Humanities, exactly! / Letteren, exact!

Afscheidscollege

27 januari 2017

door

John Nerbonne
Rijksuniversiteit Groningen

~~~~

oratiereeks Faculteit der Letteren
ISBN 978-90-367-9702-3
~~~~

Dames en heren, collega’s, familie en vrienden,

Alleen bij deze eerste zinnetjes zal ik Nederlands spreken, want er zijn ook niet-nederlandstalige onder ons, en verder hebben sommige collega’s laten weten dat ze graag von mijn rampzalige uitspraak zouden worden ontzien. Ik dank u, nederlandstalige meerderheid voor uw begrip, en ook voor uw onvolprezen tweetaligheid, een echte Nederlandse deugd!

Saying good-bye

Valedictory lectures are for saying *valé* ‘be well, good-bye’ in whatever language we use. It feels a bit hurried to be saying goodbye now, even after more than 24 years in Groningen. But it’s a worthwhile custom that I’m pleased to honor, so let’s get to it.

As I reviewed the valedictory lectures of colleagues, I noticed that they mostly describe research, and I won’t disappoint you on that score, but I’ll later rely on your indulgence as I exploit the opportunity to reflect more broadly and less rigorously not only on what my scientific career has meant, but what the experience has been like. First research, however.

Research

Like lots of late 20th century intellectuals, one of my favorite books while growing up was G.H.Hardy’s *A mathematician’s apology*, where he famously describes the pleasure of intellectual work, but where he also lays out clearly what’s expected from professors in presenting their work.

> “It is one of the first duties of a professor, in any subject, to exaggerate a little both the importance of his subject and his own importance in it” (Hardy, 1940:66).

So enjoy the following section *cum grano salis*.

Research on computational linguistics (CL) in general has taken off enormously in the thirty-five years in which I've worked in it. Even at the beginning of my period in Groningen, I was used to having to explain the field to people, talking about organizing texts, searching in them, sorting them, extracting vocabularies and word frequencies from them, describing the grammars implicit in them, … But you heard five years ago how IBM's Watson defeated the world's best players in the linguistically very sophisticated *Jeopardy!* game.[1] And now virtually everyone in this room has access to several powerful language technology products via the smart phone – minimally the Google search engine and access to Google translate, but probably a text editor with spelling correction, perhaps Siri or Cortana as a spoken language assistant, and an online dictionary with enough intelligence to find dictionary entries for inflected word forms (thus relating the word forms *lying* and *lied* to the lemma *lie*). You're up to your neck in language technology, and judging by use statistics, you love it. When I refer to CL, I'm referring to the science and engineering behind these sorts of applications. The feeling of the practitioner has been excitement, and at times bewilderment about the sporadic pace at which it's all been developed.

> "If I had asked people what they wanted,
> they'd have said faster horses." Henry Ford

I haven't been involved with developing practical applications of CL for the last several years, but Ford's remark conjures up the healthy arrogance that was never too distant, especially in Silicon Valley, the modern apogee of the inventive spirit. I am now and have always been more a scientist than an engineer, but the interplay of the two mind-sets, mixed with Californian commercialism, definitely appealed to me.

My own research has gone well, but it has been much less application-oriented than I expected twenty-five years ago. There are accidents that help to explain my turn from applied to pure research, but, at least at that time, Dutch industry was not interested in this sort of research, at least not if it meant they needed to pay for any of it. My experience agrees with that of non-academic managers among my acquaintances, who

[1] The Wikipedia summary is excellent, https://en.wikipedia.org/wiki/Watson_(computer), and see Ferucci (2012) and the other papers in the special issue of the *IBM Journal of Research and Development* for details.

often hire (part-time) employees to exploit all the subsidies (for *onderzoek* 'research') offered by the state but are more than critical about the value of the (same) state-sponsored activities. The subsidies thus fail to stimulate genuinely.[2] It will be interesting to see whether the current emphasis on "valorization" changes the situation much.

But I promised to sketch some high points of my own research, and Hardy's edict demands it!

I'm proud that my research profile is fairly broad, and all of my work in the last thirty-five years has involved CL in one way or another – in developing grammars for automatic processing; software for detailed semantic analysis; applications for natural-language database query, for a language-based appointment manager, and for interacting with speech recognition, evaluating natural-language processing systems, and computer-assisted language learning. In collaboration with colleagues, PhD students and postdocs I've been involved in work in handwriting retrieval, geo-referencing texts, automatic transliteration and text enrichment. More theoretical work has fo-

[2] I'll let the politically minded decide whether it's creeping socialism or plutocratic cronyism that is destroying motivation. I'll illustrate my frustrations with anecdotes from my own experience that I think are or were symptomatic of a larger problem. The first involved a government grant to support work in answering email automatically. The work was to be done in collaboration with a local company, and our side of it went well – after a few months Gosse Bouma and Tanja Gaustad could identify and correctly answer more than 30% of the email to the largest client of a local "customer contact company". Since the company employed dozens of people for this sort of work, we urged them to fulfill their side of the research contract, albeit belatedly, i.e. to install the software locally and begin testing on new data. The colleagues at the company were polite, even enthusiastic, but always pressed for time; couldn't the work be postponed a bit? After several back-and-forths of this sort, we offered to install the software on their servers ourselves, which they agreed to. We then took over another task, testing things on new data, which continued to look promising, but the colleagues never even made the time to examine closely how well things were going. At our final meeting they explained that they were too pressed, and that incorporating a new module into the existing work chain would be too disruptive. I had no reason to suspect deliberate bad faith in any of this, but the company didn't seem worried about not fulfilling its part of the bargain. A second frustration in pursuing applications involved a multinational information provider that cooperated with the Groningen University Library that the late Dr. Alex Klugkist brought me into contact with. The company wanted to exploit one-page description of books that they catalogued, and I wrote several proposals, each more detailed than the last, about how that could be accomplished. The final proposal was several pages long and specified the commercial software to be used and the additional algorithms to be implemented. When I couldn't find someone to start the project on time, the company cancelled the contract in less than two weeks, took the plans and worked on the topic independently. I don't think this worked out, however. The domain expertise simply wasn't represented. A third involved a grant proposal to *Stichting Technische Wetenschappen (STW)*, and like all failed grant proposals, is subject to multiple interpretations. My impression then was that the partner company was too small to convince STW, but it is the small companies drive innovation in language technology.

cused on computational lexica, especially inheritance-based lexica; several analyses of specific syntactic and semantic phenomena; learning language structure from simple data; and detecting language contact influences. I've worked on over thirty different languages.

