Language Diversity Congress Groningen

Computational Issues in Studying Language Diversity: Storage, Analysis and Inference

Mark Liberman

Jan Odijk

Arie Verhagen

Hannah J. Haynie

Gerhard Jäger

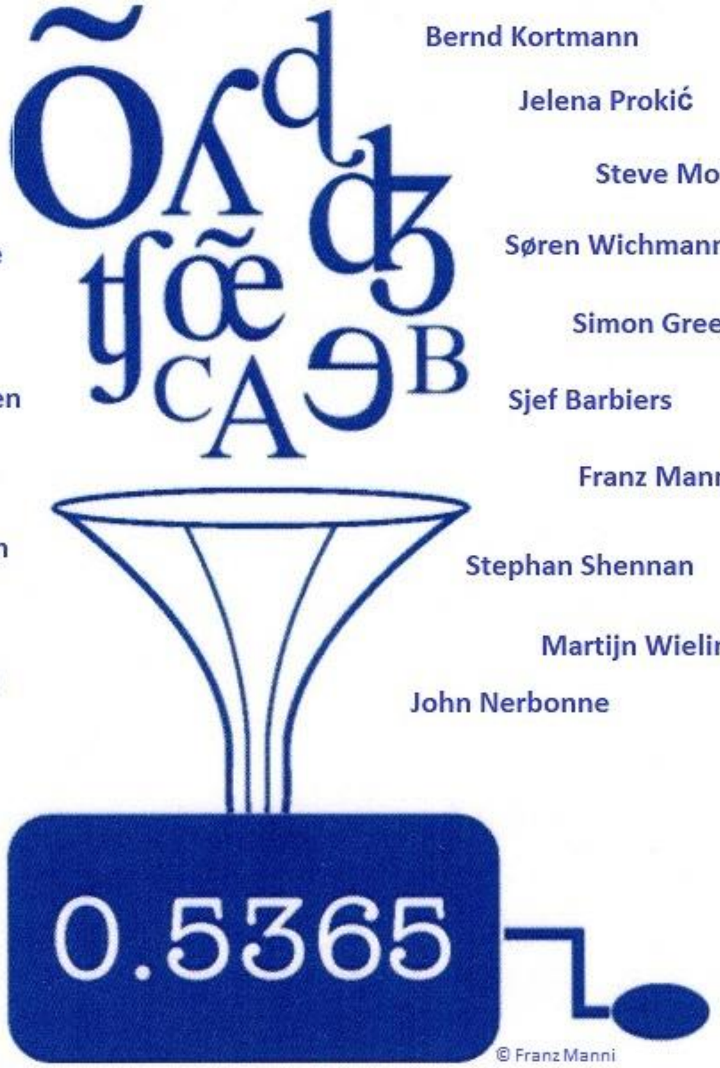
Wido van Peursen

Lars Borin

Jacob Eisenstein

Michael Dunn

Lars Johanson



© Franz Manni

Bernd Kortmann

Jelena Prokić

Steve Moran

Søren Wichmann

Simon Greenhill

Sjef Barbiers

Franz Manni

Stephan Shennan

Martijn Wieling

John Nerbonne

Thurs.-Sat. July 18-20, 2013, Remonstrantse kerk, Coehoornsingel 14, Groningen.

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

CLCG
Center for Language and Cognition Groningen

Book of Abstracts

Patterns of macro- and micro-diversity in the languages of Europe and the Middle East. *Computational Issues in Studying Language Diversity: Storage, Analysis and Inference.*
Groningen, 18-20 July 2013.

Local organisers:
Prof. dr. John Nerbonne
Dr. Karin Beijering
Mets Visser

http://metsvisser.hostei.com/lingcon

Cover image: Franz Manni ©

The conference is sponsored by the The Royal Netherlands Academy of Arts and Sciences (KNAW).

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

Table of Contents

Abstracts Posters

Language Diversity: Computational Issues

Language diversity is of importance both from historical and from cognitive perspectives. A trio of conferences on language diversity will address these issues and also the issue of what sorts of data and algorithms are relevant for the study of language diversity. The conferences are sponsored by the Royal Netherlands Academy of Arts and Sciences' (KNAW). The first conference was held in Amsterdam in Dec. 2012 with the theme 'Patterns of diversification and contact: a global perspective', the second will be held in Groningen on July 18-20, 2013 with the theme 'Patterns of macro- and micro-diversity in the languages of Europe and the Middle East. Computational Issues in Studying Language Diversity: Storage, Analysis and Inference' and the third will be held in Leiden focusing on 'Diversity and universals in language, culture, and cognition'.

The purpose of the conference series is to identify and coordinate the Netherlands' research potential in this area; including both historical and the cognitive dimensions; to sketch ties to neighboring disciplines such as archaeology, ethno-history, cognitive science, anthropology, population genetics, phylogenetics, and literary studies; to contribute further to the development of data-intensive techniques for studying language diversity; to focus the research agenda nationally and internationally, and to identify suitable partners abroad; and to communicate to the scientifically interested Dutch public about developments in this field.

The Groningen Conference on Computational Issues in Studying Language Diversity: Storage, Analysis and Inference

A lot is already known and available about the languages of Europe and the Mideast, including comparative dictionaries of entire language families, attempts at morphological and syntactic reconstruction, corpora of many languages at various stages of historical development (sometimes at great time depths), dialect atlases and dialect dictionaries, typological databases, and naturally, a good deal of archaeological, genetic and cultural information. Some of this material is even available digitally. To-date researchers have studied various relations about the data, but their approaches and research question have varied, and they have normally had to invest a good deal of time in culling relevant information from databases, corpora and word lists designed for other purposes. The challenge – both conceptually and computationally – is to coordinate the different sorts of information to address the historical and cognitive questions. This presupposes a serious interdisciplinary effort to transcend initial incompatibilities in methodology and data analysis traditions. The research questions involve language classification with respect to genealogy, typology and areal influence; language contact and genetics; linguistic history and archaeology; and the sorts of data they can be brought to bear on these questions.

Pieter Muysken, Maarten Mous and John Nerbonne are organizing the conference series, and Nerbonne has the lead on the Groningen conference.

Pieter Muysken
Maarten Mous
John Nerbonne

Programme

Thursday, July 18		
10:00-10:30	REGISTRATION + COFFEE	
10:30-10:45	Welcome, Frans Zwarts, Former Rector, U. Groningen	
10:45-11:00	Introduction, John Nerbonne	
11:00-11:40	Mark Liberman *Documenting Diversity*	Chair:Bouma
11:40-12:20	Jan Odijk *Comparative Linguistic Research in the CLARIN Infrastructure*	
12:20-13:50	LUNCH Remonstrantse kerk	
13:50-14:30	Arie Verhagen *Cross-linguistic variation in usage frequency*	Chair: Verspoor
14:30-15:20	Hannah J. Haynie *Interpreting spatial patterns in linguistic data*	
15:20-15:50	COFFEE BREAK	
15:50-16:30	Wido van Peursen *Morphological encoding of ancient Semitic texts*	Chair: Tjong Kim Sang
16:30-17:20	Jacob Eisenstein *Large-scale analysis of language variation and change in social media*	
17:20-18:30	RECEPTION Remonstrantse kerk	

Friday, July 19		
08:45-09:00	REGISTRATION	
09:00-09:40	Bernd Kortmann *Patterns of typological diversity and areality in the Anglophone world*	Chair: Schmid
09:40-10:20	Jelena Prokić & Steve Moran *Diversity of language sources: challenges in digitization, interoperation and analysis*	
10:20-11:20	POSTERSLAM,COFFEE BREAK + POSTER SESSION	
11:20-12:00	Søren Wichmann *A volcanological view on human linguistic and demographic prehistory*	Chair: Gilbers
12:00-12:40	Simon Greenhill *Quantifying the patterns of language diversification*	
12:40-14:10	LUNCH (at own expense), city centre	
14:10-14:50	Franz Manni *The comparison of linguistic and surname differences in Europe: Three examples.*	Chair: Beijering
14:50-15:30	Stephen Shennan *Demographic continuities and discontinuities in prehistoric Europe*	
15:30-16:00	COFFEE BREAK	
16:00-16:40	Lars Johanson *Four kinds of linguistic distance*	Chair: Zwart
16:40-17:20	Sjef Barbiers *Macroscale microsyntactic variation data, tools and research*	
19:30-22:00	CONFERENCE DINNER	

Saturday, July 20		
08:45-09:00	REGISTRATION	
09:00-09:40	Lars Borin *For better or for worse? Going beyond short word lists in computational studies of language diversity*	Chair: Heeringa
09:40-10:20	Martijn Wieling & John Nerbonne *Inducing and using phonetic similarity*	
10:20-10:50	COFFEE BREAK	
10:50-11:30	Gerhard Jäger *Phylogenetic inference from raw word lists*	Chair: Hoeksema
11:30-12:10	Michael Dunn *Are rates of language diversification correlated with rates of structural change?*	
12:10-12:30	CLOSING WORDS, John Nerbonne	
12:30-13:30	LUNCH, Remonstrantse kerk	

Documenting Diversity

Mark Liberman, University of Pennsylvania

Three main points:

(1) **Access is Crucial.** "Big Data" is transformative, in this as well as other areas. But the transformation requires access to the data, in order to lower barriers to entry and to permit replication; and it requires shared curation, because natural sources of "big data" often start out as a nearly-useless mess. There is an increasingly tendency for access to be restricted to a few researchers, or even to none, based on intellectual property concerns (in the commercial world) or on privacy concerns (in the public sector). It's crucial to fight these trends and to find reasonable compromises permitting appropriate research use. This is NOT equivalent to making all data "open source" or "open access". IPR and privacy concerns can be handled easily in most cases -- if the responsible parties wish to do so.

