4. Previous and Future Submission

This proposal has not been submitted before and will not be submitted elsewhere during the selection procedure.

5. Institutional setting

The research for this program will primarily be carried out at the Meertens Instituut in Amsterdam, in cooperation with the research institutes for linguistics in Utrecht (UiL-OTS), Leiden (LUCL) and Groningen (CLCG). The PhD-students will participate in the national research school for linguistics (LOT).

6. Period of funding

Run-time of the program: 5 years
Starting date: January 2, 2013.

7. Composition of the research team

Main applicant	prof. dr. Sjef Barbiers, Meertens Institute and Utrecht University
Second applicant	prof. dr. Marc van Oostendorp, Meertens Institute and Leiden University
Researchers	Project 1 (postdoc): dr. Marco René Spruit, Utrecht University Project 2-4 (PhD students): n.k.
Other researchers	prof. dr. John Nerbonne (University of Groningen) prof. dr. Hans Bennis (Meertens Institute and University of Amsterdam). prof. dr. Roberta D'Allesandro (University of Leiden). dr. Gertjan Postma (Meertens Institute).

8. Structure of the Proposed Research

Project	**Position**	**Institute**	**Name**	**Supervisors**
1. Computing regions	Postdoc	Meertens Institute U. Groningen	dr. Marco René Spruit	Nerbonne Van Oostendorp
2. Inclusion relations	PhD-student	Meertens Institute U. Utrecht	n.k.	Barbiers (promotor) Van Oostendorp (promotor) Postma (co-promotor)
3. Transition zones	PhD-student	Meertens Institute U. Leiden	n.k.	Bennis (promotor) D'Alessandro (promotor) Postma (co-promotor)
4. Grammaticometrics	PhD-student	Meertens Institute U. Groningen	n.k.	v. Oostendorp (promotor) Nerbonne (promotor) Barbiers (promotor)

9. Description of the Proposed Research

Maps and Grammar

In recent years we have observed an increasing interest in the linguistic properties of language varieties and dialects. However, a crucial issue remains to be investigated so far. It concerns the question whether the concepts 'dialect' and 'regional language variety' are contentful from a grammatical perspective, and if so, in what way. Is there any linguistic substance to the division of a language area into smaller unities?

To give an example, particular dialect groups within the Dutch speaking area such as the dialects of Lower Saxony and of the Dutch part of the Limburgian area are officially recognized as 'regional languages' under the charter of Regional and Minority languages of the European Council. We will not enter the discussion here whether or not official recognition of particular Dutch speaking areas is warranted. The issue of recognition raises a different, more fundamental issue, especially given the fact that the recognized dialect areas are far from uniform in their linguistic properties. What makes these regional languages linguistically distinct from the neighbouring language varieties?

The recognition-issue is generally decided on by historical, political or sociological arguments. The languages themselves are not taken into consideration seriously. The reason for this is that the question of defining the concept 'regional variety' has not been sufficiently discussed from a linguistic perspective. We know that there is a lot of syntactic, morphological, phonological, phonetic and lexical variation within a language area, but we don't know how much a variety should differ from other varieties to be considered a distinct variety. In this project the understanding and the definition of the concept 'linguistic region' is a central objective.

We use four ways to approach this issue. First of all, we will aggregate the existing empirical material computationally to distinguish relatively stable, homogeneous linguistic regions. Secondly, we will approach the topic from a more qualitative linguistic angle by trying to establish inclusion relations in the linguistic system which will give clues about cooccurrence patterns that might be typical for determining linguistic regions. Thirdly, we will look for geographic areas that show a relatively unstable and heterogeneous linguistic situation. This instability may indicate that these areas are transition zones between language varieties. Fourthly, we will investigate if the different regions identified in this research correspond to different grammars, i.e. abstract coherent rule or constraint systems that underly the (morpho-)syntactic and (morpho-)phonological patterns attested. Together, these four perspectives should enable us to uncover the grammatical landscape underlying the observed variation. The divisions of the map based on inclusion relations, transition zones and grammatical systems are expected to differ from the divisions resulting from the quantitative (or aggregational) method.