Focus on dialectology

But a long list doesn't leave one feeling any more insight, so, having noted that I find all these topics interesting, I want to proceed to the research that I'm best known for, the application of edit-distance measures to pronunciation transcriptions, and especially to dialect data. This research line began with a 1996 student seminar, where the students applied an analysis that Brett Kessler had introduced at a conference in 1995 (Kessler 1995). The students were able to replicate the analysis – which is non-trivial, as it is often difficult to replicate contemporary analyses, especially for students, so the project seemed blessed from the start. They digitized enough data from a large atlas of Dutch, *Reeks Nederlandse Dialectatlassen*, to make the analysis interesting.

For those unfamiliar with traditional dialectology, or who know it only through Voskuil's (1996-2000) *Het Bureau*, I'll sketch a bit of background. For several centuries scholars have remarked that languages often take on local flavors, where the local varieties ("dialects") are not incomprehensible, but clearly different. The scientific study of dialects took off in the late 19th century, when Wenker sent surveys to over 40,000 high school teachers in Germany, most of whom completed the surveys. Dialects were often studied with an eye to local peculiarities,[3] but serious scientific questions were also addressed, e.g., the nature of the geographic distribution of language variation and the role this might play in language change.[4] Still, the public image of dialectology was that of a dusty, archival sort of study with unrelenting fascination

[3] Of the sort that can result in the linguistic equivalent of curiosity cabinets ('*kunst- en rariteitenkabinet*'). I can find these charming, but their scientific utility is very limited.

[4] Wenker was criticized by the neo-grammarians, whose hypothesis of exceptionless sound change was challenged by dialectal evidence. See Bremer (1895) for the challenge and Shirmunski (1962: 79ff) for a summary of the history of the interaction.

for particularities.[5] In addition to studying geographic influences on language variation, modern variationist linguists study social influences as well.

The potential for computational analyses of pronunciation differences was clear given the scholarship in linguistics. Dialectology had been plagued with complaints that it focused on too few features without sufficient justification for their choice, while a computationally implemented comparison opened the door to analyses incorporating large amounts of data. In fact, some theorists had stated very clearly that a view based on the aggregate of differences made more sense,[6] but they hadn't been in a position to realize their ideas in the absence of computational support. Of course I expected to reap the additional advantage of employing computational techniques in digital humanities (DH), not only the ability to process much more data, but also the advantage of very explicit analyses – that are more concrete, better replicable and more easily modified.

Introducing edit distance as a means of comparing pronunciations also provides a lever to solve two major problems in dialectology. First, where earlier work had assiduously catalogued differences at a categorical level (same vs. different), edit distance assigns a numerical value to the difference, one that can sensibly be summed and analyzed in the aggregate. This can be appreciated in an example, comparing the pronunciations of *melk* 'milk' at two different Dutch sites, Grouw, where the pronunciation is [mɔlkə] and Harlem, where it is [mɛlək]. The edit distance algorithm dates to Levenshtein (1966) and finds the least costly set of operations which transforms (or "edits") one string into another. As a side effect, it provides an optimal alignment of the two strings (with respect to the admissible operations). Table (1) shows the alignment resulting from the application of the edit distance algorithm to the two pronunciations we noted above. The edit distance in this case is the sum of the edit

[5] Some professionals warned against this, too. See Coseriu's (1975:50) warning against tendencies toward "atomism" in dialectology.

[6] So Haag (1898) sketched methodology in which lines ("isoglosses") were drawn indicating boundaries between the distributions of individual features and boundaries between dialect areas should be sought where these boundaries coincided, while Shirmunski (1962) was quite explicit in advocating an aggregate view:

> Der Dialekt [...] wird nicht durch irgendein isoliertes und willkürlich ausgewähltes Merkmal, und nicht durch eine Linie als Grenze dieses Merkmals charakterisiert, sondern durch die Gesamtheit der Merkmale, die in ihren Grenzen nicht immer zusammenfallen und zum Teil auch Nachbardialekte mit erfassen.

costs, shown in the last line of the table, which in this case would be three (3). Over the past twenty years we have experimented with a myriad of variations on this algorithm, where Heeringa (2004) alone documents (and evaluates) several hundred. It turns out that if we evaluate 100 words as pronounced and recorded in several different data collection sites, the result is quite reliable (Cronbach's α > 0.9). For many purposes, namely all those where we wish to characterize the variety and not the individual word, a simple version is quite satisfactory. Still, work needs to continue on this front.[7]

1	2	3	4	5	6
m	ɔ	l		k	ə
m	ɛ	l	ə	k	
	1		1		1

Table 1. The edit-distance, or Levenshtein, algorithm calculates the least costly set of operations needed to transform one string into another, in this case a Frisian pronunciation of *melk* 'milk' and one from Harlem. As a side effect, the algorithm provides an optimal alignment of the strings, which is also shown here.

The second major problem traditional dialectology was faced with was the fact that borders between variants ("isoglosses" again) simply did not regularly coincide (or otherwise align), which made it impossible to distinguish dialect areas (or continua) without recourse to arbitrary selection of features. For the purpose of presenting to a Dutch audience, we turn to a *locus classicus* from Bloomfield's *Language* (1933:328) concerning Dutch dialects.

[7] Most of the evaluations of validity (see below) have tested whether aggregate impressions of speakers jibe well with measures of aggregate distance, and this is sufficient for ascertaining the relative dissimilarity of dialects (etc.). But minor differences in algorithms tend to be muffled in aggregate measures. There are tasks such as loan word detection, however, where the measures on individual words are crucial.

