

Final Report

Determinants of Dialect Variation

NWO 360-70-120, Sept. 2010

A. General

1. a. Researchers and organisations

Principal investigator of Determinants of Dialect Variation (DDV)

Prof. dr. John Nerbonne, Center for Language and Cognition (CLCG), Groningen,

Researchers:

Dr. Wilbert Heeringa, postdoc, CLCG

Ms. Christina Siedle, Ph.D candidate, CLCG (10/2004- 9/2005)

Dr. Therese Leinonen, Ph.D. candidate, CLCG, (4/2006-3/2010)

Dr. Marco Spruit, Ph.D. candidate, Meertens, (11/2003-10/2007)

Co-applicant

Prof. dr. Hans Bennis, Meertens Institute

Others

Prof. dr. Sjef Barbiers, Meertens, Spruit's supervisor

Prof. dr. Hermann Niebaum, CLCG, Siedle's supervisor

Dr. Charlotte Gooskens, CLCG, Leinonen's supervisor

Prof. dr. Vincent van Heuven, Leiden, Leinonen's second promotor

1.b. Project development and subsequent careers

The project began late in 2003 after Wilbert Heeringa's Ph.D. thesis was accepted. Marco Spruit, a Ph.D. student, was able to begin at roughly the same time at the Meertens Institute. Christine Siedle, came to the project nearly a year later. Unfortunately, life in Groningen turned out not to agree with her, and she left after one year although we were more than satisfied with her work. We were most fortunate in being able to attract Therese Leinonen to the project in early 2006, which entailed a shift of interest from German to Swedish, and a substantial investment in the study of acoustic techniques for analyzing vowel quality (see below).

Prof. John Nerbonne, PI, led the overall project and also the work in Groningen, where Prof. Hermann Niebaum, a specialist in German, was also involved as Christine Siedle's primary supervisor. When Siedle left the project, Niebaum, deferred to Dr. Charlotte Gooskens, who also acted as primary supervisor and *co-promotor* for Therese Leinonen. Prof. Hans Bennis led the project at the Meertens institute, and Prof. Sjef Barbiers served as Marco Spruit's immediate supervisor and *co-promotor* (Bennis and Nerbonne were first and second *promotores*).

The project employees have continued their scientific careers. Dr. Heeringa successfully applied for an NWO Veni grant, and moved to the Meertens institute to carry out the project.

Dr. Spruit accepted a lecturer position (*universitaire docent*) at Information Science, Utrecht. Therese Leinonen currently works as a postdoc on a CLARIN-NL project developing a web application for software developed in DDV, and she's accepted a position at the *Svenska litteratursällskapet i Finland* (Society of Swedish Literature in Finland, www.sls.fi) as a postdoctoral researcher. Ms. Siedle works for a digital cartography company in Düsseldorf.

2. Final Report of Main Program

a. Scientific Aims and Results

The overarching research question of the project was formulated (see proposal, "Executive Summary"):

What determines dialectal variation?

The question was made more concrete by formulating three more specific objectives:

1. quantification of the degree to which the linguistic levels pronunciation, lexis and syntax correlate
2. development and application of a measurement for aggregate syntactic distance between varieties
3. validation of phonetic distance measures within a new language, German

The first objective was achieved by Wilbert Heeringa in a three studies, one of which was a collaborative effort with Marco Spruit and the principal investigator (Spruit et al. 2009). The surprising result is discussed in detail in the presentation of the postdoc project (below).

We tackled the problem of measuring and assessing syntactic variation in exactly the way specified in the project proposal. The work was conducted on the *Syntactische Atlas van de Nederlands Dialecten* (SAND), and surprised one of the project members, who had publicly predicted that one would see little geographic structure in the data. See below for more details, including a "best young scholar award" prize awarded to Spruit in 2005.