(2) **New Algorithms = More Data**. New algorithms link existing data instances (e.g. audio/video recordings and transcripts, or texts and lexicons), and create new layers of data annotation (e.g. phonetic or grammatical analysis). The results feed back into the research process, and should be available to the research community along with the original data. This process is working well in molecular biology -- it's time to bring the same practices back to our field(s).

(3) **It's Time to Think Big.** Enormous amounts of relevant data are just out of reach -- three to six orders of magnitude more than researchers can now make use of. Some of this is the result of previous work in dialectology, lexicography, and so on. Much of it comes from non-traditional sources: for example, we can imagine historical text corpora comprising thousands of books a years over 200-300 years in many of the languages under discussion; and the audio archives of broadcasters, political and legal systems, and social science researchers include millions of hours of material. An important potential data source is implicit in the opportunity to use ubiquitous computers and near-universal internet access in creative ways, for instance as pioneered by the Ph@ttSessionz project at the BAS in Munich.

Comparative Linguistic Research in the CLARIN Infrastructure

Jan Odijk, Utrecht University

In this talk I will introduce the CLARIN[1] research infrastructure and show how it can contribute to making micro and macro-comparative linguistic research easier, faster and better (e.g. by broadening the empirical basis). It may even bring the research to a new level by making it possible to address research questions that could not be investigated before and it may give rise to completely new research questions.

I will illustrate this using inflection of attributive adjectives as a concrete example. I will focus on specific phenomena in Dutch but will show that CLARIN will offer similar functionality for the study of these phenomena in other languages as well.

The specific phenomenon in Dutch I will focus on are certain exceptions to the main rule of adjectival inflection in definite NPs, as illustrated in examples such as

1. het bijvoeglijk naamwoord, lit. the adjectival noun, ‘the adjective’
2. het medisch onderzoek `the medical research’
3. de medisch onderzoeker `the medical researcher’

The main rules of adjectival inflection would predict an –*e*-suffix on the adjective in all these cases. The exceptions to this rule appear to be not just arbitrary but to show some regularity, but it is not easy to determine the exact nature of this regularity without a large amount of relevant empirical material. I aim to show how the CLARIN infrastructure can contribute to addressing this and some related linguistic problems.

21 years ago, I could study these phenomena only by using grammars, linguistic articles, and more or less accidental encounters of relevant examples in texts I read. [Odijk 1992]
11 years ago, with the appearance of the Spoken Dutch Corpus[2] and its corpus exploitation tool COREX, it became possible to systematically search for relevant occurrences in a reasonably sized corpus . [Oostdijk et al. 2002]
1 year ago, with the appearance of the SONAR and LASSY corpora, one can search in huge corpora (SONAR has 500 million word tokens; LASSY-Groot is claimed to have more than 2 billion).
Similar resources have become available for other languages, and these resources and their interfaces are also becoming available in the CLARIN infrastructure

For the CLARIN infrastructure we are working towards

- making search in these corpora not only possible but also easy
- making search as independent of the specific form of each resource as possible
- thereby making search possible across corpora of different languages

Though we certainly have not achieved these goals yet, we are working hard to achieve them, and clear progress is being made, as I will show in the presentation.

[1] http://www.clarin.eu/
[2] http://lands.let.ru.nl/cgn/home.htm

In addition, various kinds of micro-comparative databases are being unified and are becoming available in the CLARIN infrastructure, e.g. for Dutch the MIMORE[3] interface to a set of micro-comparative databases of the Meertens Institute, This enables research into these phenomena in regional variants of Dutch.

Finally, I will also show that the CLARIN infrastructure will make it possible for linguists to automatically enrich new data with grammatical information so that a richer resources becomes available for linguistic research.

References

[Odijk 1992] J. Odijk. Uninflected Adjectives in Dutch. In R. Bok-Bennema & R. van Hout, *Linguistics in the Netherlands1992*, pp.197-208. Amsterdam: Benjamins

[Oostdijk et al 2002] N. Oostdijk *et al.* (2002). Experiences from the Spoken Dutch Corpus Project. In *Proceedings of LREC 2002* pp 340-347, Paris:ELRA. [pdf]

[3] http://www.clarin.nl/showcase/mimore/259

Cross-linguistic variation in usage frequency – Testing complex interactions in a simulation model

Arie Verhagen
Leiden University Centre for Linguistics

Languages may differ in the grammatical structures they make available to their users, but also in the relative frequencies with which structural patterns are used. The latter kind of differences are especially visible when two or more languages share structural options of combining elements, but differ in the relative usage frequencies of these options. A case in point are grammatical patterns for the formation of "names for kinds" in German, Dutch and English consisting of combinations of an adjective (A) and a noun (N). Each of the languages has two options: One is a syntactic phrase (*kleine Zehe*, *kleine teen*, *little toe*, respectively), the other is a lexical compound (*Hartholz*, *hardhout*, *hardwood*, respectively). So in terms of the available structures, these languages are the same. However, they differ systematically in the use that is made of these options: German uses the compound pattern the most, English the least, and Dutch occupies a middle position. Moreover, each of these patterns of usage frequencies appears to be relatively stable, which runs counter to the idea that synonymy is not a very stable linguistic state. This suggests that the situation observed is a result of a complex interaction between several parameters, at least some of which have different values, related to other, structural differences, in these languages (Booij 2002, 2009, Hüning 2004, 2010, and references cited there). In fact, Hüning (2004, 2010) suggests that the difference in usage frequencies of these two patterns in German, Dutch, and English may be especially due to the relatively elaborate case system of German. We constructed an agent-based simulation model (Landsbergen 2009) to test a number of versions of this idea, and to determine the role of different factors.

Running this model with a variety of settings for the parameters allows us to draw some conclusions with a relatively high degree of confidence, such as:

- While many instantiations of the two patterns may express the same meaning, *some* degree of (subtle) semantic specialization is necessary in order to reach any equilibrium;
- The presence or absence of a case system plays a less important role in the emergence of different stable equilibria than frequency-related factors (cf. Schlücker & Plag 2011).

Finally, this approach to grammatical synonymy supports a view of the 'isomorphism principle' (one form, one meaning), as far as it holds, as actually a result of competition.

References

Booij, Geert (2002), *The Morphology of Dutch*. Oxford: Oxford University Press.

Booij, Geert (2009), Phrasal names: a constructionist analysis. *Word Structure* 2: 219–240.

Hüning, Matthias (2004), 'Over woorden en woordgroepen. A+N-verbindingen in het Nederlands en in het Duits'. In: Stefan Kiedron & Agata Kowalska-Szubert (Hrsg.), *Thesaurus polyglottus et flores quadrilingues. Festschrift für Stanislaw Predota zum 60. Geburtstag*. Wroclaw: Oficyna Wydawnicza ATUT, 159–171.

Hüning, Matthias (2010), 'Adjective + Noun constructions between syntax and word formation in Dutch and German.' In: Alexander Onysko & Sascha Michel (eds.), *Cognitive Perspectives on Word Formation.* Berlin, New York: De Gruyter Mouton, 195–215 (Trends in Linguistics. Studies and Monographs, 221).

Landsbergen, Frank (2009), *Cultural evolutionary modeling of patterns in language change. Exercises in evolutionary linguistics*. Ph.D. dissertation Leiden. Utrecht: LOT Publications.

Schlücker, Barbara & Ingo Plag (2011), Compound or phrase? Analogy in naming. *Lingua* 121: 1539–1551.

Interpreting spatial patterns in linguistic data

Hannah J. Haynie, Yale University

Geographic patterns arise in the distributions of linguistic material for many reasons, including language contact, language family diffusion, and extra-linguistic factors that ultimately constrain the development of linguistic diversity. With the advance of computational methods for the analysis of linguistic diversity and language evolution, spatial dependencies in linguistic data have emerged as a both a potential complication in the modeling of linguistic relationships and a source of information about linguistic history. This talk reports on techniques for examining spatial patterns in linguistic data at multiple scales and the interpretation of spatial signals as they apply to specific genealogical and areal questions. Case studies from North America and Australia illustrate models for examining geographic patterns in dialect diversity in an ecological context and techniques for assessing the areality of linguistic features. These methods are part of a growing toolset for extracting valuable information from geographic data that might otherwise be considered a source of uncertainty or error in studying language diversity.

Morphological encoding of ancient Semitic texts

Wido van Peursen, Vrije Universiteit Amsterdam

One of the first steps In the creation of an annotated electronic corpus is the analysis at word level. Especially for languages with a poor morphology, this stage seems not very challenging, but only a necessary step towards the final aim: a corpus that can be subjected to all kinds of linguistic analysis. In languages that are morphologically poor, much information (e.g. person, number, gender of the subject of the verb form “walk”) can only be retrieved at clause level (e.g. “they walk” versus “I walk”) or even text level (e.g. when the singular/plural addressee of “you walk” becomes clear).