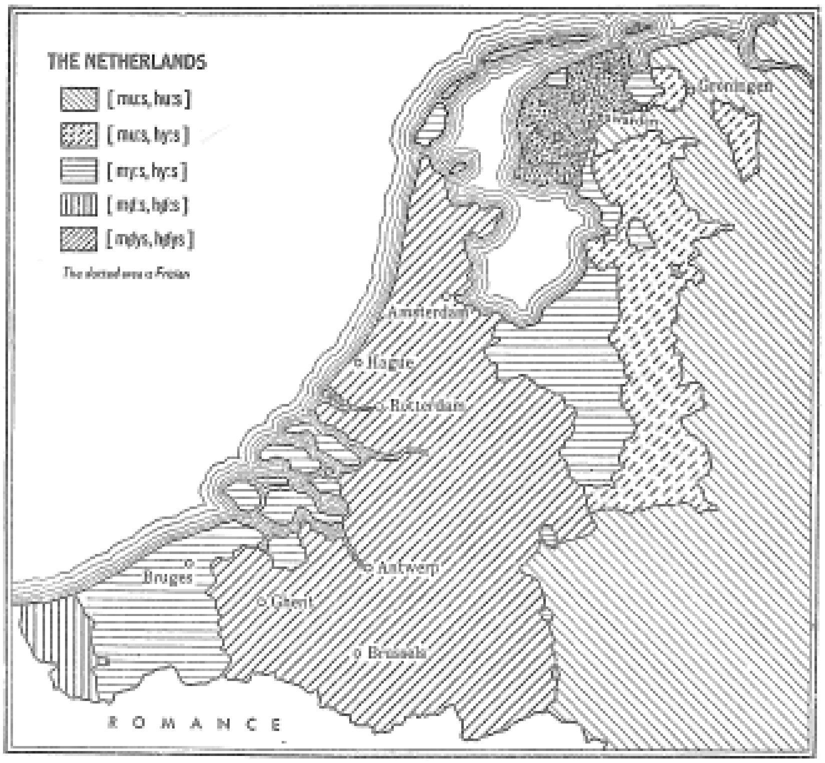

FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Kloeke.

Figure 1. Bloomfield's (1933:328) rendering of Kloeke's observations about the irregularity of historical sound developments. The vowels in the Dutch words for 'mouse' and 'house' developed from the same vowel in the middle ages, but diverged later. An aggregate view of differences need not be concerned with fickle individual developments.

In fact the more abstract view of pronunciation difference unleashed a number of novel perspectives that have since been developed further. By measuring differences and not merely cataloguing them, we were able to apply a powerful statistical technique, multi-dimensional scaling.[8] By including 125 words (with an average length of five sounds per word), we are not forced to choose which elements we think are important. Important sound differences occur more frequently in the sample.

[8] The late Joe Kruskal suggested this in an email around 1998. Wilbert Heeringa and Peter Kleiweg each credit the other for mapping the MDS dimensions to colors.

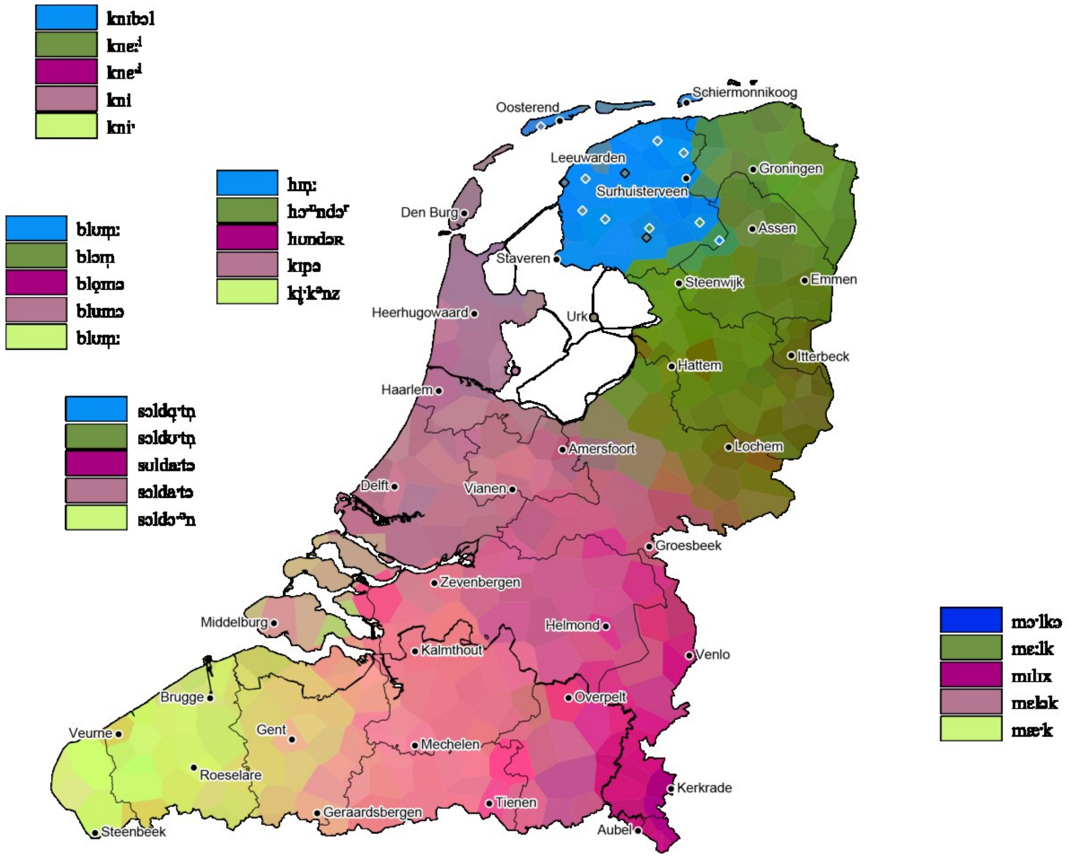

Figure 2. The mean distance of 125 Dutch words from 360 data collection sites in the Netherlands and Flanders was analyzed using MDS (see text), and the first three (most important) dimensions, accounting for over 85% of the variance, were mapped to red, green and blue. Some typical pronunciations are provided in the legends. The differences are geographically abrupt at the Frisan border, but they otherwise tend to be gradual. Following Heeringa (2004).

Wilbert Heeringa (2004) conducted the first extended study using the computational measure, and he and Charlotte Gooskens inspected the reliability and validity of the measure, introducing a more reflective element into general dialectology (Gooskens and Heeringa 2004). Nerbonne and Heeringa (2007) and Nerbonne (2010) exploited the numerical nature of the measure to re-examine theories of diffusion, concluding that earlier "gravity models" overestimated the influence of geographic distance, which was consistently sub-linear.[9] Marco Spruit (2008) examined syntax, enabling the examination of the correlation of lexical, pronunciation and syntactic differences. Spruit also examined the degree to which various syntactic features intercorrelate – a topic worth returning to. Bob Shackleton (2010) probed American and British dialect atlases trying to determine the English sources of American dialects and also showed

[9] And not quadratic as the r^2 in the gravity equation would suggest.

that both geographic distance and traditional dialect areas predict linguistic differences. The combined model is better than either independently.