The large data collections required for this kind of research are available. Dialect geography is traditionally a linguistic subdiscipline with a heavy empirical focus. 19th century dialectologists, like Georg Wenker in Germany (cf. Schmidt and Herrgen 2009) or Joseph Wright in England (Wright 1898-1905), already built relatively large databases based on extensive fieldwork. This work has continued ever since. At the beginning of the 21st century, we are equipped with a large number of databases and theoretical and technical tools allowing us to handle linguistic data in an unprecedented way. For the Dutch language area, the dialect atlases FAND (phonetics/phonology), MAND (morphology) and SAND (syntax) have been published in the past two decades, and the databases underlying these atlases are available on line, including data from more than 600 locations in The Netherlands and Belgium with several hundreds of linguistic variables (cf. DynaSAND, www.meertens.knaw.nl/sand; MIMORE www.meertens.knaw.nl/mimore; cf. the Edisyn Search Engine for some databases of other European dialects, www.meertens.knaw.nl/edisyn/searchengine).

Although the fieldwork for these databases and atlases has been inspired by linguistic questions, the geographic patterns found on the dialect maps have hardly played any role in linguistic theory so far and the maps are often little more than a visual display of the databases. With the exception of typological (i.e. macrovariation) research (cf. Nichols 1992) and dialect geography, for a long time linguistic research has been disconnected from the use of geographical maps as a tool for investigation.

The reason is that the focus of much linguistic research is on language viewed either as the property of an individual human being (the so-called internalist view of language; cf. Chomsky 1957, 1995) or as a property of a social group (the sociolinguistic view; cf. Labov 1966, 2006). In either case, the relevance of geographical information is unclear.

This program aims to bridge the gap between the fastly growing capacities of Geographic Information Systems (GIS) and the vastly enriched insights of modern linguistics. We do this by exploring the different ways in which geographical regions are relevant for the questions which interest linguists. We study the many ways in which patterns on a map can provide evidence in particular for internalist issues, and we do so from a cross-disciplinary perspective, taking into account insights and methods of dialect geography, dialectometry and theoretical syntax, morphology and phonology.

The available data will shed new light on many of the questions that modern linguistics is facing. An important question concerns the reality of grammatical structure, which has been challenged in the past years by adherents of usage-based approaches to language, who claim that the only cognitively real aspects of language are concrete instances of data plus emerging abstractions of these, and argue against supposition of ('innate') linguistic structure (e.g. Langacker 1987, 2000; cf. Bod 2006 for discussion). Such theories will entertain vastly different models of how languages vary and change. Relevant models have been tested and defended on the basis of large typological databases of macrolinguistic variation (e.g., Evans and Levinson 2009).

This program aims to complement this work by looking at micro- rather than macrovariation and to show how fine-grained linguistic analysis can be used to investigate the nature of the linguistic system, in particular with respect to the role of grammar in usage-based versus structuralist/generative theories of language. The role that microvariation research can play here is distinct from that of macrovariation research as the dialects of one language family have the large majority of their grammatical properties in common. This comes close to what Kayne (2005) calls an ideal language laboratorium that makes it possible to study the effects of minor changes or variations in a grammatical system on the other properties of that grammatical system. The correlations found in this way may shed light on the more abstract properties that constitute a grammatical system (cf. Barbiers 2008).

Starting from the concept of *geographic region*, we ask what conclusions we can draw from the fact that certain linguistic phenomena occur in each other's geographic vicinity. The most important questions are the following; a project will be devoted to each of them.

- **Computing regions** (project 1). How can we automatically detect linguistic regions in older and newer electronic databases? Which data mining techniques can be used to decide that a number of locations on a map belong together and should be considered a 'region'? How can we visualise the complex patterns (which involve at least a two-dimensional geographic space and a possibly multidimensional linguistic analysis) on a computer screen?
- **Inclusion relations** (project 2). How are we to interpret inclusion relations among linguistic regions? What interpretation can we give to the observation that some linguistic phenomenon *x* can only be found in regions that are properly included in regions in which we find *y?* Can we conclude from this, for instance, that a language system only has *x* if it also has *y*? For the discovery of inclusion patterns this project partially builds on the tools and results from Project 1.