Many Semitic languages, however, have a rich morphology. Rather than just adding labels to these forms (“tagging”), it is preferable to use a model that does more justice to all the information contained in the morphology (“encoding”). Moreover, since in the case of ancient languages such as Biblical Hebrew, no native speaker is available, a strict “form to function” approach is required. We do not know the functions of forms beforehand; establishing these functions is the purpose of our work.

Not only the functions, but also the morphemes themselves are sometimes hard to grasp. Thus without native speaker it is not always easy to establish the productivity of a morpheme. This uncertainty is reflected in the diversity of decisions taken (most often without justification) in tradtional tools such as grammars and dictionaries.

Finally, morphological variation in manuscripts raises interesting questions as to the the interaction of scribal transmission and errors (e.g. when is it allowed to speak of scribal errors and when should we describe variants in terms of orthographical or morphological variation?)

The paper will present some examples of these challenges and the way they have been dealt with in analysis of ancient Hebrew, Aramaic and Syriac texts in the Werkgroep Informatica Vrije Universiteit (WIVU).

Large-scale analysis of language variation and change in social media

Jacob Eisenstein,
School of Interactive
Computing, Georgia
Institue of Technology

An increasing amount of informal communication is conducted in written form through computer-mediated channels. With the rise of publicly readable social media platforms like Twitter, it is now possible to apply computational methods to investigate language variation on a very large scale. I will describe a series of studies that document lexical variation on Twitter across a number of different social variables. In some cases, this variation tracks spoken language dialects, but we also find that relatively novel "netspeak" terms like emoticons and abbreviations can be strongly affiliated with demographics and geography. Our recent work concerns language change over time, using a new dataset of hundreds of thousands of authors over nearly three years. Aggregating across thousands of words, we build a unified model of the geographic and demographic factors that drive the spread of words between cities.

This research was performed in collaboration with David Bamman, Brendan O'Connor, Tyler Schnoebelen, Noah A. Smith, and Eric P. Xing.

Patterns of typological diversity and areality in the Anglophone world

Bernd Kortmann, University of Freiburg (Germany)

In this talk a European language, English, will be looked at which has gone global ever since the 17th century, yielding a large and typologically diverse set of L1 and L2 varieties, on the one hand, and Pidgin and Creole languages, on the other hand. The perspective taken will be the one of a typologist surveying the morpho-syntactic variation observable in the Anglophone world. The data drawn upon in this talk are exclusively taken from the *Mouton World Atlas of Variation in English (WAVE)*. This atlas maps 235 morphological and syntactic features in 48 spontaneous spoken varieties of English (traditional dialects, high-contact mother tongue Englishes, and indigenized second-language Englishes) and 26 English-based Pidgins and Creoles in eight Anglophone world regions (Africa, Asia, Australia, British Isles, Caribbean, North America, Pacific, and the South Atlantic). The analyses of the 74 varieties are based on descriptive materials, naturalistic corpus data, and native speaker knowledge. Together with its digital companion web-site (http://www.ewave-atlas.org/) WAVE affords unique new insights into the interplay of typological diversity (notably variety type) and areality in accounting for variation in World Englishes. For the purposes of the Groningen conference, the Englishes and English-based Pidgins and Creoles spoken in sub-Saharan Africa will be zoomed in on and discussed on the basis of NeighborNet-based phenetic diagrams.

Kortmann, Bernd/Kerstin Lunkenheimer, eds. 2012. *The Mouton World Atlas of Variation in English*. Berlin/New York: De Gruyter Mouton.

Kortmann, Bernd/Kerstin Lunkenheimer, eds. 2011. *The electronic World Atlas of Varieties of English.* [eWAVE]. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ewave-atlas.org>

Diversity of language sources: challenges in digitization, interoperation and analysis

Jelena Prokić, LMU Munich
Steven Moran, LMU Munich & University of Zürich

In recent decades, the increase in digitally available language data has gone hand--in--hand with the increase in the number of quantitative studies of languages and in studies on dialectal diversity. Language and dialectal data that are available in digital format come from a variety of sources and projects that have often used home--grown solutions for storing and annotating their datasets. From a macro--level viewpoint, these data formats are disparate in encoding, which makes it difficult to undertake automatic analyses or the comparison of different datasets.

In this talk we focus on the comparison of digitally available language data at the phonetic level by a) addressing the problems of automatic preprocessing of data from different sources and b) applying quantitative methods for analyzing data.

In the first part of our talk, we present the 'orthography profile' approach that allows automatic a) cleaning of the data, b) orthographic tokenization, c) graphemic parsing, and d) standardization of the data orthographies. An orthography profile is an empirical description of a source document, where information on the orthography used in the source is interpreted and stored. This information enables a parser to automatically detect errors in the data, parse it into graphemes, return statistics on the graphemes and phonemes in the data and, if necessary, translate into other desired alphabets, such as the International Phonetic Alphabet (IPA) or other phonetic feature representation systems.

In the second part of the talk we show the strengths of this approach on the data coming from various resources on native South American languages that are being digitized as part of the QuantHistLing project at the University of Munich and the University of Marburg. We discuss general and specific challenges in preprocessing the data that comes from very diverse secondary sources. At the end of our talk we present the results of applying quantitative methods on the previously preprocessed data to uncover and clarify phylogenetic relationships between native South American languages.

A volcanological view on human linguistic and demographic prehistory

Søren Wichmann, Max Planck Institute for Evolutionary Anthropology

Since time immemorial volcanoes have inspired awe and fear, in spite of the fact that the number of casualties of volcanic eruptions has only been on the order of 1-2 people per day on average during recorded history—comparable to the number of casualties of terrorist attacks in recent decades. It is beyond doubt that the long-term benefits in terms of mineral nutrients have had a much greater effect, even if this tends to be overlooked.

Within a 200 km radius from Holocene volcanoes, modern population density can be plotted as a function of the distance to these volcanoes. In a parallel fashion, language density is an inverse function of binned distances to the nearest volcano within a 3000 km radius. This suggests that modern language density can be used as a proxy for prehistoric population densities, and that volcanoes throughout human prehistory have constituted one of the factors that have influenced population growth and human dispersal, functioning as an attractor.

Although language density cannot be equated with language diversity, there also seems to be a relationship between language diversity and the geographical concentration of volcanoes. Several of the "accretion zones" of Nichols (1997) are in areas with a high concentration of volcanoes: Caucasus, eastern New Guinea, the Northwest Coast of the Americas, and California, whereas her "spread zones" tend to be free of volcanoes: the Eurasian steppes, North Africa, southern Australia, and the Great Basin. Two cases where roughly adjacent areas are both strikingly different in terms of linguistic diversity and also strikingly different in terms of the volcanoes that they house are: eastern-most Africa (no volcanoes, low diversity) vs. the Nigeria-Cameroon borderland (volcanoes, high diversity) and Borneo (no volcanoes, low diversity) vs. eastern New Guinea (volcanoes, high diversity).

Language density and diversity provide clues to ancient population histories, which, in turn, must at least partly be explained by environmental factors. One of these factors—apparently important, but certainly not unique—seems to be volcanic activity. Other factors which—unlike volcanoes—have already been mentioned in the literature will also be discussed during the talk.

Quantifying the patterns of language diversification.

Simon Greenhill, Australian National University

There is substantial variation in the number of languages in different families. Some families like Austronesian have more than a thousand languages, others like Mayan have around 70, while some families have only a few or even one. The substantial differences in diversity hint that there is major variation in the rate at which languages are born and die. Strikingly there has been little work to quantify these diversification rates. Here I use computational phylogenetic methods on family trees of languages to investigate the pattern of language diversity across families. I will infer the rates of language diversification – and extinction - over time and in different families. The results show substantial variation in time and between different families. I will discuss this variability and attempt to explore its possible causes.

The comparison of linguistic and surname differences in Europe: Three examples.

Franz Manni, Musée de l'Homme, National Museum of Natural History, Paris (France).

I will present an overview of some research about the comparison of biological and cultural diversity in Europe, at the scale of a country. In all cases surnames have been taken as a proxy to genetic variability and dialects, and regional languages, as a proxy to cultural diversity.

To enable a deeper inference of the past genetic variability of human populations in Europe, I have been working to reduce the "noise" that internal migrations—especially those occurred during and after the Industrial Revolution—have created on preexisting patterns of genetic variability, often defacing them.

Surnames, that generally started to be in use before the Industrial Revolution, can tell us which areas experienced more or less migration. In this way, we can say where the present day population is genetically closer to the population of the Middle Ages, which, in turn, is expected to be closer—that is more consanguineous—to the population of more ancient times.

This kind of study aligns well with the moving concepts of identity. Family history, together with geographical, economical, religious, political and social factors, determines our identity and the perception the others have of it. This is why many population geneticists, myself included, consider historical linguistics and socio linguistics as sister disciplines. In fact, linguistic phenomena are also linked to populations and have been conditioned, or caused, by the above mentioned factors. A joint analysis of surname and linguistic differences usually leads to a more reliable insight into the identity of human populations and their history (Darlu et al. 2012).

