cohesion marked by οὖν and τοίνυν in Lysias. In Bakker & Wakker (eds.), 63–81.



## Chapter 40

# Variation: from dialect to pragmatics, a progress report

Annie Zaenen

Stanford University

Brian Hicks

Stanford University

Cleo Condoravdi

Stanford University

Lauri Karttunen

Stanford University

Stanley Peters

Stanford University

The paper reports on experimental studies investigating the way native speakers of English use the construction *NP be ADJ to VP* with evaluative adjectives. We show that, contrary to established linguistic theory, this construction is not always interpreted as factive. We isolated one major context in which an important minority and, for some adjectives, a majority of native speakers prefers an implicative use. We investigate whether there might be a dialect split between factive and implicative users and conclude that, contrary to what we suggested in a previous paper, this is not the case. We discuss different subclasses and end with the tentative hypothesis that the variability that is found among speakers has to do with the difference in importance that different language users attach to the non-linguistic context against which they interpret utterances.

## 1 Introduction

When we think about language variation we tend to think about dialect variation and about variation that is conditioned by sociological variables such as gender, age and

social status. But recent technological advances allow us to collect data from many more speakers than before and studies involving more subjects show unexpected areas of variation that do not fall within these traditional categories. Here we report on one such case: variation in interpretation of the infinitival complements of evaluative adjectives illustrated in (1).

- (1) John was smart to go to Groningen.

These adjectives are supposed to be factive, hence their infinitival complement is presupposed to be true even when the matrix clause is negative. Some speakers, however, will, in some circumstances, interpret the event referred to in the infinitival clause as not having happened when the matrix clause is in the negative. In other cases they interpreted it as having happened in accordance with linguistic theory. What are the reasons for this difference in interpretation? Is it a dialect difference or is something else going on?

In this short paper we give a characterization of the adjective class that exhibits the unexpected behavior and summarize the first experiment that led us to the hypothesis of two different dialects. We then describe some further experiments that show that the variation is most likely of a different nature. In conclusion, we formulate a new tentative hypothesis to account for our findings.

## 2 Two types of NP *be* ADJ *to* VP adjectives

The syntactic frame *NP be ADJ to VP* can be used with adjectives that semantically belong to different classes. Here we focus on the subclasses that have been described as factive (see Norrick (1978), for more references see Karttunen et al. (2014)). As first reported in Karttunen & Zaenen (2013), we found that emotive and evaluative adjectives behave very differently under negation. Emotive adjectives such as *outraged*, *dumbfounded*, *ecstatic*, *furious*, ... behave indeed like factives, evaluative adjectives, such as *stupid*, *smart*, *lucky*, *mean*, *nice*, *brave*, ... exhibit more complex behavior. Our initial findings, which were the by-product of a study on *lucky*, are given in percentages in Table 1.

Table 1: Emotive and evaluative adjectives.

adjective	matrix polarity					
	affirmative			negative		
	Yes	No	undecided	Yes	No	undecided
emotive	100	0	0	95.7	4.3	0
evaluative	98.9	0.9	0.3	25	64.2	10.7

The subjects were presented with sentences such as (1), with either a positive or a negative matrix clause and had to decide whether the eventuality described in the in-

finitival clause had happened or not. Some speakers/hearers interpreted the negated evaluative adjectives sometimes as implicatives, i.e. under matrix negation, the complement is also interpreted as negated (*No* answer in the table). A first hypothesis that comes to mind is that speakers simply misread the stimulus and think that in fact the sentence was (2):

- (2) John was not stupid *enough* to waste money.<sup>1</sup>

That hypothesis might be plausible for oral presentation but, in our case, the stimuli were always presented in written form. Moreover, a study of web data and of the ENTenTen-2.0 corpus<sup>2</sup> revealed that, for some of these adjectives, the implicative use seems to be the most prevalent one: e.g. for *lucky*, 9 out of 11 occurrences are implicative, for *fortunate*, 16 out of 18 are, for *stupid*, two fifths of the uses are implicative, two fifths, factive and one fifth could not be determined in the ENTenTen-2.0 corpus. To give just a couple of examples; (3) from the ENTenTen-2.0 corpus and (4) from the web:

- (3) I have a family to support and I'm not stupid to put that in jeopardy, maybe you are.  
 (4) This is my first trip to Italy, so I was not brave to venture out alone.

Moreover, in the first large experiment that we report on in Section 3, we asked the subjects that gave implicative answers whether they would say the same thing to convey the interpretation that they had assigned to the stimulus sentence. 79% answered this question positively. A further small investigation, exemplified in Table 2, quickly led to one variable that seemed to influence the judgments. (In what follows we indicate the factive with F and the implicative reading with I.)

The second column shows that the relation between the adjective and the VP complement is not the same in all cases. In some cases, *for NP to VP* would be ADJ (to choose the best piece would be clever, to waste money would be stupid) the relation is the socially expected one. We call this the CONSONANT relation. These combinations often get an implicative reading. In others, *for NP to VP* would **not** be ADJ (to choose the worst piece would not be clever, to save money would not be stupid), the relation is not the socially expected one. We call these DISSONANT. These cases have a nearly unambiguous factive reading.