- **Transition zones** (project 3). What happens in areas that are ‘in between’ larger dialect regions? It is known that in such areas we can find phenomena that are not usually found elsewhere. One of the possible explanations for this is that speakers in such transitional areas are themselves in transition from a dialect system as is spoken on the one side of their area to a dialect system as it is spoken on the other side. For the identification of transition zones this projects partially relies on the tools and results from project 1. Insight in the linguistic properties of transition zones is crucial for project 2, as this project studies transitions between sets of dialects that are subsets of each other.
- **Grammaticometrics** (project 4). Which types of linguistic data must be mapped to find regions that most faithfully reflect the ‘real’ differences between dialects? On most dialect maps thus far only superficial data have been subjected to cartography, such as individual words or grammatical constructions. What would change if we would instead map the ‘deeper’ structures underlying such data, i.e. the grammars? How do the regions differ, and what do such differences tell us? This project crucially compares a geographic division based on grammars with the division based on aggregated linguistic features, applying the tools produced in project 1. For the identification of distinct grammatical systems the results of project 2 and 3 are crucial too.

Summarizing, this program makes the following linguistic contributions. It provides quantitative and qualitative linguistic criteria for concepts such as dialect area and regional dialect. It contributes to the theoretical modeling of language variation and change and our understanding of systematic properties of languages and to the debate between internalist vs. usage based linguistics, in particular w.r.t. the question if there is evidence for the postulation of a grammatical system underling the surface linguistic patterns.

Raising the use of geographical information to a higher level of sophistication will allow linguists to interact more directly with other geographically oriented disciplines, such as social geographers and geologists. It is interesting and important from a variety of perspectives to compare the concept of ‘region’ in linguistics with regions as they are defined in other disciplines. In this project we are mainly setting up the infrastructure for such communication, but we intend to form a network with other scholars outside of linguistics, especially in the Netherlands. For this, we will organize a series of seminars in Amsterdam with representatives from different geographically oriented disciplines.

Cooperation with political geographers is particularly likely to be fruitful. Such cooperation will also open up the possibility of reaching out to regional and local, but also European, politicians. During the past decade, many provinces in the Netherlands have installed an official language policy on regional languages, e.g. hiring officials to support local groups in writing dictionaries of or theatre plays in their language. Some of the language varieties that are supported in this way are covered by the European Charter for Regional Languages or Languages of Minorities, but others are not, and in several provinces the situation is quite complicated. Linguistic insights resulting from this program will provide tools for making more informed decisions. Something similar holds at the European level, since several European institutions (including, but not restricted to the Council of Europe that authored the Charter just mentioned) invest in supporting ‘autochtonous’ regional languages. Towards the end of the project, we get together with political geographers to try to reach out to this target group.

Our results will also be important for the general public. The local language and local dialect continue to be important for identification in a changing Europe, and people tend to be highly interested in the language and culture of their 'own' region. We believe that scholars have the duty to inform people in a neutral way about the latest findings on such matters and to battle mystifications that might arise. Geographical linguists in the Netherlands, in particular those at the Meertens Instituut, have a long tradition in reaching out to the general public in this way. We aim to do that in this project too.

Project 1: Computing regions (postdoc)

To understand the concept of the linguistic region, it is necessary to develop an adequate computational idea of it, and of the connection between geographic data and linguistic theory. How can we automatically detect linguistic regions within a large set of complex and mutually dependent data? This project primarily aims to develop two such sets of tools, which will be used and tested in the other projects. It builds on existing dialectometric work in which linguistic distances between dialects are calculated and visualised, such as the work by Goebl (cf. Goebl 2005 and www.dialectometry.com) and, in the Netherlands, by Nerbonne and his group (cf. Heeringa and Nerbonne 2001, Heeringa 2004, Heeringa et al. 2006, Spruit et al 2008, Spruit 2008; cf. also www.gabmap.nl)

The first set of tools to be developed concerns the concept of a region. When can we say that a number of data points on a map corresponds to a coherent region? In linguistic fieldwork, it is unfeasible to study every individual town or village. Furthermore, in many cases the data are statistically difficult to analyse, because often the number of informants per location is not higher than one or two. Although it would of course be desirable to improve such shortcomings in new fieldwork, it is still necessary to also use older datasets, e.g., because it is interesting to compare the old situation to the new one, but also because the financial limitations and the amount of work it takes to analyse the data of even one single informant will not diminish in the near future.