The third objective first involved German resulting in Nerbonne & Siedle (2005), in the leading journal for German dialectology, which further validated the pronunciation comparison. When Siedle left for personal reasons, we shifted focus to Swedish, whose dialectology was well-known to Therese Leinonen, Siedle's replacement. This also led us to explore acoustics more than we'd expected, and to involve not only Dr. Charlotte Gooskens, Groningen, expert in Scandinavian languages, but also Prof. Vincent van Heuven, Leiden, expert in phonetics.

Determinants

We also devoted time to asking the overarching question of what influences dialect variation. The most influential earlier work on this was Trudgill's 1974 paper introducing his famous "gravity model", in which the influence of varieties is predicted to decrease quadratically (in distance) but to increase directly in proportion to population size. He buttressed this by tracing individual changes as they diffused. Our innovative step was to shift the view from individual features – such as the use of a new pronunciation – to the aggregate of differences. This, we argue, allows the research to examine the entire field of variation as a residue of the diffusion of all linguistic changes. The postdoc project (Sec.3) describes two published studies from this perspective (Nerbonne & Heeringa, 2007; Heeringa et al. 2007).

Manni et al. (2008) involved a collaboration with population geneticists, in which we investigated whether there might be an influence of family on dialect variation. We show that genetically related groups indeed speak using similar linguistic varieties, but this is most likely due to the fact that they tend to settle nearby.

b. Coherence of program

The range of co-publications and publications in common venues (special issues of journals) suggests a good degree of collaboration, which in fact was even more extensive. Dr. Wilbert Heeringa deserves special mention for his role in instructing the PhD candidates and keeping the project on an even keel. We also acknowledge that we sacrificed opportunities for easier collaboration by organizing the project at two institutes, CLCG and Meertens. But while we missed some easy “on-the-shop-floor” exchanges of the sort one-site projects enjoy naturally, we remain unremittingly enthusiastic about the collaboration, which combines Groningen’s strengths in computational and quantitative methods with the Meerten’s strengths in overseeing the range of data and in syntactic theory.

c. Added value of programmatic attack

The opportunity to approach the questions programmatically allowed us to (i) increase the confidence in the new methods, esp. in the Ph.D. project on pronunciation, and in the first-ever measurement of the influence of geography on syntax; even while we (ii) demonstrated that the quantitative approach enables dialectology to approach some deeper question of causation in novel ways. Finally, it was indispensable in (iii) allowing us to compare the influence of geography on different linguistic levels.

d. Evaluation

Both the program as a whole and also the postdoc project and the Ph.D. project exceeded the goals of the original proposal and were resounding successes. The Ph.D. project on pronunciation was also very successful, but in an unforeseen direction (acoustic phonetics).

e. Publications “Synthesizing Results”

Although the project originally envisioned a monograph as a synthesizing publication, the scientific climate has come to value articles as the more valuable contribution due to the refereeing they are normally subject to, and their greater focus. The three synthesizing articles were first, Nerbonne (2009a), “Data Driven Dialectology” which makes the case for the viewing variation from an aggregate perspective. This is published in *Language and Linguistics Compass*, a new journal offering primarily survey and programmatic articles. Spruit et al. (2009) synthesized work on lexical, pronunciational and syntactic variation, and appeared in *Lingua*. Nerbonne & Heeringa (2009) surveys techniques for measuring dialect differences, some of which emerged from the current project. It appears in a Mouton-De Gruyter handbook organized by Jürgen Erich Schmidt and Peter Auer.

f. Incidental

Naturally the day-to-day work led to incremental improvements in dialectometric technique as well which we note briefly here. Heeringa et al. (2006) surveyed the quality of string comparison algorithms. Wieling, Leinonen & Nerbonne (2007) initiated a new line of research attempting to induce the pronunciation difference between individual segments (e.g. between [i] and [e]) from a large corpus of Dutch dialect pronunciations. Wieling, Heeringa & Nerbonne (2007) applied dialectometric analysis to the data collected by Goeman, Tældeman & van Reenen in compiling the MAND and FAND collections. Nerbonne et al. (2008) examine the value of bootstrap clustering in dialectology, and also introduced noisy clustering, which gives roughly the same results (with low levels of noise) as the bootstrap.