To spell out the hypothesis explicitly: *for John to save money would not be stupid* is a dissonant combination of adjective and VP hence a factive interpretation is likely for (5):

- (5) John was not stupid to save money.

*For John to waste money would be stupid* is a consonant combination, hence an implicative reading is likely for (6):

<sup>1</sup> This example seems to entail that John did not waste money. The *be ADJ enough to VP* construction is in general systematically ambiguous between an implicative and non-committal reading. (See Karttunen (1971) and Meier (2003) for discussion.)

<sup>2</sup> <https://www.sketchengine.co.uk/xdocumentation/wiki/Corpora/enTenTen>.

Table 2: Emotive and evaluative adjectives.

STIMULUS	ADJECTIVE-COMPLEMENT RELATION	ANSWERS	CHOICE	%
R. was not clever to choose the best piece	to choose the best piece is clever CONSONANT	R. chose the best piece	F	25
		<b>R. did not choose the best piece</b>	<b>I</b>	<b>64.2</b>
		undecided		10.7
R. was not clever to choose the worst piece	to choose the worst piece is not clever DISSONANT	<b>R. chose the worst piece</b>	<b>F</b>	<b>80</b>
		R. did not choose the worst piece	I	10
		undecided		10
K. was not stupid to save money	to save money is not stupid DISSONANT	<b>K. saved money</b>	<b>F</b>	<b>78.6</b>
		K. did not save money	I	14.2
		undecided		7.1
K. was not stupid to waste money	to waste money is stupid CONSONANT	K. wasted money	F	28.6
		<b>K. did not waste money</b>	<b>I</b>	<b>66.7</b>
		undecided		4.8

(6) John was not stupid to waste money.

(5) is assumed to cohere with people’s social expectations, whereas (6) is assumed to be surprising under the factive reading, whereas an implicative interpretation would render it coherent.

### 3 Experiment 1 and the Implicative Dialect Hypothesis

Our 2014 paper (Karttunen et al. 2014) details a first large scale experiment in which we showed that the consonance/dissonance distinction is indeed involved. The experiment was run on Amazon Mechanical Turk with 206 native speakers of English who each responded to 30 test sentences which exemplified 19 adjectives (*arrogant, brave, careless, cruel, evil, foolish, fortunate, heroic, humble, lucky, mean, nice, polite, rude, sensible, smart, stupid, sweet* and *wise*). The test sentences presented the adjectives in a VP context that we judged either consonant, dissonant or neutral. The subjects had the choice between an implicative interpretation (according to the author of the sentence the event referred to in the embedded infinitival did not happen), a factive interpretation (the event happened) and *Either*. (For a detailed description of the experiment, see Karttunen et al. (2014).)

The results confirm the importance of the consonant/dissonant distinction as shown in the summary in Figure 1.

In the 2014 paper, we put forward the hypothesis there might be a dialect split between speakers of English. One group for which the factive use is the norm, another group for which the implicative use is. Both groups would adapt to the other dialect if the consonance or dissonance of the infinitival complement nudged them to do so. Our reasoning was that, if the only factor at issue was the consonance/dissonance nature of the complements, speakers should give factive judgments for neutral stimuli and only be swayed by discourse coherence pressures in case of negated consonant

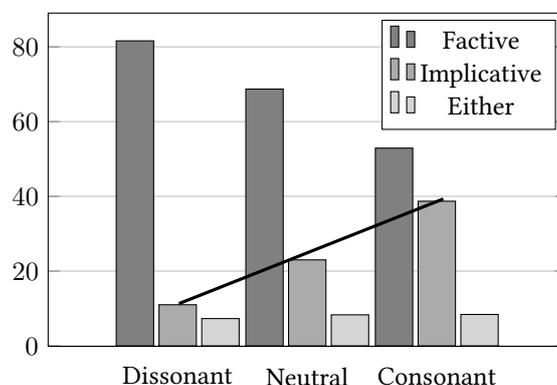


Figure 1: Results: Percentage of Factive, Implicative, and Either choices for *NP was not ADJ to VP*.

readings.<sup>3</sup> But if we have two dialects we expect there to be implicative judgments even in the neutral contexts. In fact, the neutral contexts would reveal the relative strength of the two dialects. Given the data in Figure 1, around one third of the speakers could be hypothesized to be implicative speakers.

But the results did not establish this dialect split conclusively as we had too few judgments per subject. Moreover we did not have any sociological data about our subjects. An other problem with the experiment was that we had labeled the infinitival complements as consonant or dissonant on an intuitive basis without checking whether our own judgments coincided with those of the community.