A statistical tool for using regions in geolinguistic analysis would therefore need, first, to allow us to define the concept of a region statistically. When can we decide that a specific location is likely to share a certain property with his neighbours, in the absence of real data on that location? Secondly, it would allow us to take such regions rather than individual locations as the basis for our statistics (i.e., it would be possible to aggregate over several neighbouring villages with each only a few informants so that they can count as one bigger location with enough informants to do responsible statistical analysis). An obvious place to start would be in hierarchical agglomerative cluster analysis techniques such as WPGMA (Weighted Pair Group Method with Arithmetic Mean), in combination with a set of measures that would ensure the overall stability of the system (adding or taking away a few data points should not result in dramatically different results, at least above a certain threshold).

This first set of tools will be tested in the projects on Transition and Inclusion, and will be of general use for statistical analysis on the type of geographic datasets as can be found in traditional linguistic atlases. This tool will probably also be useful in other types of geographically based research. Indeed, it would be preferable if some of the results of this tool would have an interpretation that can be used outside of linguistics proper, as one traditional application of linguistic maps is to compare them with other types of maps (e.g., depicting the location of rivers or cities, or political and other social borders.)

The second set of tools concerns visualisation techniques for studying multidimensional geographical patterns in which the third dimension consists of complex linguistic material. An obvious example (and testing bed) for this is studied in the project on Grammaticometrics, in which it is proposed that formal grammars can be seen as such. More concretely, an optimality theoretic grammar (cf. McCarthy and Prince 1993) can be seen as an ordered tuple <C1, C2,... Cn> in which Ci are constraints that are shared by all dialects; they only differ in their relative ordering. Since a tuple is a one-dimensional object, putting it on the map would result in a 3-dimensional figure. A 3-dimensional map would allow the researcher to study how a constraint moves up and down in a landscape. Similar visualisation techniques have been developed for geology (in which one can study changes in the structure of earth layers); these should be adapted to the needs of linguistic analysis, and the way in which linguistic databases are organized. This second set of tools will be tested within the project primarily by the researcher who is carrying out project 4 (on grammaticometrics). The researcher of project 3 will participate in the small but rapidly developing discussion about the value of visualisation of large amounts of data in the humanities, and in particular in linguistics. The interest in working with large amounts of data is on the rise in many disciplines within the humanities (cf. Michel et al. 2011).

Project 2: Inclusion patterns

This project investigates inclusion relations in the geographic and diachronic distribution of grammatical features. An example from SAND Volume II is verb cluster interruption in the Dutch dialects. All Dutch dialects share the property that verbs cluster at the end of the clause, e.g. the verbs *moet*, *kunnen* and *brengen* in *dat hij het boek morgen* ***moet kunnen brengen*** lit.'that he the book tomorrow must can bring'. This is called verb clustering because the verbs in a cluster tend to be adjacent. For example, modal adverbs can not occur inside the cluster and therefore sentences like *dat hij het boek moet* ***vermoedelijk*** *kunnen brengen* lit. 'that he the book must presumably can bring' are impossible.

Dialects of Dutch vary with respect to the type of constituents that may occur inside a verb cluster. In most of the dialects, including Standard Dutch, so called particles can occur in between the verbs, such as *weg* 'away' in *dat hij het moet kunnen* ***weg*** *brengen* lit. 'that he it must away can bring'. Only one third of the dialects that can have a particle inside the verb cluster can also have a bare singular noun there, e.g. *brood* 'bread' in *moet kunnen* ***brood*** *eten* lit. 'must can bread eat'. Dialects that allow temporal adverbs (e.g., *vroeg* 'early') within the verb cluster are again a subset of the dialects with bare singular nouns, and so on for bare plural objects (*appels* 'apples'), indefinite objects (*een appel* 'an apple'), prepositional phrases (*met Jan* 'with John') and definite objects (*de appel* 'the apple'), the latter only being possible in a small south western corner of the language area (French- and West-Flanders). These subset relations correspond to an onion-shaped geographic pattern, with the dialects with interruption by all types of constituents just mentioned in the core of the onion, and the dialects that only have interruption by particles at the outer shell.