See 3.a.-b. for scientific fora and international involvement.

3. International Perspective.

a. Contacts, etc. DDV was aggressive in seeking opportunities to present its work and subject it to the criticism of others. We began the project with a kickoff meeting in Groningen on Jan. 7, 2004, where Hans Goebel and Edgar Haimlerl (Salzburg) and Georges de Schutter (Antwerp) were invited speakers. Goebel is the world's authority on dialectometric technique, and De Schutter perhaps the leading expert in Dutch dialectology.

Together with Bill Kretschmar (Georgia, USA), a coordinator of American Dialect Atlas projects, we organized a special session at *Methods in Dialectology* XII, Moncton, Canada, "Progress in Dialectometry" in Aug. 2005 which about 50 researchers attended, and which resulted in a special issue of *Literary and Linguistic Computing* 21(4), 2006. The contributors included Heeringa, Spruit, Gooskens, Nerbonne, but also Goebel and Haimlerl (Salzburg), William Kretschmar, Cichocki (New Brunswick, Canada), Cynthia Clopper (The Ohio State University) and John Paolillo (Indiana, USA). We include a copy of this special issue with this report.

Together with Franz Manni (*Musée de l'Homme*, Paris), editor of *Human Biology*, we organized a small session at Digital Humanities 2006, Paris. This resulted in a special issue of *Lingua* 119(11), featuring the Spruit et al. (2009) article, on the degree to which different linguistic levels correlate. See the subproject on syntactic variation (below) for more details. A copy of the special issue is enclosed with this report.

Together with Erhard Hinrichs, Tübingen, we organized a more technically oriented satellite workshop at the joint COLING-ACL meeting in Sydney, July 23, 2006, *Linguistic Distances*. This workshop called for contributed papers on measuring linguistic distances, was deliberately broader in scope than dialectology, but it nonetheless attracted work on using string distances for identifying cognates (by Grzegorz Kondrak and Tarek Sherif, Alberta, Canada) and several proposals for measuring syntactic distance. There were 35 participants from Germany, Australia, Rumania, Ireland, the Czech Republic, India, Japan, the UK, Canada and Israel. We include a copy of the proceedings with this report.

Together with Grzegorz Kondrak (Alberta, Canada) and T. Mark Ellison (Western Australia) we organized a satellite workshop on *Computing and Historical Phonology* on June 28, 2007 at ACL 2007 in Prague. As we have only a single printed copy of this report, we refer interested readers to the online version at the ACL anthology, <http://aclweb.org/anthology-new/W/W07/#1300>, which we've mirrored at <http://www.let.rug.nl/alfa/Prague/proceedings.pdf>. Highlights were talks by Brett Kessler, Washington University, St. Louis, and Brian Joseph, The Ohio State University (then editor of *Language*). There were participants from the US and Canada, Germany, the UK and the Netherlands, India, France and Australia.

Together with Hinrichs (again) and Petya Osenova (Bulgarian Academy of Science, Sofia) we organized a satellite workshop *Computational Phonology* at *Recent Advances in Natural Language Processing 2007* on Sept. 26, 2007 in Borovetz, Bulgaria. This was also part of what developed into a satellite project sponsored by the Volkswagen foundation (see below, part b), and was attended by about 20 researchers from Bulgaria, Canada, France, Germany, Greece, Hungary, Italy, and the Netherlands. A copy of the proceedings is enclosed with this report.

Together with Sebastian Kürschner (Erlangen), Charlotte Gooskens and Renée van Bezooijen (both in Groningen) we organized a special two-day long track at *Methods XIII* on Aug.4-5, 2008 in Leeds. The proceedings appeared as *Computing and Language Variation*, a double special issue of *International Journal of Humanities and Arts Computing* 2(1-2), 2008, with

contributions from project members Leinonen, Heeringa, Gooskens and Nerbonne and international participation from Germany, the United Kingdom, Belgium, Italy, Finland, and China. We include a copy of the special issue with this report.

b. Foreign contributions to project.