## 4 Experiment 2: consonance/dissonance judgments

To address the latter criticism, we ran a norming experiment asking our subjects to judge *How ADJ it would be to VP* (e.g. *How brave would it be to fight injustice in the face of danger?*), on a sliding scale (extremely = 1, not at all = .50, completely the opposite = 0). This experiment confirmed most of our intuitive judgments (with some exceptions: e.g. contrary to us, most of our subjects thought that living in Europe was an unlucky experience). We then combined these results with the results of a new consonance/dissonance experiment of the same design as the previous one but adding the adjectives *cowardly*, *kind*, *prudent*, *right* and *wrong* and leaving out *sweet*, ending up with 23 adjectives. The results showed that the the consonance/dissonance distinction was highly significant for the choice between the two interpretations:<sup>4</sup>

<sup>3</sup> Our results focus on negated sentences: in positive sentences, there is no difference in the interpretation of the embedded complement as having happened or not between implicatives and factives. There might be more subtle differences but these our experiments do not address.

<sup>4</sup> The *p*-value is a measure of the likelihood that the result of an experiment is due to chance. An experiment with a *p*-value lower than 0.001 ( $p < 10^{-3}$ ) is commonly accepted as a statistically highly

Factive:  $p$ -value  $< 2.2 \cdot 10^{-16}$   
 Implicative:  $p$ -value  $< 2.2 \cdot 10^{-16}$   
 Cannot Decide:  $p$ -value  $= 3.261 \cdot 10^{-5}$

But the correlation between the consonance/dissonance judgments and the factive/implicative readings also showed that consonant/dissonant variable explained less than fifty percent of the variation:<sup>5</sup>

Factive: Adjusted  $R$ -Squared = 0.433  
 Implicative: Adjusted  $R$ -Squared = 0.4344  
 Cannot Decide: Adjusted  $R$ -Squared = 0.07718

In the 2014 paper, we already observed that not all adjectives in our sample were equally sensitive to the consonant/dissonant distinction. The findings reported there confirmed what we had learned from the corpus data for *stupid*, *fortunate* and *lucky* (see Karttunen et al. (2014) for details). A further inspection of our list of adjectives suggested that they fall in a few broad classes. *Arrogant*, *cruel*, *evil*, *humble*, *mean*, *nice*, *polite*, *rude*, *kind* are about character, *fortunate* and *lucky* are about good or bad luck, *brave*, *heroic*, *cowardly* are about courage, *foolish*, *prudent*, *sensible*, *smart*, *stupid*, *wise* are about judgment whereas *right* and *wrong* give an overall moral appreciation. With respect to factive and implicative readings the classes are ordered as shown in Figure 2.

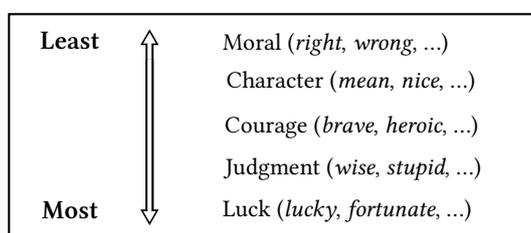


Figure 2: Adjective classes.

*Right* and *wrong* are least affected by the consonant/dissonant effect, *lucky* and *fortunate* are at the opposite end of the scale. When we add these classes to our calculations, the measure of variation accounted for improves dramatically to an adjusted  $R$ -Squared of 0.7929. Of course, this only tells us that the distinction among the classes is important but not what the distinction exactly is and certainly not how it should be explained. We are currently trying to find an operational characterization of this distinction, hypothesizing that factors like the control of the protagonist

significant result.

<sup>5</sup>  $R$ -Squared is the percentage of the response variable variation that is explained by a linear model. The adjusted  $R$ -squared is a modified version of  $R$ -squared that has been adjusted for the number of predictors in the model. The adjusted  $R$ -squared increases only if the new term improves the model more than would be expected by chance.

over the characteristic given in the adjective (one can be *mean* on purpose but one cannot be *lucky* on purpose) play a role.

## 5 Experiment 3: the Implicative Dialect Hypothesis rejected

In our third experiment, we concentrated on 39 native speakers of English. We recorded data about the age, the gender and the regional provenance but did not balance the sample with respect to these variables. We chose 6 adjectives (*lucky, fortunate, mean, nice, stupid, foolish*) from our previous set and added *right* and *wrong* to them and presented 9 sentences (three consonant, three dissonant and three neutral ones) per adjective to each subject. A logistic model of the probability of the answer as a function of the consonance score correlated, for each subject, his/her judgments with the judgments about consonance and dissonance that we had obtained in our previous experiment. The results for the negative examples confirm that there are subjects that interpret all these adjectives unambiguously as factive but they are a minority (8/39 allowing one 'mistake' per speaker, 5/39 counting very strictly). There are, however, no subjects for which these adjectives are unambiguously implicative. For most subjects, we find a mixed pattern that, as expected, contains more implicative judgments for the consonant examples but also a substantial number of implicative judgments for neutral cases. For about half of the subjects the susceptibility to the consonance/dissonance dimension is quasi-linear, for the other half the susceptibility increases the more we get to the consonant end of the scale. As a rough measure of the degree of susceptibility we calculated the difference between the intercept at consonance 0 and that at consonance 100 on the normed scale. According to this calculation (leaving out the unambiguous factive interpreters for whom the difference is of course zero), we get differences between 7.95 and 83.43 (on a scale for 1 to 100), meaning that the difference in the probability that a subject will give an implicative answer for a consonant sentence compared to the probability that (s)he will give an implicative answer for a dissonant sentence is between 7.95% and 83.43%! This is very sizable individual variation among speakers who are not reliably factive. For the speakers at the very high end, the consonant/dissonant factor explains the difference between factive and implicative judgments nearly completely but for other speakers, especially those that do not give consistent factive judgments at the dissonant end of the scale, other factors must play a role (one of these factors being the difference in susceptibility for different adjective classes).