Therese Leinonen (2010) combined a measure of vowel differences in formant space with dialectometric analysis, and introduced a correlative perspective that provides a striking picture of dialect leveling in modern Sweden (Figure 3), where dialect differences are being rapidly lost, as in the rest of Europe, too. Jelena Prokić (2010) worked on Bulgarian, focusing on the relation between MDS (above) and clustering, which had been introduced to dialectometry earlier, and examining phylogenetic inference algorithms in addition. Peter Nabende (2011) examined an application of measuring pronunciation distance, namely recognizing names from other writing systems (e.g., Urdu or Russian). These are always transliterated in a way that suggests the pronunciation. Martijn Wieling (2012) showed how to make the edit distance measure more sensitive in a data-driven way, and, by using non-linear regression techniques to gauge the influence of geography, showed how to include both geography and social factors into single statistical models. Sandra Hansen (2016) examined the relation between elicited and spontaneous data in Baden.

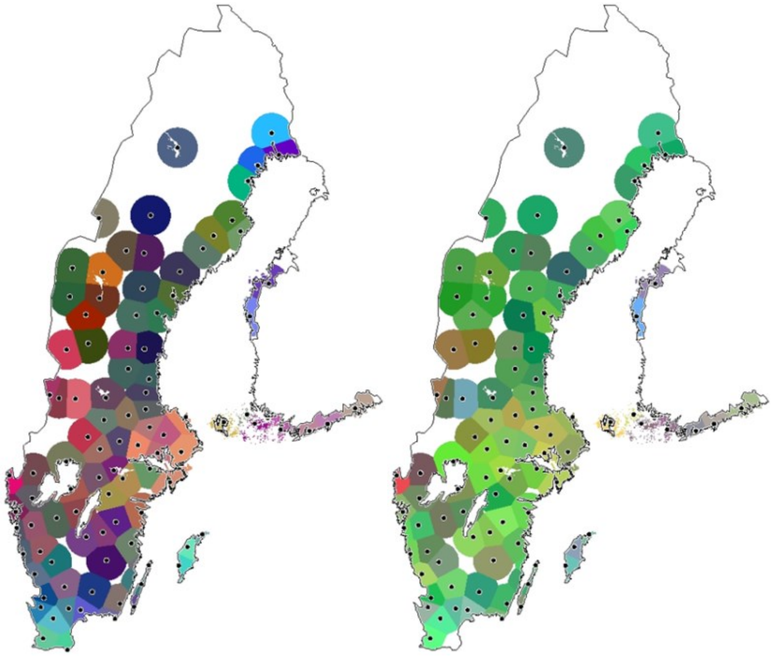

Figure 3. Leinonen (2010) analyzed six pronunciations of each of nineteen vowels from over a thousand Swedish speakers using aggregate differences in vowel quality and subjecting the result to MDS (see above). She then projected the differences from the older speakers (65 years, left) and the younger ones (27 years, right) to the two maps, respectively, dramatically capturing how much variation is being lost in contemporary dialect leveling.

> I not only use all the brains that I have, but also all that I can borrow (Woodrow Wilson).[10]

I've concentrated above mostly on the further achievements in dissertation-length projects, but colleagues in Groningen and elsewhere have collaborated a great deal in other ways, too. First and foremost among these was Charlotte Gooskens who has applied dialectometric methods to the question of the comprehensibility of (closely related) languages.[11] In Groningen some of the other colleagues who got involved in the collaboration were Leonie Bosveld, Çağrı Çöltekin, Bob de Jonge, Peter Houtzagers, Remco Knooihuizen, Sebastian Kürschner, Hermann Niebaum, and Ernst Wit. The colleagues elsewhere included Harald Baayen, Erhard Hinrichs, Bill Kretzschmar, Timo Lauttamus, Franz Manni, Philippe Mennecier, Simonetta Montemagni, Lisa Lena Opas-Hänninen, Petya Osenova, Esteve Valls, and Vladimir Zhobov.[12] It's been great benefiting from all this collaboration!

Macro-linguistics ≠ micro-linguistics

Colleagues, especially in generative and structural linguistics, have asked more than once why this line of dialect research has concentrated so much on large numbers of fairly superficial linguistic elements such as word pronunciations and vocabularies from large numbers of speakers. Shouldn't the work aim rather at comparing more abstract elements of linguistic structure, the various parameters that theoreticians postulate to be sufficient in distinguishing human languages? As Nerbonne (2017, to appear) notes, it is always interesting to examine a problem from many angles, so I don't wish to make the brief against examining dialects with respect to their more abstract "micro-structure", but this hasn't been the focus of this research line.

I have been skeptical that applying micro-linguistic insights and techniques, i.e. those developed to characterize relations *within* varieties – the insights of structural and generative linguistics – ought to carry over uncritically to the characterization of

[10] In a speech to the National Press Club. https://en.wikiquote.org/wiki/Woodrow_Wilson

[11] Charlotte collaborated in turn with Renée van Bezooijen, Jelena Gobubovic, Vincent van Heuven, Jens Moberg, Anja Schüppert, and Femke Swarte. See http://www.let.rug.nl/gooskens/

[12] And I might add Groningen students such as Wybo Wiersma, Bart Alewijnse and René van der Ark, all of whom contributed work that was published.

relations *among* them, macro-linguistics. The structure of molecules such as H_2, O_2, CO_2 or MH_4 is described categorically – in terms of the respective valences of their component atoms which determine their ability to share electrons. Valences are categorical. But the properties of large aggregates of these molecules in gaseous form is standardly described within statistical mechanics (aka statistical thermodynamics, Wikipedia). Language varieties can be characterized internally by categorical laws, much like those of chemistry, but the relations among varieties may best be understood statistically.[13]

Arguments from analogy always fail, so I do not claim to prove anything about the aggregate analyses of linguistic variation from an analogy in statistical mechanics. The point is rather to drive home the insight that aggregate properties are often not simple extensions of individual ones.[14]

The success of analyzing larger samples of language material in order to infer the relations among varieties suggests that dialectology may be in a similar position. In addition, the law-like relations between linguistic dissimilarity and distance, which I've elsewhere dubbed "Seguy's curve" (Nerbonne 2010), depends on adopting an aggregate perspective. Another excellent candidate for a law-like relation would be the relation between linguistic dissimilarity and comprehensibility,[15] how well speakers of different dialects understand one another. This too benefits from an aggregate perspective.

Once one adopts this position, then it is a virtue that one avoids additional assumptions, but not at the cost of explanation, of course. Valls et al. (2012) compare phonetic and more abstract phonological characterizations of dialects, and shows that one needn't incorporate more abstract structure if the goal is to characterize the relations among varieties. At the risk of repetition, I note that I still advocate such stud-

[13] James Burridge (Portsmouth) has work ("How do dialects get their shape?") in which he proposes analyzing varieties via statistical mechanics and sketches a system of equations for doing this. In Burridge's work individual speakers are the units over which one aggregates, not linguistic features, which is what the Groningen work has aggregated over. Burridge takes inspiration from Vitányi's (2013) essay suggesting that Tolstoy's philosophy of history, emphasizing as it does the masses' importance rather than leaders', calls for a construal of history along these lines. Burridge's aggregation over individual speakers seems worthy of further development.