We were fortunate enough to be invited to participate in the Volkswagen Foundation's Unity and Diversity in Europe program. The presence of the NWO project in Groningen may have enhanced our chances of obtaining this. The project was conducted in collaboration with the *Seminar für Sprachwissenschaft, Eberhard Karls Universität, Tübingen*, and the Bulgarian Academy of Sciences, Sofia. Prof. Erhard Hinrichs, Tübingen and Prof. Vladimir Zhobov, Sofia, were the other site leaders. The Groningen participation in the project consisted of a Ph.D. project conducted by Jelena Prokić from Oct. 1, 2006 through Sept. 30, 2010. Her Ph.D. will be awarded on Nov. 29, 2010 for thesis *Families and Resemblances*, investigating dialectology not only from a synchronic point of view (the focus of DDV) but also from a diachronic, historical one. Prokić collaborated with Heeringa, Leinonen and Nerbonne extensively during her Ph.D. period. See <http://www.sfs.uni-tuebingen.de/dialectometry/> for a project web site and <http://www.let.rug.nl/~prokic/> for publications.

There is also interest in dialectology on the part of population geneticists, which led to John Nerbonne's visiting the *Musée de l'Homme* as guest researcher in 2006 and 2007, and which resulted in a studies which complement DDV work first by examining more exotic languages (see Alewijnse et al. 2007 and van der Ark et al 2007); second, by examining the degree to which linguistic and genetic signals of relatedness correlate (see Manni et al. 2008); and third, by comparing techniques (see Nerbonne et al 2008c).

Subprojects

B.1. Subproject on Syntactic Variation (Ph.D. project)

B.1.1 Researcher: Dr. Marco Spruit, Meertens Institute.

B.1.2 Intended publication: The aim of the subproject was to apply dialectometric techniques to syntactic variation for the first time, and this was accomplished in Spruit (2005, 2006a) and in more detail in Spruit (2008). A second goal was to investigate the degree to which syntactic, pronunciational and lexical variation co-vary, the focus of Spruit et al. (2009). Finally, it was also a goal to investigate techniques for detecting the degree to which syntactic variables are associated. The work on this topic is quite original, and a first version was published in Spruit (2007).

Marco Spruit defended his dissertation (Spruit, 2008) at the University of Amsterdam on Mar. 26, 2008. A copy was sent to NWO.

B.1.4 Final Report. The subproject on syntactic variation was conducted at the Meertens Institute by Marco Spruit. The aim of the subproject was to apply dialectometric techniques to syntactic variation for the first time. When we first presented the project, one leading syntactician predicted that because syntactic variation is organized along typological lines, it would be conditioned by geography to a far lesser extent than pronunciation. This turned out to be wrong.

This subproject investigated three quantitative perspectives on syntactic variation in Dutch dialects. The first perspective applied a simple measure of difference (essentially Hamming distance) to the syntactic properties noted in the *Syntactische Atlas van de Nederlands Dialecten* (SAND) and uses these to classify the Dutch dialect varieties syntactically (Spruit,

2006a/b). This classification is compared with—and highly resembles—the traditional, perceptual classification based on subjective judgements first developed by Weinen and Daan (Spruit, 2005).. Spruit showed thus that syntactic variation patterns are geographically coherent. He was awarded an award as “excellent young researcher” by the Association for Literary and Linguistic Computing at the 2005 Methods XII conference in Canada for the presentation of his paper (Spruit, 2006a).

In collaboration with Heeringa and Nerbonne Spruit then quantified the degrees of association between pronunciation, lexical and syntactic differences (Spruit et al. 2009). The analysis reveals that the linguistic levels of pronunciation, lexis and syntax are genuinely albeit modestly associated. It also turned out – surprisingly – that syntactic and pronunciation differences only as strongly associated with each another as each is with lexical differences. We had expected that the well known volatility of the lexicon would lead to rather different patterns. We comment on this central line of research more fully in the summary of the postdoc project (below).