The computational analysis of linguistic corpora extracted from linguistic atlases, started about twenty years ago, has favored a tighter collaboration among population geneticists and computational linguists, thus making possible the exchange of methodologies, approaches and a better integration of their results. I will present three examples of such joint studies: (*i*) the Netherlands (Manni et al. 2006, 2008); (*ii*) Italy (Boattini et al. 2012) and (*iii*) Spain (Rodriguez-Diaz et al. 2013)

Some research questions remained unanswered, because available linguistic data do not mirror the same demographic processes that geneticists are used to investigate in their discipline. I will review some problematic aspects and suggest ways to obtain a deeper integration of population genetics and (population) linguistics.

References:

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D and F Manni. 2012. General Method to Unravel Ancient Population Structures Through Surnames. Final Validation on Italian Data. *Human Biology* 84(3): 235-270.

Darlu P, Bloothooft G, Boattini A, Brouwer L, Brouwer M, Brunet G, Chareille P, Cheshire J, Coates R, Dräger K, Desjardins B, Hanks P, Longley P, Mandemakers K, Mateos P, Pettener D, Useli A and F Manni. 2012. The family name as socio-cultural feature and genetic metaphor: from concepts to methods. *Human Biology* 84(2):169-214.

Manni F, Heeringa WJ, Toupance B and J Nerbonne. 2008. Do surname differences mirror dialect variation? *Human Biology* 80: 41-64

Manni F, Heeringa WJ and J Nerbonne. 2006. To what extent are surnames words? Comparing the geographic patterns of surname and dialect variation in the Netherlands. Numero special de *LLC Literary and Linguistic Computing* "Progress in Dialectometry: Toward Explanation" 21:507-27

Rodriguez-Diaz R, Manni F and MJ Blanco Villagas. 2013. Usefulness of Isonymy methods to describe the genetic structure of large populations. The case of Spain. Submitted to *American Journal of Human Biology*.

Demographic continuities and discontinuities in prehistoric Europe

Stephen Shennan, University College London

Fundamental to understanding the history of linguistic diversity is the demographic history of the populations of language speakers. The paper will outline some of the results of an ongoing project that has been reconstructing the population history of temperate Europe from Ireland to Poland for the period 8000-4000 BP that covers the introduction of farming. It will show that farming did result in population increase, as always assumed, but that regional increases were not generally maintained. The general pattern is one of demographic boom and bust. These booms and busts can be shown to have consequences for material culture and probably social institutions. It seems reasonable to suggest that they could have had consequences for language patterns as well.

Four kinds of linguistic distance

Lars Johanson (Mainz)

The talk deals with four dimensions of linguistic distance.

Genealogical distance can be demonstrated with the tools of comparative linguistics, regular sound-meaning correspondences. Contact-induced changes do not invalidate the results of genealogical classifications even if extensive copying may make certain relations difficult to recognize.

Typological distance, which may also be used as the basis for grouping of languages and dialects, is in principle independent of genealogical classifications. Even closely related varieties may be typologically different.

Lexicostatistical distance may also serve as a basis for grouping of languages and dialects. The new automatic measuring techniques developed to go beyond the comparative methods yield other results than the latter since they measure distances of other kinds. Methods applicable to lexical databases have been highly successful in dialectometry. The annotation of the data is, however, crucial. Measuring phonetic distances requires transcriptions, since official orthographies are too idiosyncratic to serve as a base for comparisons.

Intelligibility distance has opened a new field of research: efforts to investigate how speakers of different languages understand each other without deliberately engaging in language studies. Interlingual comprehension may be problematic even between closely related or neighboring languages. Native speakers sometimes practice “mother tongue talk in more than one language”.

The relevant data available can be combined to yield novel insights only when these four dimensions are kept apart.

Macroscale microsyntactic variation data, tools and research

Sjef Barbiers – Meertens Instituut and Utrecht University

Recent years have seen a significant increase in the (on-line) digital availability of dialect data for syntactic research, both at the national and the European level (cf. www.dialectsyntax.org for an overview of projects and data collections). To facilitate such research, tools have been developed to search, visualize and analyze the data (e.g. www.meertens.knaw.nl/edisyn/searchengine, www.meertens.knaw.nl/mimore). This infrastructure is quite a step forward and makes it possible to ask research questions that were previously beyond reach. In this talk I will first describe the infrastructure that is currently available and the challenges that one has to face when setting up or contributing to such an infrastructure. I will make clear that there is still a lot to wish for and a lot of work to do concerning data collections, on-line availability and research tools. I will then present some examples of successfull research that was carried out with this infrastructure and finally outline a research program that is made possible by it, focussing in particular on the research program Maps and Grammar (NWO) that will start at the Meertens Instituut this year.

For better or for worse? Going beyond short word lists in computational studies of language diversity

Lars Borin - Språkbanken, University of Gothenburg, Sweden

In order to study language diversity computationally, we need to be able to group individual linguistic behaviors and linguistic products – including entire language systems – into larger aggregates, or categories, by characterizing them as being “the same” or, conversely, “different”. This is equivalent to defining a formal distance measure, which should give us not only the means for determining when two individual language systems or linguistic products should be considered to represent the same language – using some kind of (motivated) threshold – but also for grouping languages in more encompassing categories and placing them relation to each other in some kind of abstract space. Since there are many bases for grouping languages, there are also many potential linguistic distance measures, which could be used alone or in combination to achieve a categorization.

In the field of linguistics, the most commonly encountered method for measuring linguistic distances is based on the use of a small diagnostic word list – a so-called Swadesh list – where the words expressing the same concept in the two languages are compared using some variant of Levenshtein distance (also referred to as edit distance). This methodology has been successfully applied in a large number of studies. Notably, it is at the core of several recent studies purporting to uncover prehistoric movements of proto-language communities from their original homelands, as well as to establish the time depths of the corresponding language families.

In this presentation, I will attempt to elucidate which aspects of linguistic diversity can and cannot be studied using this kind of methodology, and point to alternative and complementary computational tools available from the field of computational linguistics/language technology.

Inducing and using phonetic similarity

Martijn Wieling, Department of Quantitative Linguistics, University of Tübingen, Germany
John Nerbonne, Department of Humanities Computing, University of Groningen

Structuralists famously observed that language is "un systême oû tout se tient" (Meillet, 1903, p. 407), insisting that the system of relations of linguistic units was more important than their concrete content. In the first part of this presentation, we will illustrate how to derive content from relations, in particular phonetic (acoustic) content from the distribution of alternative pronunciations used in different geographical varieties. The material consists of six dialect atlases each containing the phonetic transcriptions of the same sets of words at hundreds of sites. We collect the correspondences via an alignment procedure (i.e. using the Levenshtein algorithm), and then apply an information-theoretic measure, pointwise mutual information, assigning smaller segment distances to segments which frequently correspond. We iterate alignment and information-theoretic distance assignment until both stabilize and we evaluate the quality of the phonetic distances obtained by comparing them to acoustic vowel distances. For all dialect data sets we find strong significant correlations between the induced phonetic distances and the acoustic distances, illustrating the usefulness of the method in deriving valid phonetic distances from dialectal pronunciations.

In the second part of the presentation, we focus on the English accents of hundreds of speakers from across the world. The data source we use is the speech accent archive (http://accent.gmu.edu), which contains phonetic transcriptions of a single paragraph of text (69 words) from a large number of speakers. We use the PMI-based Levenshtein distance to determine the linguistic distance between the average native U.S. English pronunciation and the accented pronunciation of hundreds of non-native English speakers. We validate the computational distances with a perception study and find a good match between perceptual and computational pronunciation distances. These results illustrate the usefulness and applicability of the PMI-based Levensthein algorithm to determine linguistically sensible pronunciation and sound distances.

Phylogenetic inference from raw word lists

Gerhard Jäger
University of Tübingen

The talk addresses the task of inferring a phylogenetic tree of languages from the collection of word lists made available by the Automated Similarity Judgment Project. This task involves three steps: (1) computing pairwise word distances, (2) aggregating word distances to a distance measure between languages and inferring a phylogenetic tree from these distances, and (3) evaluating the result by comparing it to expert classifications. For the first task, weighted alignment will be used and a method to determine weights empirically will be presented. For the second task a novel method will be proposed that tries to minimize the bias resulting from missing data. For the third task, several methods from the literature will be applied to a large collection of language samples to enable statistical testing. It will be shown that the language distance measure proposed here leads to significantly more accurate phylogenies than a method relying on unweighted Levenshtein distances between words.

Are rates of language diversification correlated with rates of structural change?

Michael Dunn, Max Planck Institute, Nijmegen

Language diversification, or the rate of linguistic "speciation", is hard to measure in absolute terms, since our information about past (and even present) languages is often fragmentary, and the evidence for language diversification is also influenced by the rate of language extinction. But bursts of diversification are clearly evident in the branching structure some tree topologies -- for example, the sudden growth of the Oceanic subgroup of Austronesian. An acceptable measure of the rate of language diversification can be inferred from a time-calibrated family tree, as produced by Bayesian Phylogenetic Inference methods, using a range of plausible extinction rates.