[14] Something we knew from the fallacy of composition in any case. What's true of wholes need not be true of their parts (Hansen, 2015).

[15] See the remarks above on Charlotte Gooskens' work, and also http://www.let.rug.nl/gooskens/.

ies, as the point has not been proven definitively. But I can’t resist recalling the pleasure I felt when Wieling et al. (2012) showed that some aspects of acoustic structure (relative vowel pronunciation) could be induced from large-scale comparisons of alternative dialect pronunciations.

A lot of linguistics can be characterized discretely, as sets of elements combined in potentially complex, even recursive rules. A lot of my work has served to show that the properties of aggregates are better approached statistically rather than discretely.

The future of this research line

As Casey Stengel once quipped, “Predictions are tough – especially about the future.” But there are a few dozen researchers using these techniques, there is software available, and there is a regular conference where this research line is being pursued, the series *Methods in Dialectology*. There are also research questions galore for the next generation. To what extent does linguistic structure influence aggregate differences? Can we develop better measures of syntactic differences? Im morphology, should we measure allomorphy and morphotactics independently? How can we measure allomorphic variation independently of phonetic and phonological variation? Can we automate the detection of these differences well enough to enable corpus-based measurements? Can we bring this social perspective on language into closer contact with the dominant cognitive perspective of linguistics? (Wieling and Nerbonne 2007, Wieling et al. 2014).

Digital Humanities

The connection to digital humanities (DH) is simple: dialectology is a branch of linguistics that enjoys a particular affinity to cultural studies.

I reported on the computational work on dialectology at various professional meetings, but in particular, also at the annual ADHO meetings[16] starting around 2003. The work was welcome there as it represented a strand of computational work that

[16] These were then known as “joint meetings of the Association for Computers in the Humanities and the Association for Literary and Linguistic Computing”.

was already being recognized as fruitful within a traditional humanities field, and it launched a separate, more abstract line of publications reflecting on the place of computation within the Humanities (Nerbonne 2005, 2007a, 2015). Attending the Digital Humanities (DH) conferences, often for the sake of just a couple of sessions on language variation, I was attracted to the sessions on stylometry, and tried my hand at this a bit, too (Nerbonne 2007b, 2013, 2014, Klaußner et al. 2015).

It should be clear based on my remarks on dialectology that I am optimistic about the opportunities for the use of computational techniques in studies such as literature and history, but I won't extend my speculation to the form this is likely to take.

I turn now to other aspects of my career in Groningen and its preparation, and the presentation will ramble a bit.

A biographical interlude

As I noted above, I came to Groningen after eight years of work in industrial research, which explains my earlier interest in applications, and I definitely benefited a great deal from my experience in industry. For example, the attitude toward unforeseen expenses is certainly more relaxed in industry, undoubtedly due to better funding, but it is also more sensible, recognizing that some material expenses can translate into more efficient use of time.[17]

At the same time I do not want to suggest that industry is a promised land where research finally flourishes for its own (inherent and practical) value. Industrial work is fraught with its own set of demanding diversions, things like using the very newest (and otherwise untested) equipment manufactured by your employer, trying out the programming language the next department over has just developed, or contacting development partners in other parts of the company who are willing to invest time and money in research ideas. I left the large and well-established Hewlett Packard

[17] I once wrote a page to Hewlett-Packard Labs about wanting to use a new $20K software package (Refine from Reasoning Systems), noting that it would save me months of work and that the experiment might be useful to others, and that was enough. We acquired the package. I won't outline how much work is normally needed at a university to spend just 10% of that, but I will add that the problem is not just having less money. There is an additional element of tradition that refuses to imagine that new work might require different levels of support.

company in 1990 to join a small research company, DFKI,[18] where I also liked the work and the colleagues, but where the disadvantages of trying to work within a small company run purely on research were also apparent. Facing the prospect of needing to shift foci every few years depending on what was trendy and what larger companies wanted, I was pleased when the position in Groningen opened up. I hadn't realized it then, but I sensed that the freedom to pursue research lines that I found promising would be more worthwhile. Universities are fantastic in this respect.

University instruction

Although I'd worked for a while in industry, my employers had always been generous in allowing me to accept invitations to teach, both at the Stanford Symbolic Systems department and later at the *Universität des Saarlands*, Computational Linguistics Department, so the segue into the instructional duties of a university position was not difficult or even uncomfortable. I genuinely enjoy teaching and the interaction with the students.

Several years ago I compiled a list of courses I taught, and it was varied, including courses in programming, Logic, Linguistics, Statistics and a range of specialized subjects within those, such as string algorithms, feature logics, syntactic theory, dialectology and machine learning. In addition there has been a large number of specialized, project-related courses, which I also enjoy teaching. The temporal sequence of subjects – Logic, Language, Computation and Statistics – has reflected the development of my own interests, which began in philosophy and logic, developing from there to semantics and syntax, turning at the end of graduate school to computation, and about ten years later to statistics and inductive processes. This is development I shared with a large number of students of language of my generation, e.g., Erhard Hinrichs, Mark Johnson and Hans Uszkoreit,[19] and I think it reflects our basic interest in studying language using exact methods, and then also our willingness to take advantage of opportunities as they arise.

[18] Deutsches Forschungszentrum für Künstliche Intelligenz, https://www.dfki.de/

[19] The four of us corresponded about German syntax, especially in phrase-structure models, in the early 1980s. Many others followed similar intellectual paths.

The overall impression of my teaching above may well surprise those who've noted only the last fifteen years of my teaching in Groningen, which has concentrated on statistics. Most of my students would never suspect my beginnings in the heroic age of categorical reasoning that Hans Solo evokes.

"Never tell me the odds!"
– Hans Solo, 'The Empire Strikes Back'

I started teaching statistics nearly twenty years ago, when the linguistics department was dissatisfied with a course elsewhere that they were sending their students to. My department also wasn't offering enough instruction in this field at the time, and the rise in interest in statistical methods for language processing was well underway, so I designed and taught a semester-long introduction, which I've continued in different forms ever since. A specialized seminar for research master's students followed around 2005. I've also enjoyed this work, as a rudimentary understanding of statistics sharpens the critical mind and is also useful professionally. I can teach statistics today with the same conviction with which I taught logic in the 1970s. I feel like I'm continuing the enlightenment ideal of education contributing to the autonomy and articulateness of the individual, the sort of ideal that derived from Kant and inspired Humboldt (Hofmann 2010). A lot of specialized education is linked only indirectly to that goal.