The aim of this experiment was to see whether there are any reliable implicative interpreters. The answer to that question is *no*. That doesn't mean that the differences among speakers that we found might not correlate with sociological variables. We recorded the regional provenance, the age and the gender of the participants. 39 subjects, however, is a very small number to study these variables. For age, we put the subjects in four buckets (see Table 3).

One can argue that unambiguous factive interpreters are less common under

Table 3: Age related differences.

	number of subjects	number of unambiguous factive interpreters
19-30 years	23	4
31-40 years	11	2
41-50 years	2	1
51-60 years	3	1

younger subjects but it seems rather hazardous to draw conclusion from such a small sample. The subjects came from four regions in the U.S. The distribution is given in Table 4.

Table 4: Regional differences.

	number of subjects	number of unambiguous factive interpreters
Northeast	10	4
Midwest	10	2
South	12	1
West	5	0

Again, there is a possible trend but not enough data. As far as gender goes, we had 21 males among them 4 unambiguous factive interpreters and 18 females again with 4 unambiguous factive interpreters among them.

In all the categories, the unambiguous factive interpreters were a minority. However, one possibly important factor we do not have information for, is level of formal education.

As before, there was a difference in the adjective classes. Given we have only 2 adjectives per class, it is not possible to conclude much from the data here but overall the data confirmed the ordering of the classes that we gave before (except for the courage class that was not represented). The trends are clear from Table 5 (in percentages).

*Right* and *wrong* are nearly always factive in all conditions. The *luck* class is interpreted as implicative in most cases when it is in a consonant context, and even in a neuter one but that result might not be significant. The two other classes give rise to a majority of factive interpretations but there is a big difference, especially in the consonant class, between the judgment class and the character class. We have yet to do a formal susceptibility calculation per class but the data is very limited.

Table 5: Judgments per adjective class.

class	dissonant			neuter			consonant		
	yes	no	?	yes	no	?	yes	no	?
Moral	33	0	0	27	0	0	30	2	1
Character	31	0	2	27	3	3	24	7	2.5
Judgment	27	4	2	23	8	2	17	14	2
Luck	27	5	1.5	15	16	2	11	21	1

## 6 Conclusion

The experiments described above show that native speakers of English differ in their interpretation of negative sentences with evaluative adjectives and infinitival complements. They also indicate that this is most likely not a dialect variation although it is possible that age plays a role. We have at this point no good explanation for the difference but the importance of the consonant/dissonant factor suggests that it might be the case that it is not so much a linguistic difference than a difference in the way different language users evaluate the importance of linguistic structure versus the importance of arriving at interpretations that are coherent with their overall beliefs about the state of the world. A similar result was found in a study by Wason & Reich (1979) (See also Cook & Stevenson (2010)). The preference for an interpretation coherent with the previous state of beliefs is most likely more common when there is an additional load on sentence processing such as negation. But it remains astonishing that one finds these effects even in a setting where there is no time pressure or any other factor that would explain degraded performance.

It remains also to be explained why we find the effect not only in understanding but also in production as shown by the corpus data. Here the fact that the construction under investigation is very close to one that can have the intended implicative meaning (exemplified in (2)) may play a role.

## References

- Cook, Paul & Suzanne Stevenson. 2010. No sentence is too confusing to ignore. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, 61–69. Uppsala, Sweden.
- Karttunen, Lauri. 1971. Implicative verbs. *Language* 47(2). 340–358.
- Karttunen, Lauri, Stanley Peters, Annie Zaenen & Cleo Condoravdi. 2014. The chameleon-like nature of evaluative adjectives. In Chris Piñòn (ed.), *Empirical Issues in Syntax and Semantics*, vol. 10, 233–250. CSSP-CNRS.
- Karttunen, Lauri & Annie Zaenen. 2013. Factive investigations. In *Structure and Evidence in Linguistics*. Stanford.

*Annie Zaenen, Brian Hicks, Cleo Condoravdi, Lauri Karttunen & Stanley Peters*

- Meier, Cécile. 2003. The meaning of *Too*, *enough*, and *So...That*. *Natural Language Semantics* 11(1). 69–107.
- Norrick, Neal. 1978. *Factive Adjectives and the Theory of Factivity*. Tübingen: Niemeyer.
- Wason, Peter & Shuli Reich. 1979. A verbal illusion. *Quarterly Journal of Experimental Psychology* 31. 591–597.