The rate of disparification of languages -- the amount of change within a given structural domain within a language -- also varies. The members of a particular subfamily may be more similar or more disparate, for example, in their levels of lexical retention and innovation, variation in morphological and syntactic structures, or phonological complexity. Some elements of language change can be shown to proceed in bursts according to the "Punctuated Equilibrium" model popularized by Dixon (1997). Atkinson et al. (2008) demonstrated this effect statistically for lexical evolution, and the evolution of other elements of language structure is expected to follow this same pattern. One suggested motivating factor for this is that whatever social factors cause language splitting events also drive lexical and other linguistic change. But it has not in fact been shown that the rate of linguistic diversification ("speciation") is correlated with the rate of change in language structure. It has been shown in some biological contexts that rates of species diversification and the rates of structural (in biology, "morphological") change evolve independently, such that rapid diversification can occur with little morphological change, and vice versa (Adams et al. 2009).

In this paper I test the correlation between rate of language diversification and the rate of change in different areas of language structure, including phonology, lexicon, and morphosyntactic organization, and discuss the implications for our understanding of the factors driving the macro-processes of language change.

Obstruents and latitude: acoustic adaptation, thermal influence or spurious statistics?

Damian Blasi[1,2] and Steven Moran[3,4]

1: Max Planck Institute for Mathematics in the Sciences

2: Max Planck Institute for Evolutionary Anthropology

3: University of Zurich

4: University of Marburg

Recently, the study on non-linguistic factors shaping phonological systems has experienced a revamp. Notorious example of this are:

- Negative correlation between languages distances from Africa and a proxy for phonological complexity (Atkinson 2011).
- Increase of phoneme inventory size with population of speakers (Hay and Bauer 2007)
- Positive correlation between altitude where a language is spoken and the number of ejectives it has (Everett 2013)

If the reported trends are a symptom of actual mechanisms of co-variation, then overestimating these results is hard. For instance, they might provide a potential solution to the inherent limitations in the comparative method: our (admittedly partial and not always firm) understanding of population sizes, areas and migrations goes beyond the 10000 ybp usually regarded as the non plus ultra in the field.

However, any of the models and correlations before mentioned has been free of criticisms (e.g. Cysouw et al. 2012, Moran et al. 2012). Given the increasing number of publications in that direction, it is important to ask whether there is any gold standard for the assessment of such statistical patterns.

In this poster we analyze the case study of the startling negative correlation between number of consonants and latitude (Maddieson 2011) and specifically the number of obstruent consonants and latitude (Moran and Blasi forthcoming). We do so with the support of several databases, including WALS (Haspelmath et al. 2008), PHOIBLE (Moran 2012) Ethnologue (Lewis 2009) and CDIC (Olson et al. 1985). We present the results of standard statistical tests and sampling techniques, and we outline some caveats underlying any (negative or positive) automatic interpretation of these.

References

Atkinson, Q. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332.6027: 346-349.

Everett, C. 2013. Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives. PLoS ONE, 8(6).

Haspelmath M. and M. Dryer and D. Gil and B. Comrie. 2008. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. http://wals.info/

Hay, J. and Bauer, L. 2007. Phoneme Inventory Size and Population Size. Language, 83:388400.

Moran, S. 2012. Phonetics Information Base and Lexicon. PhD thesis, University of Washington.

Moran, S. and D. Blasi. Forthcoming. Crosslinguistic Comparison of Complexity Measures in Phonological Systems. In Frederick J. Newmeyer and Laurel Preston (eds), Measuring Grammatical Complexity.

Moran, S., D. McCloy and R. Wright. Revisiting Population Size vs Phoneme Inventory Size. Language, 88(4): 877893.

Olson, J., J. Watts, L. Allison, R. Millemann and T. Boden. 1985. Major World Ecosystem Complexes Ranked by Carbon in Live Vegetation: A Database. U.S. Department of Energy.

Maddieson, I. 2011. Phonological Complexity in Linguistic Patterning. Proceedings of the 17th International Congress of Phonetic Sciences, 2834.

Lewis, M. P. 2009. Ethnologue: Languages of the world sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www. ethnologue. com.

M. Cysouw, D. Dediu, S. Moran 2012. Comment on Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. Science 335, 657

Technology for Endangered Languages Data: The Language Archive

Sebastian Drude, Max Planck Institute for Psycholinguisitcs

Language documentation in the modern sense is concerned with creating lasting records of language in the natural environment by building annotated multi-media corpora, among other digital resources. A crucial point is to ensure that the data are archived in a sustainable way – they ought to be available and usable for years and decades to come, as the basis for further research, educational projects or language revitalization activities. Generally there is yet little awareness of the fact that the data about endangered languages are endangered themselves. This talk presents the activities and solutions being developed at The Language Archive at the Max-Planck-Institute for Psycholinguistics in Nijmegen that aim at providing tools and an infrastructure that supports the creation and long-term archiving of precious language data. By now, The Language Archive is one of the largest world-wide, hosting data on more than 150 languages – among these 70 languages documented in the DOBES (Documentation of Endangered Languages) program 2000–2015. Tools developed by TLA, such as ELAN and the web-based language archiving technology, are now widely used and have an exemplary character in the field.

Finding dialect areas by means of bootstrap clustering

Wilbert Heeringa

Jain and Dubes (1988, p. 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 mean to find groups given a set of local dialects and their mutual linguistic distances. Goebl (1982) introduced cluster analysis in the field of dialectometry (see also Goebl 1984, 1993).

The weakness of cluster analysis is its instability; small differences in the distance matrix may strongly change the results (Jain et al. 1999, Nerbonne et al. 2008). Kleiweg et al. (2004) introduced composite cluster maps, which are obtained by collecting chances that pairs of neighbouring 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 at al. 2008 use clustering with noise and bootstrap clustering to overcome instability.

Both the work of Kleiweg et al. (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 Daan and Blok (1969).

We introduce a new flavour 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 perform a procedure consisting of four steps. First, we randomly select 1000 *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 neighbour clustering since we prefer this method reflecting 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. The number will vary between 0 (never) and 1000 (always). 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.

We apply the procedure to distances in the sound components measured with Levenhstein distance between a set of 86 Dutch dialects. We use material which was collected in the period 2008-2011. Recorded transcriptions of male speakers aged 60 years or older are used, 125 words per speaker.

Figure 1 shows the distances, a network map and an area map.

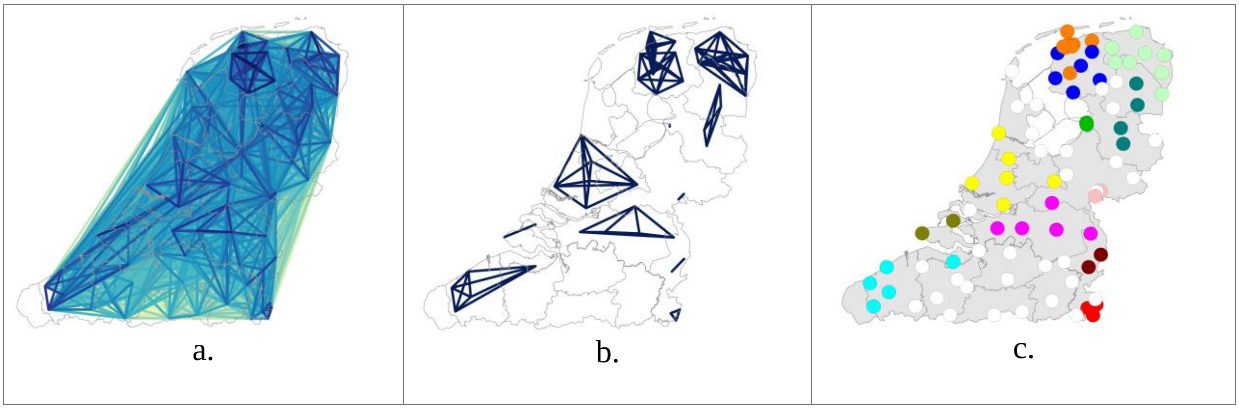


Figure 1. a) the distances: darker lines represent larger distances, b) network map, c) area map.

Literature

Daan, J. and Blok, D. P. (1969). Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde. *Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, volume XXXVII. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Goebl, Hans (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Austrian Academy of Science, Wien.

Goebl, Hans (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Vol. 3, Max Niemeyer, Tübingen.

Goebl, Hans (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In: Viereck, W. (ed.), *Proc. Internat. Congress of Dialectologists*, Vol. 1. Franz Steiner Verlag, Stuttgart. 37-81.

Heeringa, Wilbert (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen Dissertations in Linguistics 46. PhD thesis, University of Groningen.

Jain, A.K. and R. C. Dubes (1988). *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ.

Jain, A. K., M.N. Murthy and P.J. Flynn, Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 264-323.