But I confess that I've mentioned Humboldt's ideas partially to provide an occasion to remark on the frequently invoked Humboldtian concept of the university (*Bildungsideal*). Like a lot of older literature, it's cited very selectively, so no one follows Humboldt today in insisting that universities be financially independent (including independent of state support), but where details are offered, then what's cited as worthwhile is the unity between research and instruction (*Einheit von Forschung und Lehre*), best represented as revolving around a researcher and a group of students learning from him.

The unity seems within grasp on good days, say while supervising student projects leading to bachelors' and masters' theses. What needs to be added is that so very much of teaching involves nothing close to research. We're constantly simplifying to keep things understandable, adding motivational – e.g., interactive – elements to teaching programs, keeping careful book on who has done which exercises, validating that exams are comprehensive and fair, etc. When we're planning instruction, we refer constantly, not to the *Bildungsideal*, but rather, e.g., to the semester schedule, the course objectives as these have been approved by the administration, the justification of the number of credit hours, the skills and competences to be developed, alumni advice, and the position of courses with the curriculum, which needs to progress in complexity and difficulty. All of these are sensible, and please don't misinterpret me to be saying otherwise, but we might acknowledge that all these improvements take real time, including our time as scholarly staff, and we might require that the research process become part of the check list of desiderata that we consider during planning curricula and courses.[20]

I'm optimistic that universities will continue to unite research and instruction. Instruction has long since overtaken research as the primary function of universities, but research develops from the curiosity that lots of us share, and will continue, rhizome-like, even under adverse conditions (Deleuze and Guattari 1988), where each part is capable of supporting new organisms.

Management

> "You may not be interested in war, but war may be interested in you." – Leon Trotsky

I've spent a lot more of my time on what is often referred to as "academic self-management" that I would have imagined twenty-five years ago. I hasten to add that I'd always supposed that I'd be involved at low-level management, that I promised to become department chair when I came, and that I assumed those duties enthusiastically. I heard once in the mid-nineties that I was part of a single-digit percentage of

[20] So the Teaching and Examinations regulations stipulate that "students become familiar with the theory and practice of academic research from the very start of their Bachelor's degree programme", but none of the "indicators" against which curricula and courses are measured, mentions becoming familiar with research. I suspect that my colleagues like me have always included an examination of research and its logic somehow, but it would be only correct to include this explicitly. http://myuniversity.rug.nl/infonet/medewerkers/let/onderwijs/examencommissies/oeren/2016-2017/bacheloropleidingen/ba-oer-deela-16-17_eng.pdf

professors in my faculty that actually turned in written reports of the performance reviews which all chair holders are supposed to conduct annually. I like this level of management, trying to see how personal ambitions might more smoothly mesh with the needs of the organization, and it's a natural step to take as one gains in experience.

The game changer for me was an official reorganization in the Faculty of Arts in 1996-1998, which nearly closed the department I chaired. This was unexpected, time consuming and unpleasant, so I won't relate the details at this festive occasion. [21] I became convinced that my department's low political profile was partly to blame, so I resolved during that "reorganizational" process to become more involved in faculty decision making. Notice that this means that, like most academics, I took up academic administration *for all the wrong reasons*. I actively sought the position of Linguistics research director, director of the Center for Language and Cognition, Groningen (CLCG) when it became vacant. I eventually held the position from Jan. 1, 1999 through Aug. 2012, or nearly fourteen years. So even though I took the management work on for the wrong reasons, I definitely enjoyed some aspects of the work.

Being the director of the Groningen research institute CLCG involved representing the interests of the linguistic research of the faculty, which meant responding to is-

[21] But since I'll never say anything about it otherwise, I'll add some details in this footnote. I add them to buttress a proposal that universities make much less frequent use of reorganizations (less than two in twenty-five years). Coming from outside the Netherlands, I hadn't even realized that academic staff with contracts for an indeterminate period *could* be terminated on the grounds that their employing unit needed to "reorganize" (to stay financially solvent). The practices of the Dutch universities show less respect for tenure and job security (than other systems I've seen in action). The US AAUP condones closing entire departments during financial emergencies, but these are not frequently invoked. The German colleagues are *unkündbar*, ignoring malfeasance, and I've never heard of a German academic being fired for financial reasons. My own unit consisted then of myself and four assistant professors, all of whom were serious about teaching and research, and all somewhat successful. Two of them, both very accomplished given the stage of career they'd reached, and the two who I collaborated most closely with, received termination notices, because they'd been hired last. I'm not going to review all the steps taken in the campaign to trying to reverse that decision, all the letters written, all the committee meetings attended, all the conversations and emails with potential allies. Some of these are probably even now confidential, but the politicking occupied an enormous amount of time, not in the least because the process was fairly opaque even to those immediately affected. If the faculty had carried out its plans, more than half of the department, myself included, would have left, but the campaign to reverse the decision on departmental cutbacks was ultimately successful. I think that the fact that we were offered jobs elsewhere was decisive, but this definitely makes one wonder about how serious the original plans were, since our leaving for other jobs would have have had the same financial consequences.

sues involving research, with or without specific invitations to respond.[22] There were usually about 45 faculty members and about 60 graduate students and postdocs in CLCG during my directorate. The work involved reporting annually on the work being done, keeping track of whether minimal publication standards were being met, and consulting regularly with the dean and/or faculty board, about research policies. The "interests" of linguistic research involved issues such as serving on recruitment committees, the evaluation of ongoing research, measures for stimulating research, but also the policies with respect to Ph.D. students whose four-year financing had been exhausted before their dissertations could be completed and defended. With respect to the last group, we once debated for weeks on how to deal with a proposed governmental measure that would have made it incorrect for former Ph.D. students to receive unemployment compensation and to continue to work om their dissertations – a completely ridiculous stricture that would have relegated essentially all the Ph.D. candidates to a "bad faith" category, if they continued to work on their dissertations after the end of their official project. After a colossal waste of committee time, the proposed measure was withdrawn.

But there was also room in this position for innovations that did not require university or national approval. I introduced a regular check on Ph.D. projects after a year, requiring a substantial position paper outlining the central research questions of the thesis, the approach to be taken, preferably a pilot study or pilot experiment, and a sketch of existing work with similar aims. I found it gratifying that the succeeding instance responsible for monitoring the performance of graduate students, the Graduate School of Humanities (GSH), adopted essentially the same procedures.