Strikingly, this onion-like patterning corresponds to gradual judgements by individuals that are in the outer shell. Despite the fact that in the dialects of such speakers only particles can occur within the verb cluster, they have clear but gradual intuitions about the acceptability of other interrupting constituents, i.e., a bare singular noun will be more acceptable than a temporal adverb, which will be more acceptable than a bare plural object, and so on.

A correspondence between gradual grammaticality and onion-like geographic distribution is also found with vowel reduction in subject pronouns. Similar correspondences can be traced in diachronic corpora, e.g. the correspondence in 14th century Dutch between frequencies of cliticisation of *to/te* 'to', as in *to/te doene* 'to do' and *ton/ten hofstede* 'at the farm', and the contexts that allow for cliticisation. The onion-like structure is to be formulated as inclusion relations between more and less "tolerant" dialects and corresponding higher frequencies.

The questions this project will ask about the relation between gradual grammaticality and geographic distribution include the following:

(i) For which (morpho-)phonological and (morpho-)syntactic phenomena in our databases does it hold that speakers across the language area show gradual grammaticality judgements?

(ii) Do the phenomena identified in (i) have an "onion"-like geographic distribution, with the least acceptable (and the least frequent) variant of a phenomenon occuring in the smallest geographical area and the most acceptable variant occuring in the largest area? Mutatis mutandis for frequency levels in diachronic corpora.

(iii) If the answer to (ii) is yes for one or more phenomena, how do we explain this correspondence?

(iv) If the answer to (ii) is no for certain phenomena, what other geographic and diachronic patterns do we find, and how do we explain these patterns?

(v) A speaker can have a grammaticality judgement on a variant that does not belong to his own dialect on the basis of familiarity with the other dialect, and this may give rise to graduality (the more familiar, the more acceptable and the higher its frequency). It can also be that a speaker does not give a plain positive or negative judgement on a particular variant, due to grammatical factors inside his own dialect. The question is how we can develop a method to tease these two possibilities apart.

(vi) What do these inclusion relations tell us about the locus of linguistic variation, i.e. which part of the variation that we find is determined by linguistic factors and which part is determined by extra-linguistic factors such as geographic distribution and frequency?

The research for this project will be based on the data in the DynaSAND- and GTR-databases for synchronic variation (www.meertens.knaw.nl/mimore) and on historical corpora such as Corpus Gyseling (CG), Corpus Van Reenen Mulder (CRM) and the Drenthe corpus. These historical corpora are digitally available at the Meertens Institute.

Project 3: Grammar in transition zones

It is evident from the maps in FAND, MAND and SAND that transition zones, e.g. the zone between the Frisian and the Low-Saxon dialects, show more variation, are linguistically more complex and show a less clear geographic distribution of linguistic features than non-transition zones. This complex situation invites a number of intriguing descriptive and theoretical questions about the interaction between dialects and the language varieties that result from this interaction.

The research questions include:

(i) How do we define transition zones and distinguish them from non-transition zones? As this is a relative, not an absolute distinction, a quantitative measure will be developed by the postdoc in project 1 on the basis of the entire data set of the DynaSAND- and GTR-databases. The proportion of accidentally absent grammatical features may be one way to distinguish between transitional and non-transitional zones. Grammatical features characterized as [+grammatical, -realized] are features that do not violate any principle of the grammatical system of the relevant dialect but nevertheless do not occur in that dialect. E.g., a syntactic construction may be possible in a dialect but accidentally absent, just like a word can be accidentally absent in a certain dialect. The project will develop methods to establish if a feature is accidentally absent or because it is ungrammatical. The expectation is that accidentally absent features are more characteristic for non-transition zones, as transition dialects may absorb the grammatically possible features that are present in their environment.