Finally, Spruit developed a technique for searching for associations among syntactic variables using a data mining technique based on geographical co-occurrences (Spruit 2007, 2009). This approach contributes to the validation of existing typological hypotheses and facilitates the identification and exploration of relations among linguistics variables in general.

As an additional activity flanking this subproject, Nerbonne co-organized a workshop (with population geneticist Dr. Franz Manni, Musée de l’Homme, Paris) on comparing syntactic databases at the 2006 Digital Humanities meeting in Paris. Proceedings appeared in 2009 as *The Forests behind the Trees* in a special issue of *Lingua* (Nerbonne & Manni, 2009). It included the article by Spruit et al. but also articles by Michael Dunn (Max Plank Institute, Nijmegen), by Bendikt Szmrecsányi and Bernd Kortmann (Freiburg), by Sjef Barbiers (Meertens), by Pino Longobardi (Trieste), and by Jan-Wouter Zwart (Groningen). Nerbonne (2009a) introduced the special issue.

Spruit was awarded the Ph.D. in Amsterdam on March 26, 2008. A copy of the dissertation was sent then to NWO.

B.2. Subproject on Pronunciation Variation (Ph.D. project)

2.1 **Researcher:** Dr. Therese Leinonen, CLCG

2.2 **Publications** were *intended* applying the techniques Nerbonne and Heeringa had developed for Dutch to German. Christine Siedle worked on this subproject for one year, analyzing data that we had obtain from the *Deutscher Sprachatlas* (Marburg). While this resulted in a good publication in the leading journal for German dialectology (Nerbonne and Siedle, 2005), Ms. Siedle never settled comfortably in Groningen and left the project after a very good first year. Leinonen (2010) applies *inter alia* the Groningen techniques to Swedish data, and realizes this objective, but she goes on to apply acoustic techniques to the same data (see below).

Therese Leinonen was awarded the Ph.D. degree at the University of Groningen on July 1, 2010. A copy of the dissertation is included with this report.

2.4 **Final Report** The goal of the subproject on pronunciation variation was to replicate studies that had been conducted earlier by Nerbonne and Heeringa on Dutch on another language. As we note above, this was accomplished.

We attempted to make the best of the difficult situation of Ms. Siedle's leaving the project by bringing Therese Leinonen to it, who very much wished to work on Swedish dialects (Ms. Leinonen had worked on Swedish dialects at *Kotimaisten kielten tutkimuskeskus, The Research Institute for the Languages of Finland*). We therefore contacted Prof. Anders Eriksson, Göteborg, a phonetician who had worked on the most recent large collection of Swedish dialect material, SweDia. The nature of the SweDia data was certainly challenging given the range of analytical techniques we had worked with in Groningen until then.

The data from the SweDia dialect database were recorded at 98 rural sites in Sweden and the Swedish speaking parts of Finland around year 2000. At each site approximately twelve speakers were recorded: three older women, three older men, three younger women and three younger men. But unlike the pronunciation data we had analyzed earlier from the Netherlands, Germany and Norway, which consisted of phonetic transcriptions of entire words, the SweDia data consisted of acoustic recordings of vowels alone.

The vowel quality of nineteen different vowels was analyzed acoustically by means of principal component analysis (PCA) of Bark-filtered spectra. The two extracted principal components can be interpreted roughly in terms of vowel height and advancement. A correlation with formant measurements of a subset of the data showed high correlations. Separate PCAs of vowels produced by male and female speakers effectively reduced the variation in the acoustic measures related to anatomical/physiological differences between men and women.

Dialectal variation was studied both in each vowel and on an aggregate level. Both methods contributed to the understanding of the dialectal variation and were shown to complement each other. The thesis includes maps that display the pronunciation of each vowel in the two age groups at each site. Co-occurring vowel features were identified by a factor analysis. The aggregate analysis showed that when it comes to vowel pronunciation the Swedish dialects form a linguistic continuum without abrupt dialect borders. Within the continuous distribution of vowel features, however, some more coherent dialect areas can be identified. These areas coincide to a large extent with classifications that have previously been proposed for Swedish dialects.