## Chapter 41

# Talking about beliefs about beliefs without using recursion

Denise Zijlstra

University of Groningen

Marieke Wijnbergen

University of Groningen

Margreet Vogelzang

University of Groningen

Petra Hendriks

University of Groningen

It has been argued by Roeper and colleagues that second-order beliefs (beliefs about beliefs) can only be represented using an overt recursion device, such as sentence embedding. We investigated this claim in a comprehension experiment with Dutch adults. For sequences of three simplex sentences linked by demonstrative pronouns (e.g., “Computers are intelligent. John thinks that. Mary knows that”), participants frequently accepted a second-order belief (“Mary knows that John thinks that computers are intelligent”). As predicted, they did so more often when the two demonstrative pronouns differed in form (*this* vs. *that*) than when they had the same form (either *this* or *that*). We conclude that second-order beliefs do not require syntactic recursion and can be constructed and understood via pronominal reference.

## 1 Introduction

Children have been found to master *first-order Theory of Mind* (ToM), which is the ability to attribute mental states (such as beliefs) to other persons, by the age of 4 or 5 (e.g., Wellman, Cross & Watson 2001; Wimmer & Perner 1983). One view is that children’s development of first-order ToM depends on their mastery of the grammar of syntactic complementation (de Villiers & Pyers 2002). According to this

language-first view, the development of the syntactic means to embed a possibly false proposition (e.g., “Computers are intelligent”) under a mental state verb (as in “John thinks that computers are intelligent”) or verb of communication (such as *says*) is necessary to mentally represent another person’s false belief. *Second-order ToM*, or the ability to attribute beliefs about beliefs to other persons, is mastered much later and perhaps not before the age of 8 or 9 years old (Hollebrandse et al. 2008). However, if second-order ToM merely involves a second application of ToM, then why are children so much delayed in their development of second-order ToM? Roeper (2007) argues that this delay is due to the additional difficulty of second-order beliefs caused by syntactic recursion.

Roeper’s argument is based on the following observation (2007: 265): suppose we said to you that “John told a lie. The Statue of Liberty was turned upside down,” and we would ask you whether the Statue of Liberty was turned upside down. You would probably answer “no”, because it was a lie that John told. In this case, it is possible to infer that the sentence expressing the false belief is subordinate to the sentence referring to John telling a lie, because “it is possible to convey an embedded meaning across a sentence boundary” (Hollebrandse et al. 2008: 269). Thus, this sequence of two simplex sentences yields the same interpretation as the complex embedded structure “John lied that the Statue of Liberty was turned upside down”. An alternative way of expressing a first-order belief is by using a demonstrative pronoun such as *that* (Hollebrandse et al. 2008; Hollebrandse & van Hout 2015):

- (1) a. Computers are intelligent. John thinks that.
- b. John thinks that computers are intelligent.

Here, *that* in the second clause of (1a) refers to the situation expressed by the first clause of (1a). In other words, the sequence of two simplex clauses in (1a) can effectively make the same claim as an embedded proposition (1b). This means that, to express a first-order belief, a speaker can use a recursive rule system involving sentence embedding, as well as a non-recursive rule system (Hollebrandse et al. 2008; Hollebrandse & van Hout 2015). For second-order (false) beliefs, however, things are argued to be different. Consider the following example:

- (2) Mary told a lie. John told a lie. The Statue of Liberty was turned upside down.

Here, it is much more difficult to make a guess, or inference, about what Mary was lying about. When asked what these sentences assert, all people consulted by Roeper (2007: 265) answered with confusion and uncertainty and mostly concluded that Mary and John told the same lie. The difficulty here lies in putting a guess inside a guess without an overt recursion device such as sentence embedding. However, if we would say to you that “Mary lied that John lied that the Statue of Liberty was turned upside down,” you would understand the second-order belief, because this belief was expressed explicitly by using two sentence embeddings. The same appears to be the case for sequences of three simplex sentences linked by demonstrative pronouns:

- (3) a. Computers are intelligent. John thinks that. Mary knows that.  
b. Mary knows that John thinks that computers are intelligent.  
c. Mary knows that computers are intelligent.

It is claimed (Hollebrandse et al. 2008; Hollebrandse & van Hout 2015) that the third sentence in (3a) can be interpreted as (3c), but not as (3b). On the basis of the patterns of interpretation in (2) and (3), it is argued that the representation of second-order beliefs requires syntactic recursion. Only by syntactically embedding the false belief inside a proposition inside another proposition, a listener will understand that Mary knows that John thinks that computers are intelligent. Thus, in contrast to first-order beliefs, second-order beliefs require an overt recursion device, such as sentence embedding, and are therefore considered to be syntactically different from first-order beliefs (Roeper 2007; Hollebrandse et al. 2008; Hollebrandse & van Hout 2015).