(ii) Dialect contact: Which grammatical features from neighbouring dialects are part of the dialects in a transition zone, are there any differences in this respect between the levels of phonology, morphology and syntax (cf. Trudgill 1999, Winford 2003, Taeldeman 2008) and do transitional dialects show more optionality, i.e. alternations of two or more equivalent structures, than non-transitional dialects? Are the speakers in a transition zone best characterized as bi- or multidialectal, or do they have one 'transitional' grammar? How to distinguish these possibilities? What is the best way to model these grammatical systems and the synchronic transition and diachronic change from one grammatical system to another, e.g., by constraint ranking as in Optimality Theory (McCarthy and Prince 1993, Kager 1999) or by parametrization of lexically defined morphosyntactic features as in Minimalism (Chomsky 1995, Kayne 2005)?

(iii) To which extent do we find linguistic properties that are typical for transition zones, i.e. properties that do not occur in any other part of the language area? Grammatical features that are characterized as [-grammatical, +realized] are features that violate the grammatical principles of a dialect system but that do nevertheless occur. Such features are expected to be unstable and subject to change or disappearance (cf. Postma 2010). A possible example is the complex anaphor *zich eigen* 'SE own' which is generally excluded in the Dutch dialects (cf. Barbiers and Bennis 2003) but occasionally shows up in transition zones that have *z'n eigen* 'his own' and *zichzelf* 'SE self' in the neighbouring dialects. It is a goal of the project to find more cases like this, analyse them theoretically and check if they arise and shortly after disappear in historical corpora.

(iv) What are the parallels and differences between the linguistic patterns found in geographic transition zones and those in transitional stages in the history of Dutch, that is stages in which the dialects show a relatively rapid change? Does Kroch's Constant Rate Hypothesis for diachronic change (Kroch 1989) apply to synchronic transition?

(v) Which similar and different linguistic and geographic patterns do we find when we compare different transition zones and how does this vary with their intralinguistic and extralinguistic environments? W.r.t. the latter, one question would be about the effect of different types of political borders (regions, provinces, countries) on the linguistic variation in a transition zone.

The research in this project will be based on the databases mentioned in project 2. Verbal inflection will be one of its major foci, as this is at the interface of phonology, morphology and syntax and shows abundant variation in the Dutch language area. The Meertens Institute has a separate digital database on verbal inflection for more than 600 locations based on the DynaSAND and GTR data.

Project 4: Grammaticometrics

Recent years have seen considerable advances in the mathematical modeling of geographic language variation: refined statistical and data mining techniques have been developed which allow the researcher, first, to determine quite precisely how different forms in different dialects really are, second, to aggregate such differences over whole dialects to get a global picture of differences, third, to visualise the data thus found in a variety of ways, and fourth, to compare different dimensions to each other (e.g. syntax to phonology, or linguistic to geological maps) (cf. project 1 and references cited there).

This project aims to literally add another dimension to these maps. So far, the comparison between dialects has been based on linguistically rather 'superficial' data: comparing surface forms of individual words in the case of phonology, or constructions in the case of syntax. E.g., the difference was calculated between one dialect saying 'paard' for horse, and another saying 'peerd' (on a Levenshtein scale, this difference would be 1, since one needs to replace one vowel for another), or a dialect with the word order *dat ik dat boek wil lezen* lit. 'that I that book want read' was compared to a dialect with the word order *dat ik dat boek lezen wil* lit. 'that I that book read want'. Several types of modern grammatical theory – e.g. usage-based grammar or construction grammar – would predict that these traditional dialectological maps are exactly at the right level: such theories claim that the individual construction is all there is to language structure. However, another tradition in grammatical microvariational analysis – the one usually identified as structuralist and generative grammar – claims that differences between language varieties tend to be systematic, just because languages are more than collections of individual words or grammatical constructions. They are regulated by formal 'grammars', which in turn are responsible for the fact that vowels are fronted in certain contexts, or that the auxiliary follows the main verb. Such a tradition would predict that we would get interesting geographic distributions if we do not put (zero-dimensional) individual constructions on the map, but rather (one-dimensional) grammars.