The analyses indicate a large-scale leveling of Swedish dialects. The linguistic distances between sites based on vowel pronunciation are significantly shorter for younger speakers than for older speakers. In central Sweden the aggregate distance in vowel pronunciation between older and younger speakers is large. In some peripheral areas the linguistic change in apparent time is significantly smaller.

The ongoing change in vowel pronunciation in central Sweden is connected to an ongoing chain shift in front mid-vowels. This language change is the result of dialect contact in combination with convergence to Standard Swedish. For Standard Swedish the vowel shift means that a phoneme system which has been very difficult to describe structurally is being simplified in the course of time, as the vowel inventory becomes both smaller and more symmetrical.

Therese Leinonen was awarded the Ph.D. degree on July 1, 2010. A copy of the dissertation is included with this report.

B.3. Postdoc project on the determinants of dialectal variation.

This project focused on the determinants of linguistic differences (geography, population sizes) and the relationships among linguistic levels (pronunciation, lexis, syntax, prosody).

1) Determinants of linguistic distances

The first study is based on a high density network of 52 settlements which all lie within the Lower Saxon dialect area of the northern Netherlands. Pronunciation distances among the dialects of these settlements are calculated with Levenshtein distance on the basis of transcriptions, taken from the *Reeks Nederlandse Dialectatlassen*, 125 words pronunciations per dialect. Geographic distances between settlements are obtained on the basis of longitude-latitude coordinates and 'as the crow flies'. Geography accounts for 59% of the variance in the aggregated pronunciation distances when using a linear regression model. The populations of the different settlements were taken from the *Geschiedkundige atlas van Nederland; Het koninkrijk der Nederlanden 1815-1931* Ramaer (1931) and date from around 1815.

Geography and populations sizes are put in one model according to Trudgill's linguistic version of the gravity model (where the constant is omitted): the population size product of two locations divided by the geographic distance between the two locations. The gravity model did not add explanatory power in comparison to geography only. The contribution of the population product independently is moreover positive, contradicting the predictions of the gravity model! (Nerbonne & Heeringa, 2007).

A second study is based on 27 varieties in the Netherlands and Flanders. The data had been collected by Renée van Bezooijen in 2001, pronunciations of 100 nouns per variety. The findings were similar: geography correlated significantly with the pronunciation distances again and explains 33% of the variance, but population size information, although statistically significant, had only a minor effect. (Heeringa, Nerbonne, Van Bezooijen and Spruit, 2007).

2) Relationships among linguistic levels

In the period between 1999 and 2002 Jørn Almborg and Kristian Skarbø compiled a database which consists of recordings and phonetic transcriptions of translations of the fable 'The North Wind and the Sun' in about 50 Norwegian dialects. On the basis of 15 of these recordings, pronunciation distances, lexical distances and prosodic distances are measured among 15 recordings. The pronunciation level significantly correlates with the lexical level ($r=0.49$) and prosodic level ($r=0.43$), and the lexical level significantly correlates with the prosodic level ($r=0.18$). The three levels are correlated with perceptual distances, i.e. distances among dialects as perceived by the speakers themselves. The perceptual distances were obtained by an experiment carried out by Charlotte Gooskens in the spring of 2000. The three linguistic levels correlate significantly with the perceptual distances, but highest correlation was found for the pronunciation level. Within the pronunciation level a distinction was made between consonants and vowels on the one hand, and between substitutions and insertions/deletions on the other hand. When correlating the separate levels with perception and using multiple linear regression analyses it appears that especially consonant substitutions play a major role (Gooskens & Heeringa 2006).