Kleiweg, Peter, John Nerbonne and Leonie Bosveld (2004). Geographic Projection of Cluster Composites. In: Alan Blackwell, Kim Marriott and Atsushi Shimojima (eds.), *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004.* Cambridge, UK, March 2004 Lecture Notes in Artificial Intelligence 2980. Springer, Berlin. 2004. 392-394.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa and Franz Manni (2008). Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, Reinhold Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society.* Berlin: Springer. 647-654. (Studies in Classification, Data Analysis, and Knowledge Organization)

Syntactic differences among Germanic languages

Wilbert Heeringa, Charlotte Gooskens, Anja Schüppert,
Femke Swarte, Vincent van Heuven

University of Groningen & University of Oldenburg

Sometimes readers are confronted with texts written in a language which is unknown to them. In order to understand the texts, readers will use their knowledge of the languages they are familiar with, especially their mother tongue(s). When the unknown language has many cognates written in a strongly related spelling system, this will ease the understanding of a text.

Another aspect which is likely to play a role in understanding texts in unknown languages is the sentence structure. Words may have 'moved', i.e. occur at a different position in the sentence than expected by the reader. For example:

(1)

```
Dutch   text : Wanneer geen  hulp  gegeven        kan  worden ...
German reader: Wenn    keine Hilfe gegeben werden kann        ...
```

Dutch *worden* corresponds with German *werden*. A native speaker of German expects this word between 'gegeven' and 'kan', but it has 'moved' to the position following on 'kann'.

Words may also be 'added' or 'removed' in comparison to the native language of the reader. For example:

(2)

```
German text  : Das sieht    gut  aus!
Dutch  reader: Dat ziet  er goed uit!
```

A native speaker of Dutch will expect (a German equivalent of) *er* following on *sieht*, but it is lacking in the German sentence.

In our research we aim to model such syntactic variation between closely-related languages. This study takes place in the context of a larger research program which aims to find non-linguistic and linguistic determinants of mutual intelligibility within the Germanic, Romance and Slavic language groups. Intelligibility scores of written and spoken language are obtained with a large-scale web-based experiment. The project aims to find the predictors of these intelligibility scores.

As far as we know, procedures for measuring syntactic distances between two corpora have hardly been developed. Ground-breaking work, however, was done by Nerbonne & Wiersma (2006) (see also: Wiersma, Nerbonne & Lauttamus 2011). They introduced a computational technique for measuring the aggregate degree of syntactic difference between two language varieties. The authors created frequency vectors of *n*-grams (trigrams for example) of part-of-speech tags, and then compared and analysed them using a permutation test, which resulted in both a general measure of difference and a list with the *n*-grams that are most responsible for the difference. The measure was applied to the English of Finnish immigrants in Australia to look for traces of Finnish grammar in their

English.

We apply three syntactic distance measures to a set of five Germanic languages, i.e. Danish, Dutch, English, German and Swedish. An English text of 16 sentences was translated into each of the other languages resulting in five texts in all. The texts will be used as stimuli in the web-based intelligibility experiment. Each of those texts was then translated into each of the other languages as literally as possible. These translations model the knowledge which is likely to be used by the subjects (the readers) in the web-based intelligibility test. For each of the five stimulus texts we obtain four subject texts. When applying the syntactic measures, for each stimulus text we measure the syntactic distance to the four subject texts which are derived from it.

In the first measure we compute the mean logarithmic number of positions that a word in a sentence in the stimulus text has moved compared to the corresponding word in the corresponding sentence in the subject text. In Example (1) *werden* in the German sentence has moved 2 positions forwards in the Dutch sentence where *werden* is translated as *worden*. We call this measure *movement.*

In the second measure we establish the mean number of words in the stimulus text which do not have a counterpart in the subject text (i.e. the number of *in*serted words seen from the perspective of the reader) plus the mean number of words in the subject text that do not have a counterpart in the stimulus text (*del*eted words). In Example (2) we find that Dutch *er* does not have a counterpart in the German sentence, which will be experienced as a deletion by the Dutch reader. We call this measure *indel.*

Our third measure has been developed by Nerbonne & Wiersma (2006). In our data, word classes are coded manually and the distance between a stimulus text and a subject text is found by calculating 1 minus the Pearson's correlation coefficient between the histograms corresponding with the two texts, where each histogram plots the frequencies of the word class trigrams in the text. The significance of the correlation is found by means of a Mantel test. The trigram measure has some advantages over the two measures mentioned above. While both *movement* and *indel* require the aligning of sentences using a procedure which needs to know which word in the stimulus language corresponds to which word in the subject language, this is not required by the trigram measure. Even a parallel corpus is not required when the samples are sufficiently large.

In our investigation we will answer the following questions:

1. How well do the three measures of syntactic distance correlate with each other?
2. When modelling written mutual intelligibility as measured by the web-based intelligibility test, do we need to include all of the three measures?

Preliminary results obtained on the basis of Danish, Dutch, English and German show a significant correlation between movement distances and trigam distances (*r*=0.83, *p*<0.01), but we do not find correlations between indel distances and trigram distances, or between movement distances and indel distances. This suggests that both movement distances and indel distances should be included in the model. We expect that trigram distances can be used as an approximation of movement distances when they can be obtained more easily.

Literature

John Nerbonne & Wybo Wiersma (2006). A Measure of Aggregate Syntactic Distance. In: J.Nerbonne & E.Hinrichs (eds.) *Linguistic Distances* Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006, 82-90.

Wybo Wiersma, John Nerbonne & Timo Lauttamus (2011). Automatically Extracting Typical Syntactic Differences from Corpora. *Literary and Linguistic Computing* 26(1), 2011, 107-124.

Exemplar spaces in corpus-based dialectology and typology

Natalia Levshina

Philipps University of Marburg

Introduction

Exemplar-based semantic maps, which represent the probabilistic semantic space of a semantic and/or formal category as a cloud of exemplars, are a relatively new tool for exploring language diversity. They allow one to detect common and language-specific dimensions of language variation, identify grammaticalization paths, explore how different languages 'carve up' the conceptual space, etc. (Wälchli 2010; Levshina 2012; Wälchli & Cysouw 2012). In this presentation, I would like to propose a unified approach, which can be used for typological, dialectological and register-specific comparisons of lexical and grammatical constructions. The data can be collected from parallel or comparable corpora. The distances between the exemplars can be based on their contextual similarity, in accordance with the distributional approach, or on the surface forms of linguistic expressions, according to the principle of iconicity. Different multivariate techniques, such as cluster analysis or multidimensional scaling, can be used to explore the structure of the data. In addition, the same data can serve as input for obtaining the distances between constructional exemplars, and for measuring the (dis)similarities between languages or lects. For illustration, I will report the results of two case studies, a typological and a variational one, which have been completed recently.

Case study 1. The art of GIVING from the Biblical and Hollywood perspectives

In this quantitative onomasiological study of the semantic field of GIVE in ten genetically diverse European languages I explore a probabilistic exemplar space of GIVE verbs, interpret its common dimensions of variation and compare the form-meaning mapping of language-specific GIVE verbs onto this common space. The data are collected from two parallel corpora of different registers, namely, Bible translations and film subtitles. The parallel corpora are Bible translations (cf. www.paralleltext.info), word-aligned with the help of GIZA++ tool (Och & Ney 2003), and film subtitles, collected from free online resources and sentence-aligned on the basis of timing information (cf. Tiedemann 2003).

A set of verbs of giving was selected with the help of the FrameNet, e.g. *give*, *hand*, *sell*, *bequeath*, *donate*, *pass*, etc. Next, I collected the instances of those verbs in the English segment of the corpora. As a result, 77 multilingual contexts were found in each corpus (154 exemplars in total) and the lexical equivalents of GIVE were coded for each language. The underlying assumption of the approach is that the cross-linguistic similarity between the form of a pair of exemplars can be interpreted as an indication of similarity of the exemplars' meanings (cf. Cysouw 2010). These aggregate similarities between all pairs of exemplars were represented in a distance matrix, which then served as input for Multidimensional Scaling.

The analyses suggest that the semantic space of GIVE is organized along two principal dimensions. The first dimension relates to the relative prominence of the Recipient in the event of transfer. The second dimension corresponds to the distinction between physical and non-physical transfer. As a hierarchical cluster analysis shows, the languages carve up the semantic space in the way that strongly correlates with the genealogical relationships between the languages. Finally, there is strong evidence that the Bible translations exhibit significantly less variation in the expression of GIVE than the subtitles, but at the same time vary along a highly specific dimension related to YIELD category.

Case study 2. Synchronic variation as a time machine: Variation of Dutch causative *doen*

This case study focuses on the Dutch causative construction with *doen* "do" in the Netherlandic and Belgian varieties of Dutch, as in *Je kapsel doet me denken aan een vogelnest* "Your hairstyle makes me think of a bird's nest". The construction expresses causative events that are construed as direct causation (e.g. Verhagen and Kemmer 1997). The data are 731 observations of the causative *doen* from written and spoken corpora of Netherlandic and Belgian Dutch, which represent three registers at different levels of formality (newspapers, Usenet discussion groups and informal conversations). The observations were coded for 35 categorical semantic and formal variables. The structure of the exemplar space was explored with the help of multidimensional scaling and hierarchical clustering with bootstrap.