I may have acquired the reputation of being especially interested in academic management when I acceded to my dean's request in 2004 to head the section of English linguistics. This went fine until the two other professors in the department left – one for a position elsewhere, and one for health reasons. Faced with the wish that a professor continue to chair the department, I again acceded, but only after verifying that an excellent associate professor would assume my duties in Information Science. But "the best laid plans of mice and men often go awry," and my successor in Information

[22] *Gevraagd of ongevraagd …*

Science needed to withdraw after a family member fell seriously ill. I needed to take over again, which left me the head of two departments for several years. Not to be recommended![23]

One of the elements that saved me was the additional organizational support allotted to the CLCG at least during this period. The time allotted was raised minimally in view of the additional responsibilities, and the additional support was a godsend, because it came in the form of extra time for the CLCG secretary and administrative assistant, Wyke van der Meer. Wyke assumed from me tasks such as arranging for visitors to come to Groningen (and handling their travel and lodging needs), coordinating appointments and meetings, organizing flights and hotels. In fact, the years when Wyke assisted me were my most productive in publications and in supervising Ph.D. students. *Wyke, ik sta bij jou in het krijt!!*[24]

And all the rest…

The rest can be summarized briefly. It was a great privilege and a source of satisfaction to me that I could direct a large NUFFIC project with Prof. Venasius Baryamureeba training over sixty computer scientists and awarding nearly twenty PhDs. The curiosity, industry and intelligence of the Ugandan grad students were inspiring. The loyal collaboration of colleagues in Groningen, Nijmegen and Eindhoven as well as those at Makerere (Kampala), Gulu, Mbarara and Kyambogo was essential, in par-

[23] And I'm omitting lots of management and management-like tasks I took on in professional societies such as the Association for Computational Linguistics (ACL) and its European chapter (EACL). See https://www.aclweb.org/archive/officers_new.html, 2002 and 1997-98), both of which I headed at different points, or the Alliance of Digital Humanities Organizations (ADHO, http://adho.org/administration/steering), or its European sub-organization (European Association for Digital Humanities, EADH, see https://eadh.org/people/president-officers). I also served on the advisory board of the *Institut für Deutsche Sprache*, Mannheim (http://www1.ids-mannheim.de/) for ten years, and for shorter, but still multi-year periods on advisory boards for STEVIN (http://over.taalunie.org/organisatie/netwerk/stevin), CLARIN (http://www.clarin.nl/), CLARIAH (http://www.clariah.nl/), and the Swedish GSLT (http://www.gslt.hum.gu.se/). I enjoyed this work just as I did refereeing for over forty journals and countless conferences and evaluating research proposals for about twenty organizations. It broadened my perspectives.

[24] And is there a policy issue lurking here? Yes, at least if one concedes that coordinating appointments, finding lodging, and administering travel costs, etc., can be carried out by non-scientific staff. If this is possible, then it should be more efficient to delegate the activities and costs to non-scientific staff. I won't presume to make proposals to current policy-makers as to exactly how this ought to be managed. I leave them with only the remark that current policy seems to make poor use of existing resources.

ticular, that of Gerard Renardel and Henk Sol, as was the support of the International Office in the person of Erik Haarbrink.

As an international member of staff, my proposals were often countered by objections beginning “Maybe that’s how it’s done in Anglo-Saxon lands/ in the USA and Canada / ....”, essentially a geographical *ad hominem* that one most profitably ignores in favor of a focus on the issues at hand. In general, I found it no serious disadvantage to work as a foreigner at this Dutch university. On the contrary, I have been supported well by the university and by its governing boards. And it has been a privilege to work under faculty boards of recent years, especially because of Ger de Haan and Gerry Wakker, who have been great deans in difficult times. I’m pleased that thanks to them and also to Carel Jansen computational linguistics / alfa-informatica is now part of the department Communications and Information Sciences.

Let me not conclude before praising the colleagues in the Dutch systems! For 25 years I have enjoyed the privilege of attending our (bi-)national conference in computational linguistics, Computational Linguistics in the Netherlands (CLIN), and I’m annually overjoyed to meet a large number of Dutch and Belgian colleagues – both senior and junior – who are eager to argue the pros and cons –technically, linguistically and practically – of various strategies for analyzing language computationally. This stands in contrast to national conferences elsewhere! I was asked recently what the advantages of working in the Netherlands have been, and number one on my list was the presence of a great community seriously interested in CL research. For various profound and accidental reasons, it is often difficult to protect research time effectively enough for sustained work. The Dutch situation ensures enough collegial interest for important work to go on. To that I might add that the level of popular interest in science and research is excellent. The science supplements of the *NRC* and the *Volkskrant* are an excellent indication of that, and it’s hard to find comparable quality elsewhere.

Absolutely essential to my professional well-being in Groningen were discussions with my two favorite colleagues in Groningen, Gertjan van Noord and Gosse Bouma, and later with Johan Bos at the weekly meetings of our *leesclubje*, where we normally, almost always, focused on research – reading and discussing recent articles, sometimes textbooks in areas of emerging interest, and often our own ideas – ripe or unripe. It's been great to have Gregory Mills, Malvina Nissim, Barbara Plank, and Martijn Wieling strengthen the reading club! Leonie Bosveld and George Welling were indispensable in instruction and in broadening our instructional offerings, enabling a better and more interesting cooperation with the colleagues in Communication, with whom we eventually fused. I always had the feeling that we respected one another, and I'm grateful to you all for that!

My wife Ellen has supported me in innumerable ways in the last quarter century here, e.g., by having foreign grad students over on St. Stephen's day (Dec. 26), by accepting visitors and entertaining them, not in the least by reminding me – frequently – that there's more to life than work. I'll say today that she's been right more than I usually admit. But … it might be better if you kept that to yourselves and didn't pass that on to her!

Closing

Valete, Groningen!

Literature

Bremer, Otto (1895). *Beiträge zur geographie der deutschen Mundarten: in form einer Kritik von Wenkers Sprachatlas des deutschen Reichs* (Vol. 3). Breitkopf & Härtel.

Bloomfield, Leonard (1933) *Language*. New York: Holt, Rinehart and Winston.

Coseriu, Eugenio (1975) *Die Sprachgeographie*. Tübingen: Gunter Narr.

Deleuze, Gilles and Félix Guattari (1988) *A thousand plateaus: Capitalism and schizophrenia*. London: Bloomsbury.

Ferrucci, David A. (2012) "Introduction to 'This is Watson'" *IBM Journal of Research and Development* 56 (3-4): 1-1.

Gooskens, Charlotte and Wilbert Heeringa (2004) "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data"*Language Variation and Change* 16(3): 189-207.

Haag, Karl (1898) *Die Mundarten des oberen Neckar- und Donaulandes*. Reutlingen: Buchdruckerei Egon Hutzler.