According to the above reasoning put forward by Roeper (2007) and colleagues, the second-order false belief interpretation is not available for (3a) because of the absence of an overt recursion device in this sentence. However, an alternative explanation for the unavailability of the second-order false belief interpretation in (3a) is that this interpretation is strongly dispreferred due to the two identical demonstrative pronouns. When a language has different pronominal forms, these different forms tend to refer to different referents (e.g., personal versus demonstrative pronouns in Dutch and German, see Ellert 2010, and null versus overt pronouns in Italian, see Carminati 2002). According to Diessel (1999), all languages have at least two demonstrative forms that are deictically contrastive and allow speakers to refer to referents nearby versus referents at some distance. For example, Dutch distinguishes between the proximal demonstrative pronoun *dit* ('this'), which is used for near deixis, and the distal demonstrative pronoun *dat* ('that'), which is used for remote deixis. So when the speaker uses two identical demonstrative pronouns, these pronouns can be expected to refer to the same referent. It is thus conceivable that the second-order belief interpretation for the second demonstrative pronoun in (3a) is blocked by the interpretation assigned to the first demonstrative pronoun. If the first pronoun is interpreted as referring to the situation denoted by the first clause, the second pronoun, which has the same form as the first pronoun, is perhaps preferably interpreted as referring to this same situation and hence the third clause is interpreted as representing a first-order belief as well.

In this study, we investigate whether it is possible to represent a second-order belief without an overt recursion device in examples such as (3a). Additionally, we investigate whether it is easier to represent a second-order belief when the demonstrative pronouns differ in form, compared to when the demonstrative pronouns are identical. To investigate these questions, we carried out a comprehension experiment with adult speakers of Dutch.

## 2 Methodology

### 2.1 Participants

29 unimpaired Dutch adults (fourteen men, fifteen women) participated in this study, with a mean age of 22 years old (age range 19-31). They were all students of the University of Groningen or the Hanze University of Applied Sciences.

### 2.2 Materials

To investigate whether it is possible to represent a second-order belief without sentence embedding, we used a referential choice task. In this task, participants read sequences of three simplex sentences. An example is given in (4a). The first sentence expresses a proposition that is, or could be, false. The second and third sentence each consist of a referential subject, a mental state verb (such as *thinks* or *knows*) and a potentially ambiguous demonstrative object pronoun (either *dit* 'this' or *dat* 'that'). The subject of the third sentence is always someone with authority, in (4a) a dentist, to increase the plausibility of the arguably more complex second-order belief. We did this as we were interested in the possibility of the second-order false belief interpretation, rather than in a preference for either the first-order or second-order belief interpretation. If a second-order belief can be represented without sentence embedding, it should be possible for the sequence of simplex sentences in (4a) to receive the same interpretation as the complex sentence in (4b).

- (4) a. Snoep is gezond. Marie denkt dat. De tandarts weet dit.  
'Candy is healthy. Mary thinks that. The dentist knows this.'
- b. De tandarts weet dat Marie denkt dat snoep gezond is.  
'The dentist knows that Mary thinks that candy is healthy.'

The participants were asked to indicate what the underlined demonstrative pronoun in the third sentence refers to. They should indicate their interpretation by selecting an answer out of three answer possibilities: an answer corresponding to a *second-order belief*, e.g., that the dentist knows that Mary thinks that candy is healthy (5a), an answer corresponding to a *first-order belief*, e.g., that the dentist knows that candy is healthy (5b), and an answer corresponding to a plausible but non-mentioned belief, such as that the dentist knows that Mary eats candy (5c). Participants were allowed to select more than one answer.

- (5) a. Dat Marie denkt dat snoep gezond is.  
'That Mary thinks that candy is healthy.'
- b. Dat snoep gezond is.  
'That candy is healthy.'
- c. Dat Marie snoept.  
'That Mary eats candy.'

The second and third sentence in each item contain one of the two demonstrative pronouns *dit* ‘this’ or *dat* ‘that’ and one of the three mental state verbs *denken* ‘to think’, *geloven* ‘to believe’ (both non-factive verbs), and *weten* ‘to know’ (a factive verb). The three verbs were selected for their high frequency. The four different combinations of demonstrative pronouns *dit-dit* ‘this-this’, *dit-dat* ‘this-that’, *dat-dat* ‘that-that’, and *dat-dit* ‘that-this’ were equally divided over the 36 test items in our experiment, resulting in nine items per demonstrative pronoun combination. Additionally, we used all nine possible combinations of the three mental state verbs. Each verb combination was used once with each of the four demonstrative pronoun combinations, and therefore was used four times in total. On the basis of these materials, four different lists were construed such that for each sentence the lists only differed in the pronoun combination used. Each participant only received one list.

The 36 test items were preceded by two practice items, which were identical for each of the four lists. Practice items consisted of a sequence of two clauses, the second one containing a mental state verb and a demonstrative pronoun (e.g., *Het regent. Jan denkt dat.* ‘It is raining. John thinks that.’), and were included to ensure that participants understood the task.

### 2.3 Procedure

All participants were tested individually at the university. Participants received the referential choice task on paper. They were instructed that participation was voluntary and that they could stop the experiment at any moment. As a token of gratitude, the participants received a small reward upon completion of the task.

### 2.4 Data analysis

Three participants were excluded from the analysis because they failed to respond to all items. The responses of the remaining 26 participants were analysed.