In a second study the levels of lexicon and pronunciation are considered, using as data 360 Dutch sites from the *Reeks Nederlandse Dialectatlassen*, 125 words per dialect. Lexical distances were measured using Goebel's 'gewichteter Identitätswert' (GIW), a method in which the coincidence of rarely used words counts more heavily than those of more frequent ones. Pronunciation differences are measured using Levenshtein distance. The two levels correlate significantly ($r=0.63$, $p<0.001$). The measurements of the two levels are combined by taking the average of the normalized lexical and pronunciation distances. The results are similar to the traditional dialect map of De Schutter, which is considered by the author to reflect the 'communis opinio' of traditional dialectologists at the end of the 20th century. (Heeringa & Nerbonne 2006)

In a third study the pronunciation level, the lexical level and the syntactic level are compared to each other and to geography using data of 70 Dutch dialects. This research is based on two Dutch dialectal data sources: the Reeks Nederlandse Dialectatlassen (pronunciation and lexis) and the first volume of the Syntactische Atlas van de Nederlandse Dialecten (syntax). Lexical and syntactic levels were measured using GIW. The three levels correlate significantly with each other. Pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and syntax is much more strongly associated with pronunciation (42%) than with lexis (25%). The three levels were correlated to geographic distances, which are measured in kilometers on the basis of longitude-latitude coordinates and ‘as the crow flies’. For all levels significant correlations were found. Pronunciation and syntax are more strongly associated with geography (47% and 45%, respectively) than lexis is (33%). Next the levels are correlated to each other again, but the influence of geography is filtered away as a factor of influence underlying the associations among the linguistic levels under investigation. Some influence between pronunciation and syntax remains (12%), although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (3%). (Spruit, Heeringa & Nerbonne 2009).

Publications

- Alewijnse, B. Nerbonne, J. van der Veen, L. & Manni, F. (2007) A Computational Analysis of Gabon Varieties In Petya Osenova et al. (ed.) *Proceedings of the RANLP Workshop on Computational Phonology*. Borovetz. 3-12.
- van der Ark, R. Mennecier, P. Nerbonne, J. & Manni, F. (2007) Preliminary Identification of Language Groups and Loan Words in Central Asia In Petya Osenova et al. (ed.) *Proceedings of the RANLP Workshop on Computational Phonology*. Borovetz, 13-20.
- Gooskens, Ch. & Heeringa, W. (2006) The relative contribution of pronunciation, lexical and prosodic differences to the perceived distances between Norwegian dialects. In: J. Nerbonne & W. Kretzschmar, Jr. (eds.), *Literary and Linguistic Computing*, special issue, *Progress in Dialectometry: Toward Explanation*, 21(4), Oxford University Press, Oxford, 477-492.
- Heeringa, W.J., Nerbonne, J., Bezooijen, R. van, & Spruit, M.R. (2007). Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied. *Tijdschrift voor Nederlandse taal- en letterkunde*, 123(1), Uitgeverij Verloren, Hilversum, 70–82.
- Heeringa, W. & Nerbonne, J. (2006) De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak. *Nederlandse Taalkunde*, 11(3), 218-257.
- Leinonen, Th. (2008). Factor Analysis of Vowel Pronunciation in Swedish Dialects. *International Journal of Humanities and Arts Computing* 2(1-2), 189-204.
- Leinonen, Th. (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. PhD thesis, University of Groningen. GRODIL 83.
- Leinonen, Th. (submitted). Aggregate Analysis of Vowel Pronunciation in Swedish Dialects. Submitted to *Oslo Studies in Language (OSLa)* March, 2010.
- Manni, F., Heeringa, W., Toupance, B. & J. Nerbonne (2008) Do Surname Differences Mirror Dialect Variation? *Human Biology* 80(1), Feb. 41-64.
- Nerbonne, J. (2009a) Introduction to *The Forests behind the Trees*. *Lingua* 119 (11) Spec. issue *The Forests behind the Trees* ed. by John Nerbonne and Franz Manni. 1581-1588.
- Nerbonne, J. (2009b) Data-Driven Dialectology. *Language and Linguistics Compass* 3(1), 2009, 175-198. DOI: 10.1111/j.1749-818x.2008.00114.x