In order to test, then, usage-based vs. grammar-based approaches to language variation, we would first need to work out some of the predictions of the grammar-based approaches, and therefore create sample maps in which grammars are mapped rather than constructions. Given the fact that it has so far not been possible to construct a complete grammar of any language under any theory, we will have to severely restrict the grammatical domain, and construct fragments of grammars for a well-defined subdomain, such as clitic formation or verbal inflection. Our expectation is that the regions which are mapped in this way would show a 'deeper' level of linguistic variation. In as far as these regions appear at all, they would give a new type of evidence supporting grammar-based views of language variation. On the other hand, if it proves to be impossible to draw such maps, this could be an indication that construction-based approaches are on the right track.

An example from Weinreich (1954, 392-393) serves to illustrate the different analytical results of the construction-based and grammar-based approaches. Suppose four speakers of a language pronounce the word *man* in the following ways: (1) [man], (2) [man], (3) [mån] and (4) [mån]. On the basis of these pronunciations alone, a non-structural analysis would conclude that (1) and (2) should be grouped together and distinguished from the group formed by speakers (3) and (4). Suppose now the following hypothetical system. Informant (1) speaks a variety in which vowel length is a distinctive feature. The form of informant (1) is then phonemically /măn/. Informant (2) does not distinguish vowel length and has /man/. In the system of informant (3), the pronunciation of [å] in [mån] is a positional, non-distinctive rounded variant of [a] showing up between nasals. Phonemically, informant (3) has /man/. The fourth variety does not have this positional variation. Informant (4) has /mån/. Thus, according to this structural analysis, system (2) and (3) are the same and group together, while systems (1) and (4) are different. Superficial observation of this hypothetical system yields a map of the /a/ area that is split into two parts by the isogloss between /man/ and /mån/. According to the structural analysis, however, this isogloss does not exist at all: a clear difference in prediction.

Similar 'contradictory' results can be expected in other parts of the grammatical system. A good candidate is the verbal *–t* suffix in Dutch dialects that distinguishes second person plural from first and third person plural in some dialects, while it expresses plural in first, second and third person in other dialects. Like in project 3, verbal inflection will be one of the central domains of investigation in this project.

Appendix: References

Barbiers, S. & H. Bennis, 2003. Reflexives in Dutch Dialects. In: Koster, J. & H. van Riemsdijk (ed.). *Germania et alia*. Groningen : Universiteit Groningen, 2003, pp. 25-44. http://odur.let.rug.nl/~koster/DenBesten/contents.htm.

Barbiers, S, 2008. Locus and Limits of Syntactic Variation. *Lingua* 119, 1607-1624. [Special Issue The Forest Behind the Trees, eds. J. Nerbonne and F. Manni].

Bod, R. 2006. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review* 23, 291-320.

Chomsky, N. 1957. *Syntactic structures*. 's-Gravenhage: Mouton (Janua Linguarum; Series minor, 4).

Chomsky, N., 1995. *The minimalist program*. Cambridge, Mass.: MIT Press (Current studies in linguistics series, 28).

Evans, N. and S. Levinson, 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32 (5), 429-492.

FAND I. Goosens, J. J Taeldeman en G. Verleyen 1998. *Fonologische Atlas van de Nederlandse Dialecten. Deel I: Het korte vocalisme*. Gent: KANTL.

FAND II & III. Goosens, J. J Taeldeman en G. Verleyen 2000. Fonologische Atlas van de Nederlandse Dialecten. *Deel II: De Westgermaanse vocalen in open syllaben. Deel III: De Westgermaanse lange vocalen en diftongen*. Gent: KANTL.

FAND IV. De Wulf, C., J. Goossens en J. Taeldeman 2005. Goosens, J. J Taeldeman en G. Verleyen 1998. *Fonologische Atlas van de Nederlandse Dialecten. Deel IV: De consonanten*. Gent: KANTL.