Hansen, Hans, "Fallacies" *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2015/entries/fallacies/

Hansen-Morath, Sandra (2016) *Regionale und soziolinguistische Variation im alemannischen Dreiländereck. Quantitative Studien zum Dialektwandel.* PhD Diss. Albert-Ludwigs Universität, Freiburg.

Hardy, G.H. (1940, 1990) *A Mathematician's Apology*. New York: Cambridge University Press.

Heeringa, Wilbert J. (2004) *Measuring dialect pronunciation differences using Levenshtein distance*. PhD Diss. Univ. Groningen. hdl.handle.net/11370/6ff6bbca-842f-4a90-9c6e-a0d3cce748da

Hofmann, Jürgen (2010) "Welche Bedeutung hat das Humboldt'sche Erbe für unsere Zeit?" 225. *Veranstaltung der Humboldt-Gesellschaft am 08.01.10*. http://www.humboldtgesellschaft.de/inhalt.php?name=humboldt (consulted 12 Dec. 2016)

Kessler, Brett (1995) "Computational dialectology in Irish Gaelic"*Proc. 7th Conf. European Chap. Association for Computational Linguistics*. Morgan Kaufmann Publishers Inc.

Klaussner, Carmen, John Nerbonne and Çağrı Çöltekin (2015) "Finding characteristic features in stylometric analysis" *Digital Scholarship in the Humanities* 30.suppl 1: i114-i129.

Leinonen, Therese (2010) *An acoustic analysis of vowel pronunciation in Swedish dialects*. PhD Diss., University of Groningen. hdl.handle.net/11370/51096135-2965-4d55-882e-4f078cd52057

Levenshtein, Vladimir I. (1966) "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet physics doklady* 10(8):707-710.

Nerbonne, John (2005) "Computational Contributions to the Humanities" *Literary and Linguistic Computing* 20(1), 25-40.

Nerbonne, John (2007a) "Crosstalk in humanities computing" *International Journal of Humanities and Arts Computing* 1.2: 85-96.

Nerbonne, John (2007b) "The exact analysis of text" Foreword in: Frederick Mosteller and David Wallace. *Inference and Disputed Authorship: The Federalist Papers*. Stanford: CSLI. ^{3}2007.

Nerbonne, John (2010) "Measuring the diffusion of linguistic change" *Philosophical Transactions of the Royal Society B: Biological Sciences, 365* (1559): 3821-3828.

Nerbonne, John (2014) "Review of J.Pennebaker: *The Secret Life of Pronouns. What our Words Say about us*" *Literary and Linguistic Computing*, 139-142. DOI:10.1093/llc/fqt006.

Nerbonne, John (2015) "Die Informatik als Geisteswissenschaft" *Zeitschrift für digitale Geisteswissenschaften* 1. Avail. at http://zfdg.de/die-informatik-als-geisteswissenschaft. Constanze Baum & Thomas Stäcker (eds.) Grenzen und Möglichkeiten der Digital Humanities. Sonderband der *Zeitschrift für digitale Geisteswissenschaften.* DOI: 10.17175/sb001_003

Nerbonne, John (2016) "Rezension von W. Abraham and E. Leiss (eds.) *Dialektologie im neuen Gewand. Zu Mikro-/Varietätlinguistik, Sprachenvergleich und Universalgrammatik. Linguistische Berichte,* Sonderheft 19" *Zeitschrift für Dialektologie und Linguistik* 82: 88-92.

Nerbonne, John (to appear, 2017) "Strukturelle quantitative Dialektologie" In: Elisabeth Leiss and Sonja Zeman (eds.) *Die Zukunft von Grammatik – Die Grammatik der Zukunft. Festschrift für Werner Abraham anlässlich seines 80 Geburtstages.* Tübingen: Stauffenberg.

Nerbonne, John and Wilbert Heeringa (2007) "Geographic distributions of linguistic variation reflect dynamics of differentiation." In: Sam Featherston and Wolfgang Sternefeld (eds.) *Roots: Linguistics in search of its evidential base*. 267-297.

Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten and Willem van de Vis (1996). "Phonetic distance between Dutch dialects". In: *CLIN VI: Proceedings 6th CLIN meeting*. Antwerp: Centre for Dutch Language and Speech (UIA). 185-202.

Nabende, Peter (2011) *Applying dyanamic Bayesian networks in transliteration and detection*. PhD Diss. Univ. Groningen. hdl.handle.net/11370/2eb9dbf9-0e9c-4b8c-a3e2-410b30215b16

Prokić, Jelena (2010) *Families and resemblances*. PhD Diss. Univ. Groningen. hdl.handle.net/11370/be920d0e-2f88-417a-89d2-7704f962b9e4

Schirmunski, Viktor M. (1962). *Deutsche Mundartkunde. Vergleichende Laut-und Formenlehre der deutschen Mundarten*. Berlin: Akademieverlag.

Shackleton Jr, Robert G. (2010) *Quantitative assessment of English-American speech relationships*. PhD Diss. Univ. Groningen. hdl.handle.net/11370/b8a69e64-7f7f-4643-98b4-aa0097c5cf20

Spruit, Marco René (2008) *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD Diss. University of Amsterdam.

Valls, Esteve, John Nerbonne, Jelena Prokic, Martijn Wieling, Esteve Clua, and Maria Rosa Lloret (2012) "Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches" *Verba: anuario galego de filoloxía, 39*: 35-61.

Vitányi, Paul (2013) “Tolstoy’s mathematics in *War and Peace*” The Mathematical Intelligencer 35(1):71-75.

Voskuil, J.J. (1996-2000) *Het bureau*. Vol.1-7. Amsterdam: Van Oorschot.

Wieling, Martijn B. (2012) *A quantitative approach to social and geographical dialect variation*. PhD Diss. Univ. Groningen. http://hdl.handle.net/11370/cd637817-572f-4826-98c1-08272775fb64

Wieling, Martijn, Eliza Margaretha and John Nerbonne (2012) "Inducing a measure of phonetic similarity from pronunciation variation" *Journal of Phonetics* 40.2: 307-314.

Wieling, Martijn and John Nerbonne (2007) "Dialect pronunciation comparison and spoken word recognition" *Proc. of the RANLP Workshop on Computational Phonology*. 71-78.

Wieling, Martijn, John Nerbonne, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa and R. Harald Baayen (2014). "A cognitively grounded measure of pronunciation distance" *PloS ONE*, *9*(1): e75734.

Wikipedia (2017) "Statistical mechanics" https://en.wikipedia.org/wiki/Statistical_mechanics