- Nerbonne, J., Gooskens, C. Kürschner, S. & van Bezooijen, R. (2008a) Language Variation Studies and Computational Humanities. *International Journal of Humanities and Arts Computing*, Special Issue on *Language Variation* ed. by J. Nerbonne, C. Gooskens, S. Kürschner, and R. van Bezooijen. 1-18. DOI: 10.13366/E1753854809000287
- Nerbonne, J. & Heeringa, W. (2007) Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: S. Featherston and W. Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base, Studies in Generative Grammar 96*, Mouton De Gruyter, Berlin, 267-297.
- Nerbonne, J., & Heeringa, W. (2009) Measuring Dialect Differences In: J. E. Schmidt and P. Auer (eds.) *Language and Space: Theories and Methods* in series *Handbooks of Linguistics and Communication Science*. Berlin: Mouton-De Gruyter, Ch. 31, 550-567
- Nerbonne, J., Heggarty, P., van Hout, R. & Robey, D. (2008b) Panel Discussion on Computing and the Humanities. *International Journal of Humanities and Arts Computing*, Special Issue on *Language Variation* ed. by J. Nerbonne, C. Gooskens, S. Kürschner, and R. van Bezooijen. 19-37. DOI: 10.13366/E1753854809000299
- Nerbonne, J., Kleiweg, P., Heeringa, W. & Manni, F. (2008c) Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.) *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society* Berlin: Springer. 2008. 647-654. (*Studies in Classification, Data Analysis, and Knowledge Organization*)
- Nerbonne, J. & Manni, F. (eds.) (2009) *The Forests behind the Trees*. Special issue of *Lingua* 119 (11) 1581-1706.
- Nerbonne, J., & Siedle, Ch. (2005) Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2), 129-147.
- Spruit, M.R. (2005). Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited. In: Doetjes, J., Weijer, J. van de (eds), *Linguistics in the Netherlands, 2005*, John Benjamins, Amsterdam, 179-190.
- Spruit, M.R. (2006a). Measuring syntactic variation in Dutch dialects. In J. Nerbonne & W. Kretzschmar, Jr. (eds.), *Literary and Linguistic Computing*, 21(4), spec. issue, *Progress in Dialectometry: Toward Explanation*, 493-506.
- Spruit, M.R. (2006b). Tellen met Taal. Het meten van variatie in zinsbouw in Nederlandse dialecten. In: Gerritsen, D., Verburg, A. (eds), *Respons: Mededelingen van het Meertens Instituut*, 8, Meertens Instituut, Amsterdam, 12-16.
- Spruit, M.R. (2007). Discovery of association rules between syntactic variables. Data mining the Syntactic atlas of the Dutch dialects. In: Dirix, P., Schuurman, I., Vandeghinste, V., Eynde, F. van (eds), *Computational Linguistics in the Netherlands 2006. Selected papers from the seventeenth CLIN meeting*, LOT Occasional Series, Utrecht, 83-98.
- Spruit, M.R. (2008). *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD thesis, University of Amsterdam, LOT Dissertation Series 174, LOT, Utrecht, 157 pp.
- Spruit, M.R. (2009). Towards linguistic knowledge discovery in language variation databases. *Zeitschrift für Dialektologie und Linguistik – ZDL-Beiheft 138*, Low Saxon Dialects across borders, Stuttgart: Franz Steiner Verlag, 179-193.
- Spruit, M.R., Heeringa, W., & Nerbonne, J. (2009). Associations among linguistic levels. *Lingua*, 119 (11), The forests behind the trees, Elsevier, 1624-1642.
- Wieling, M., Heeringa, W., & Nerbonne, J., (2007) An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-van Reenen-Project Data. *Taal en Tongval* 59(1). 84-116.
- Wieling, M., Leinonen, Th. & Nerbonne, J. (2007). Inducing Sound Segment Differences Using Pair Hidden Markov Models. In J. Nerbonne, M. Ellison & G. Kondrak (eds.). *Computing and Historical Phonology: 9th Meeting of ACL Special Interest Group, Computational Morphology and Phonology, ACL Workshop*