The results reveal substantial differences in the structure of the lectally specific constructions with *doen*. Most importantly, the cluster structure is more evident in the Netherlandic data, where the main cluster is represented by the construction *doen denken aan* "make think, remind of". Moreover, this cluster is also more prominent in less formal registers. Since the Belgian variety and more formal registers usually reflect a more archaic stage of the development of Dutch, the results tie in well with the hypothesis of the gradual decay and specialization of *doen* as a causative auxiliary (e.g. Speelman & Geeraerts 2009). In addition, a lectometric analysis based on Behavioural Profiles (Gries 2012) of lectally specific *doen* reveals that the national variants of *doen* are more dissimilar in informal spoken Dutch than in more formal registers, in accordance with previous accounts of convergence and divergence of the diaglossic continua in Flanders and in the Netherlands (Geeraerts et al. 1999).

References

Cysouw, M. (2010). Semantic maps as metrics on meaning. *Linguistic Discovery, 8*(1), 70–95.

Geeraerts, D., Grondelaers, S. & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Gries, S. T. (2012). Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. In G. Jarema, G. Libben, & C. Westbury (Eds.), *Methodological and analytic frontiers in lexical research* (pp. 57–80). Amsterdam/Philadelphia: John Benjamins.

Levshina, N. (2012). Let in translation': A typological study of the concept of LETTING in a parallel corpus of film subtitles. Paper presented at the European Summer School in Logic, Language and Information (ESSLLI). Opole, Poland, August 2012.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Speelman, D., & Geeraerts, D. (2009). Causes for causatives: the case of Dutch "doen" and "laten". In T. Sanders & E. Sweetser (Eds.), *Causal Categories in Discourse and Cognition* (pp. 173–204). Berlin/New York: Mouton de Gruyter.

Tiedemann, J. (2007). Improved Sentence Alignment for Movie Subtitles. *Proceedings of RANLP '07.* Borovets, Bulgaria, 2007.

Verhagen, A., & Kemmer, S. (1997). Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics, 24,* 61–82.

Wälchli, B. (2010). Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery, 8*(1), 331–371.

Wälchli, B., & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics, 50*(3), 671–710.

Surname and linguistic diversity of Spain: Towards productive inquiry.

Franz Manni (1), Roberto Rodríguez-Díaz (2) and María José Blanco-Villegas (2)

(1) Musée de l'Homme, National Museum of Natural History, Paris (France) **<manni@mnhn.fr>**
(2) Área de Antropología Física, Departamento de Biología Animal, Facultad de Biología, Universidad de Salamanca (Spain).

In patrilineal societies surnames are transmitted, almost unchanged, from generation to generation. As they are a reliable demographic marker of identity, their variability makes possible the estimation of migrations occurred after their introduction, that is at the end of the Middle Ages. At that time, each surname was "autochthonous" of the province where it started to be in use but later migrations made them spread all over the country. This phenomenon has been uneven, and some provinces remained more isolated than other ones.

According to the 2008 Spanish Census, we have analyzed the geographic variability of more than 33,753 surname types, corresponding to 51,419,788 individuals. Such variability (frequency of each surname type in the 47 Spanish provinces) has been summarized in a pairwise distance matrix between provinces, meaning that the similarity in the surname composition of a given couple of provinces is proportional to the distance we computed (Hedrick's *H*, 1971). Patterns of variability have been represented as *i)* a Multidimensional Scaling plot and as *i)* a barrier dividing neighboring provinces that exchanged less migrants over the time.

One of the goals of the study has been to relate surname diversity to dialect and language diversity in Spain, according to a preliminary dialectometrization of the ALPI (Linguistic Atlas of the Iberian Peninsula) that yields a very accurate cartography of language differences (Goebl 2010). Our surnames-based barrier corresponds to the separation between the dialects of Castilla-Aragon and Andalusia in the South and Catalunia on the East. If the match of the barrier is almost perfect concerning the frontier with Andalusia, the surname barrier does not correspond very well with the frontier between the dialects of the Castillan and Catalan group because it is located more inland and less close to the Mediterranean coast.

It does not seem that linguistic factors have played a constant role in the pattern of surname diversity of Spain, meaning that demographic and cultural phenomena have sometimes been divergent. This preliminary conclusion is challenged by a number of alternative interpretations that we will discuss.

References:

Goebl H. 2010. La dialectometrización del ALPI: Rápida presentación de los resultados. 26th CILFR (*Congrés Internacional de Lingüística i Filologia Romàniques*). 6-11 September, Valencia (Spain).

Hedrick PW. 1971. A new approach to measuring genetic similarity. *Evolution* 25: 276-280.

Marta Meneguzzo (University of Verona)

ONSET CLUSTERS IN STANDARD GERMAN AND IN TYROLEAN DIALECTS

Tyrolean dialects and dialects of Trentino (northern Italy) deploy a rather complex syllable structure not found in the corresponding standard languages (Tyrolean dialects: *[kf]rok*, <gefragt>, 'asked'; *[ks]agt*, <gesagt>, 'said'; *[kʃ]lofen*, <geschlafen>, 'slept'; *[pʃ t]ellt*, <bestellt>, 'booked'; dialects of Trentino: *gra[nt]*, <grande>, 'tall'; *fo[rt]*, <forte>, 'strong'; *fo[rn]*, <forno>, 'oven'; *fia[ŋk]*, <fianco>, 'hip').

I will analyze the onset clusters of such varieties and present the very first results of my PhD research project, which focuses on consonant clusters in the Germanic-Romance linguistic contact area of northern-central Italy. Languages vary to some extent as to the consonant clusters that they allow. The analysis of such peculiatities in varieties which are in contact geographically as well as in languages which do not belong to the same family allows to detect universal aspects of the sonority scale and its language-specific realization. Furthermore, such processes turn out to be a rich soil for answering the question whether languages in contact influence each other by allowing similar consonant clusters.

The empirical basis for the analysis of Tyrolean dialects is represented by Georg Wenker's questionnaires (*Wenkerbögen*, 1888-1923, which have been processing in 2001, cf. www.diwa.info), out of which I have analyzed 150 interviews, in which 20 items containing onset clusters occur. Consonant clusters of standard German will be compared to those of five Tyrolean dialects (Bozen, Bruneck, Mareit, Meran, Salurn). On the basis of the sonority indexes suggested by Parker (2011), sonority distances (SD) will be fixed and consonant clusters will be analyzed from an optimality theoretic point of view, investigating those constraints which play an important role for the issues in question.

(1) Tableau 1: SD: two-member onset clusters of standard German (adapted from Krämer 2009: 146)

/gn/ SD: 3	*< 3 DifSon	FAITH	*< 4 DifSon	*< 5 DifSon	*< 6 DifSon
→ a. [gn]			*		
b. Ø		*			

(2) Tableau 2: SD: two-member onset clusters of standard German (adapted from Krämer 2009: 146)

/kf/ SD: 2	*< 3 DifSon	FAITH	*< 4 DifSon	*< 5 DifSon	*< 6 DifSon
→ a. [kf]	*				
b. Ø		*			

(3) Tableau 3: two-member onset clusters (adapted from Krämer 2009: 146): ranking for Tyrolean dialects

/gn/ SD: 3	*<2DifSon	FAITH	*<3DifSon	*<4DifSon	*<5DifSon
→ a. [gn]				*	
b. Ø		*			

(4) Tableau 4: two-member onset clusters (adapted from Krämer 2009: 146): ranking for Tyrolean dialects

/kf/ SD: 2	*<2DifSon	FAITH	*<3DifSon	*<4DifSon	*<5DifSon
→ a. [kf]			*		
b. Ø		*			

With respect to microvariation and sonority distance, standard German and Tyrolean dialects differ slightly in constraint ranking with regard to the faithfulness constraints (FAITH). Tyrolean dialects are more tolerant than standard German because they allow sonority threshold 2 (*[kf]allen*, <gefallen>, 'fallen'; *[gv]isst*, <gewusst>, 'known') in comparison with 3 ([gn]ade, <Gnade>, 'boy', 'fellow').

The status of sibilants deserves special attention. The traditional analyses for standard German propose an extrasyllabic status motivated by reasons of sonority, whereas sibilants in Tyrolean dialects act as "wildcards", due to the fact that they occur onset-internally (standard German: *[ʃ v]er*, <schwer>, 'heavy'; *[ʃ pR]ung*, <Sprung>, 'jump'; Tyrolean dialects: *[kʃ m]olzen*, <geschmolzen>, 'melted'; *[kʃ t]orben*, <gestorben>, 'passed away'). Romance dialects *not* in contact with Germanic ones show similar onset clusters. In Bolognese dialect, for instance, sibilants occur onset-internally as well (*[mst]ir*, <mestiere>, 'job'; *[tst]an*, <stupido >, 'silly'; *[dstr]ozzer*, <distruggere>, 'to destroy'; *[psk]adaur*, <pescatore>, 'angler', cf. Pascoli 2012: 24; 26).

Linguistic contact is not available as an explanation for the considered languages. In fact, structural loans are not found in the realm of syllable structure in the analyzed linguistic contact area (Tyrolean dialects and dialects of Trentino) as regards consonant clusters. Similar clusters emerge, on the contrary, between varieties which are *not* in contact (Tyrolean dialects and Bolognese dialect).