## 3 Results

Table 1 shows the number of responses per belief response type on the referential choice task. The most frequently chosen response type was the second-order belief response (326). This type of response was chosen more often than the response type according to which both a second-order belief and a first-order belief are possible (313), which was chosen more often than the first-order belief response (277). As expected, the other belief response and further responses including the other belief were hardly ever selected. On the 936 items in total, participants selected an answer corresponding to the second-order belief 654 times ( $326 + 313 + 3 + 12$ ), an answer corresponding to the first-order belief 603 times ( $277 + 313 + 1 + 12$ ), and an answer corresponding to the other belief 20 times ( $4 + 1 + 3 + 12$ ). Thus, the participants found a second-order belief response to be possible more than two-third of the time.

Table 1: Distribution of responses per response type in absolute numbers.

Response type	Number of responses
second-order belief	326
first-order belief	277
other belief	4
second-order + first-order belief	313
second-order + other belief	3
first-order + other belief	1
second-order + first-order + other belief	12
Total	936

For analysis, we excluded all responses which contained an other belief interpretation. In total, 20 responses were excluded, so that 916 responses remained for analysis. Next, we looked at the effect of the forms of the two demonstrative pronouns on the response type (i.e., whether a second-order belief, a first-order belief or both were chosen, see Table 2). As we used two demonstrative pronouns (*dit* ‘this’ and *dat* ‘that’), there were four different pronoun combinations.

Table 2: Distribution of responses (second-order vs. first-order vs. both first- and second-order belief) per demonstrative pronoun combination in absolute numbers.

Pronoun combination	Second-order belief	First-order belief	Both beliefs	Total number of responses
<i>dat-dat</i>	56	87	84	227
<i>dat-dit</i>	99	50	79	228
<i>dit-dit</i>	69	90	74	232
<i>dit-dat</i>	103	50	76	229
Total	326	277	313	916

A Pearson’s Chi-squared analysis was carried out to investigate the effect of the forms of the two demonstrative pronouns on the participants’ belief responses. The analysis indicated that the form of the demonstrative pronouns significantly influences the participants’ responses ( $\chi^2(6, N = 916) = 41.820, p < .001$ ). Post-hoc pairwise chi-square analyses with Bonferroni corrections were done to compare all combinations of pronouns. The results show that responses to the conditions in which the two demonstrative pronouns were the same (i.e., *dat-dat* and *dit-dit*) did not differ ( $\chi^2(2, N = 459) = 1.791, p = .408$ ). Moreover, the two conditions in which the two demonstrative pronouns were different (i.e., *dat-dit* and *dit-dat*) also showed no dif-

ference ( $\chi^2(2, N = 457) = 0.135, p = .935$ ). All other comparisons showed differences between pronoun combinations (all  $p < .008$ ). So, there is indeed a difference in interpretation of our simplex sentences when the demonstrative pronouns differ in form, compared to when the demonstrative pronouns are identical.

To investigate this difference between pronoun combinations, we analysed the distribution of first-order and second-order belief responses in more detail. McNemar tests show that when the two pronouns were the same (i.e., *dat-dat* and *dit-dit*), participants selected the first-order belief response more often than the second-order belief response ( $p = .003$ ). In contrast, participants selected the second-order belief response more often than the first-order belief response when the pronouns differed (i.e., *dat-dit* and *dit-dat*,  $p < .001$ ).

Table 3: Distribution of responses (second-order vs. first-order vs. both first- and second-order belief) per verb combination in absolute numbers.

Verb combination	Second-order belief	First-order belief	Both beliefs	Total number of responses
<i>denken-denken</i>	34	36	32	102
<i>denken-geloven</i>	45	24	35	104
<i>denken-weten</i>	26	39	36	101
<i>geloven-denken</i>	42	29	32	103
<i>geloven-geloven</i>	47	26	31	104
<i>geloven-weten</i>	39	22	35	96
<i>weten-denken</i>	33	37	32	102
<i>weten-geloven</i>	30	32	41	103
<i>weten-weten</i>	30	32	39	101
Total	326	277	313	916

Finally, although it was not the aim of our study, we wanted to see whether the different mental state verbs had a different effect on the participants' belief responses. Three mental state verbs were used in the study (*denken* 'to think', *geloven* 'to believe' and *weten* 'to know'), resulting in nine different verb combinations. The responses per verb combination are shown in Table 3.

A Pearson's Chi-squared analysis revealed that the choice of verbs did not influence the participants' responses ( $\chi^2(16, N = 916) = 23.452, p = .102$ ).

## 4 Discussion and conclusion

It has been argued that an overt recursion device such as sentence embedding is required for representing second-order beliefs, but not first-order beliefs, and that this explains children's delayed development of second-order ToM compared to first-order ToM (e.g., Roeper 2007; Hollebrandse et al. 2008; Hollebrandse & van Hout

2015). This study investigated the first part of this reasoning, namely whether syntactic recursion is truly necessary for second-order beliefs. Using a referential choice task, we tested whether second-order beliefs can also be represented by a sequence of simplex sentences linked by demonstrative pronouns.