The present research can (and will) be extended in different ways. First, interviews in the field with the help of questionnaires will contribute to extend the samples to analyze. Secondly, the analysis of consonant clusters will regard not only the onset, but also the coda. Moreover, attention will be devoted to phonological processes which lead to the making up of consonant clusters (epenthesis, deletion, resyllabification). Last but not least, the analysis will focus on the determining of the constraint rankings for the analyzed varieties.

Comparative study of string similarity and vector similarity measures for Bulgarian dialect classification.

Taraka Rama

University of Gothenburg & University of Groningen

June 14,2013

Abstract

Prokic (2010) applied weighted Levenshtein distance (Levenshtein 1965) to the Bulgarian dialect dataset for the task of dialect classification. The vanilla Levenshtein distance assigns uniform substitution cost between a pair of symbols whereas, the weighted Levenshtein distance assigns diferential substitution costs between a pair of symbols. The distance between a pair of dialects is the aggregate of the weighted Levenshtein distance between the pair of words belonging to the same concept. The pair-wise dialect distance matrix obtained from this step is supplied as an input to a standard clustering algorithm. The output of the clustering algorithm is then evaluated through a comparison with the gold standard classification. Prokic (2010) shows that the weighted Levenshtein distance outperforms the vanilla Levenshtein distance. However, the
field of computational linguistics boasts of more than a dozen string similarity measures. [1]

In this poster, I evaluate the performance of the above mentioned string similarity measures for the task of Bulgarian dialect classification. An alternate approach consists of representing a dialect word list as a boolean or a numeric vector of n-grams (extracted from the word list) and the application of one of the various vector similarity measures (Rama & Kolachina 2012) for the purpose of computing a pair-wise dialect distance matrix. At this stage, I evaluate the performance of a vector/string similarity measure through a direct comparison of the distance matrix with the gold standard classification by employing a correlation measure known as point- biserial correlation. The correlation score is always in the range of -1 to 1. The preliminary experiments suggest that there are few vector/string similarity measures which perform at the same level of weighted Levenshtein distance.

References

Levenshtein, V. (1965), 'Binary codes capable of correcting spurious insertions and reversals', *Cybernetics and Control Theory* **10**, 707-710.

Prokic, J. (2010), Families and Resemblances, PhD thesis, Ph. D. thesis, Rijksuniversiteit Groningen.

Rama, T. & Kolachina, P. (2012), How good are typological distances for determining genealogical relationships among languages?, *in* 'Proceedings of the 24th International Conference on Computational Linguistics'.
URL: *http://aclweb.org/anthology/C/C12/C12-2095.pdf*

[1] http://www.coli.uni-saarland.de/courses/LT1/2011/slides/stringmetrics.pdf

Productivity of Dutch verbal inflection patterns

OSCAR STRIK
REMCO KNOOIHUIZEN

Rijksuniversiteit Groningen

Germanic languages have two general types of verbal inflection paradigms: weak verbs, which form past tense and participle forms by means of a dental suffix, and strong verbs, which employ vowel alternations but no dental suffix in these forms. It is generally stated that the weak forms are productive in language change and neologisms to the detriment of strong verbs (cf. Salverda 2006). However, changes from weak to strong forms have been attested in Dutch and other Germanic languages (van Haeringen 1940, among others).

We present an experimental and computational study of the role of analogy in changes in the verbal paradigms of Dutch, inspired by Albright & Hayes (2003). We first present the results of three experimental studies into the use of weak and strong forms:

- an elicitation task asking for past and perfect forms of nonce verbs (cf. van Santen 1997 for existing verbs)
- an acceptability judgment task with weak and strong forms of nonce verbs
- an elicitation task asking for forced strong forms of existing weak verbs

Although confirming the productivity of mainly weak inflections, the experimental results show that certain strong classes, in particular I and II, appear to be relatively productive as well, as is a pattern not previously described for Dutch in which the vowel [o:] is used to mark strong-like past and perfect forms.

Computational analogical modelling using the 900 most frequent forms in Dutch (SUBTLEX-NL, Keuleers et al. 2010) with two different models (Minimal Generalization Learner, Albright & Hayes 2003, and Analogical Modeling, Skousen 1989) confirms the analogical productivity patterns found in the experimental study, underlining once more the role of analogy in language variation and change. The spread of the [o:] vowel cannot be modelled, however, as this change occurs through a different mechanism.

REFERENCES

Albright, A. & Hayes, B. (2003). Rules vs. analogy in English past tensen: a computational/experimental study. Cognition 90. 119–161.

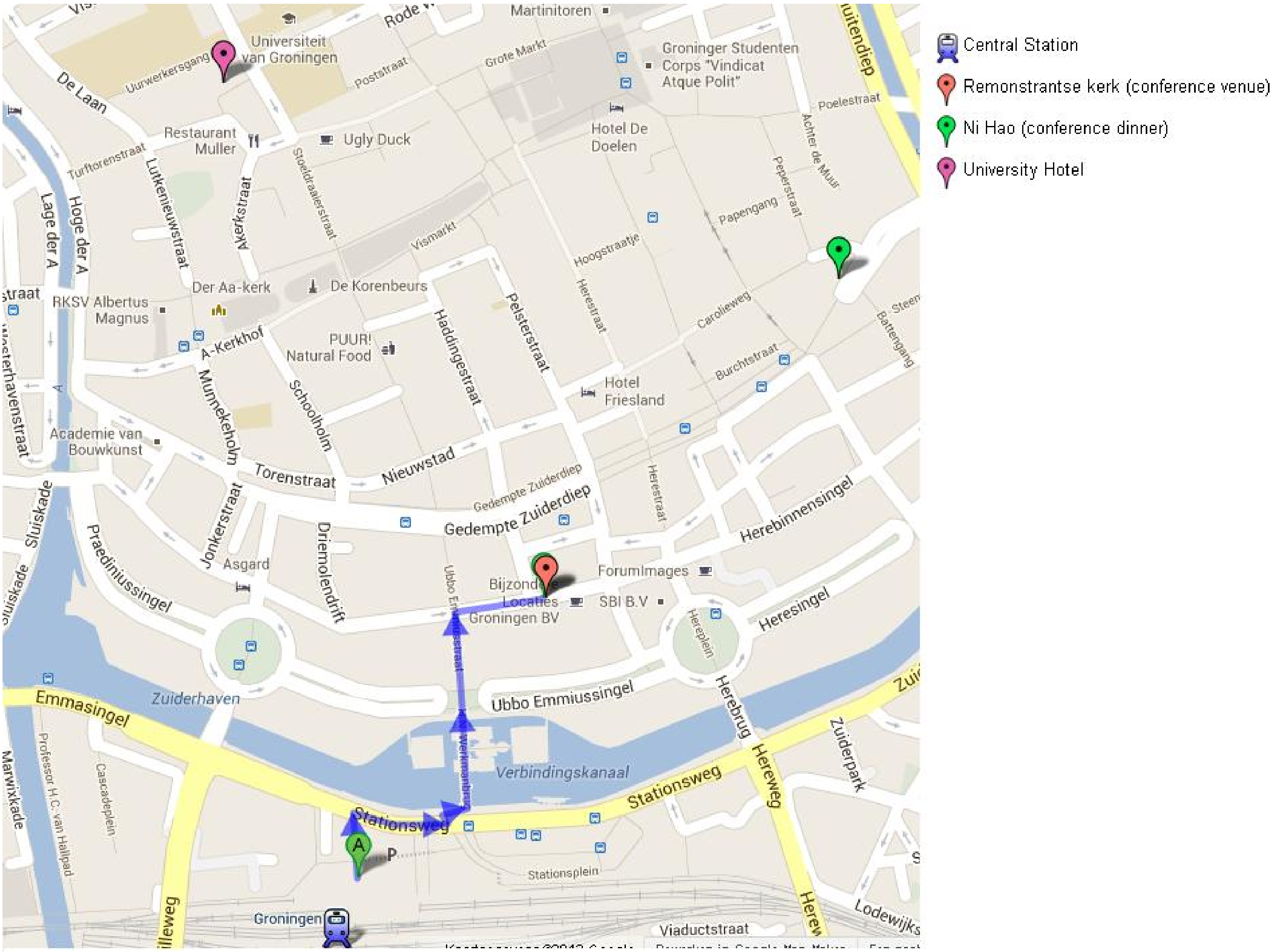
van Haeringen, C.B. (1940). De taaie levenskracht van het sterke werkwoord. De Nieuwe Taalgids 34. 241–255.

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: a new frequency measure for Dutch words based on film subtitles. Behavior Research Methods 42. 643–650.

Salverda, R. (2006). Over de sterke werkwoorden in het Nederlands, Engels en Duits. In Hüning, M. et al. (eds.), Nederlands tussen Duits en Engels. Leiden: SNL. 163–181.

van Santen, A. (1997). Hoe sterk zijn de sterke werkwoorden? In van Santen, A. & van der Wal, M. (eds.), Taal in tijd en ruimte. Leiden: SNL. 45–56.

Skousen, R. (1989). Analogical Modeling of Language. Dordrecht [etc.]: Kluwer.



Conference Location

Remonstrantse Kerk, Coehoornsingel 14, 9711 BS Groningen.

Conference Dinner Location

NI HAO Wok Kattendiep, Gedempte Kattendiep 122, 9711 PV Groningen, on Friday, July 19, 19:30.