Our study first of all shows that Dutch readers allow a demonstrative pronoun that is the object of a mental state verb in a simplex sentence to refer to a first-order belief referred to by a demonstrative pronoun in the preceding simplex sentence, resulting in the representation of a second-order belief. Thus, it is possible to represent a second-order belief without a syntactic recursion device such as sentence embedding. In our study, participants selected this referential option in more than two-third of the cases. This shows that the representation of second-order beliefs by non-recursive pronominal reference rather than recursive sentence embedding is a viable option in languages such as Dutch. This finding contradicts the earlier claim that overt syntactic recursion is necessary for the representation of second-order beliefs (Roepers 2007). Additionally, this finding sheds doubt on the explanation of children's late development of second-order ToM as resulting from the complexity of syntactic recursion (Hollebrandse et al. 2008).

As hypothesized, participants' selection of a second-order belief response was influenced by the form of the two demonstrative pronouns used. The participants in our study were more likely to choose a second-order belief response if the two demonstrative pronouns differed in form. At the same time, participants were less likely to choose a second-order belief response when the two demonstrative pronouns were identical in form. This finding is in accordance with the view that pronouns that have the same form tend to refer to the same thing, while pronouns that have a different form (such as *this* versus *that*) tend to refer to different things. This feature of pronominal reference may explain why previous studies on the representation of second-order false beliefs, using sentences with identical demonstrative pronouns (e.g., Hollebrandse et al. 2008) incorrectly concluded that it is impossible to express second-order beliefs without syntactic recursion.

In our study, we used three different mental state verbs for introducing the two beliefs in the sequences of three sentences: two non-factive verbs (*denken* 'to think' and *geloven* 'to believe') and one factive verb (*weten* 'to know'), which were evenly distributed across the second and third sentences of the test items. As the first sentence of each item always expressed a proposition that is false or could be false, it might be expected that using a factive verb instead of a non-factive verb in the third sentence, as in "Candy is healthy. Mary thinks that. The dentist knows this." would lead to mainly second-order belief responses. Factive verbs such as *know* presuppose the truth of their sentential complement (e.g., Karttunen 1971), meaning that it is impossible to know something that is false. Thus, if the third sentence contains a factive verb, a first-order belief interpretation (e.g., "The dentist knows that candy is healthy") should be practically impossible. However, the verbs used did not influence the participants' belief responses, suggesting that verb factivity did not play a role in the participants' interpretation of the demonstrative pronouns. Three possible explanations come to mind. One possibility is that the subject of the third sentence

being someone with authority (e.g., a dentist) has a similar effect on interpretation as having a factive verb in the third sentence and, by increasing the plausibility of the second-order belief response across the board, masks potential effects of verb factivity. A second possibility is that readers not always agreed on the falsity of the first sentence (e.g., Candy is healthy). Finally, a third possibility is that the verb *weten* ('to know') perhaps is not strictly factive in ordinary language use and may allow uncertainty regarding the truth of its sentential complement (cf. Hazlett 2010). However, more research is needed to clarify this issue.

To conclude, our study shows that the expression of second-order false beliefs does not require the presence of an overt recursion device such as sentence embedding. Like first-order beliefs, second-order beliefs can also be expressed by a sequence of simplex sentences linked by demonstrative pronouns. This suggests that first-order ToM and second-order ToM are not fundamentally different.

## References

- Carminati, Maria Nella. 2002. *The processing of Italian subject pronouns*. Amherst, USA: University of Massachusetts PhD thesis.
- de Villiers, Jill G. & Jennie E. Pyers. 2002. Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development* 17(1). 1037–1060.
- Diessel, Holger. 1999. *Demonstratives: form, function and grammaticalization*. Amsterdam: John Benjamins.
- Ellert, Miriam. 2010. *Ambiguous pronoun resolution in L1 and L2 German and Dutch*. Nijmegen: Max Planck Institute PhD thesis.
- Hazlett, Allan. 2010. The myth of factive verbs. *Philosophy and Phenomenological Research* 80(3). 497–522.
- Hollebrandse, Bart, Kathryn Hobbs, Jill G. de Villiers & Thomas Roeper. 2008. Second order embedding and second order false belief. In A. Gavarró & M. J. Freitas (eds.), *Proceedings from GALA 2007: Language acquisition and development*, 270–280. Newcastle, UK: Cambridge Scholar Press.
- Hollebrandse, Bart & Angeliek van Hout. 2015. Comprehension and production of double-embedded structures: A kindergardenpath in long-distance dependencies in Dutch school-aged children? In C. Hamann & E. Ruigendijk (eds.), *Proceedings from GALA 2013: Language acquisition and development*, 190–204. Newcastle, UK: Cambridge Scholar Press.
- Karttunen, Lauri. 1971. Implicative verbs. *Language* 47(2). 340–358.
- Roeper, Thomas. 2007. *The prism of grammar: how child language illuminates humanism*. Cambridge, UK: MIT Press.
- Wellman, Henry M., David Cross & Julianne Watson. 2001. Meta-analysis of Theory of Mind development: the truth about false-belief. *Child Development* 72(3). 655–684.

*Denise Zijlstra, Marieke Wijnbergen, Margreet Vogelzang & Petra Hendriks*

Wimmer, Heinz & Josef Perner. 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13(1). 103–128.