Goebl, H. 2005. Dialektometrie (Art. 37), In R. Köhler et al. (eds.) *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch/An International Handbook*, Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, vol. 27) 2005, 498-531. https://www.sbg.ac.at/rom/people/prof/goebl/docs/HSK 2005 Dialektometrie.pdf

Heeringa, W. and J. Nerbonne, 2001. Dialect areas and dialect continua. *Language Variation and Change* 13, 375-400

Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenshtein distance.* PhD Dissertation University of Groningen.

Heeringa, W., P. Kleiweg, C. Gooskens, J. Nerbonne, 2006. Evaluation of String Distance Algorithms for Dialectology. In J. Nerbonne and E. Hinrichs (eds.) *Linguistic Distances Workshop at the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, July 2006. 51-62.

Kager, R., 1999. *Optimality Theory*. Cambridge; Cambridge University Press.

Kayne, R. 2005. Some notes on Comparative Syntax, with Special Reference to English and French. In G. Cinque and R. Kayne (eds.), *Oxford Handbook of Comparative Syntax*. New York: Oxford University Press, 3-69.

Kroch, A. 1989. Reflexes of Grammar in Patterns of Language Change. *Language Variation and Change* 1, 199-244.

Labov, W. 2006. *The social stratification of English in New York City*. Cambridge: Cambridge University Press [PhD Dissertation 1966].

Langacker, R. 1987. *Foundations of cognitive grammar: Theoretical prerequisites.* Stanford, CA: Stanford University Press.

Langacker, R. 2000. A dynamic usage-based model. In M. Barlow and S. Kemner (eds.), *Usage-Based Models of Language*. Stanford: CSLI, 1-63.

MAND I. De Schutter, G., B. van den Berg, T. Goeman and T. de Jong, 2005. *Morphological Atlas of the Dutch Dialects Volume I.* Amsterdam: AUP.

MAND II. Goeman, T., M. van Oostendorp. P. van Reenen, O. Koornwinder and B. van den Berg, 2009. *Morphological Atlas of the Dutch Dialects Volume II.* Amsterdam: AUP.

McCarthy J. and A. Prince, 1993, Prosodic Morphology. Constraint Interaction and Satisfaction. Rutgers University Cengter for Cognitive Science, Technical Report 3.

Michel, J.-B. et al. 2010. Quantitative Analysis of Culture using millions of digitized books. Science 331 (614), 176-182.

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago and London: University of Chicago Press.

Postma, G.J., 2010. The impact of failed changes. In C. Lucas et al. (eds.), *Continuity and Change in Grammar*. Amsterdam/Philadelphia: John Benjamins. 269-302.

SAND I. Barbiers, S., H. Bennis, G. De Vogelaer, M. Devos and M. van der Ham 2005. *Syntactic Atlas of the Dutch Dialects, Volume I.* Amsterdam: AUP.

SAND II. Barbiers, S., J. van der Auwera, H. Bennis, E. Boef, G. De Vogelaer and M. van der Ham, 2009. *Syntactic Atlas of the Dutch Dialects, Volume II.* Amsterdam: AUP.

Schmidt, J. and J. Herrgen (eds.), 2001. *Digitaler Wenker-Atlas (DiWa).* Marburg: Deutscher Sprachatlas.

Spruit, M., W. Heeringa and J. Nerbonne, 2008. Association among Linguistic Levels. *Lingua* 119, 1624-1642. [Special Issue The Forest Behind the Trees, eds. J. Nerbonne and F. Manni].

Spruit, M. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD Dissertation University of Amsterdam. LOT Dissertations 174,

Trudgill, P. 1999. Dialect contact, dialectology and sociolinguistics. *Ciioderiios de Filologio Iiigleso*, vol. 8, 1-8.

Weinreich, U. 1954. Is a structural dialectology possible? *Word* 14: 388-400.

Winford, D. 2003. *An introduction to contact linguistics*. Oxford, Blackwell.

Wright, J. 1898-1905. *English Dialect Dictionary*. 6 volumes. Oxford: Oxford University Press.

10. Summary in key words

Dialect geography, dialectometry, visualisation, theoretical linguistics, grammaticometrics

15. Summary for non-specialists (in Dutch)

Grammatica op de kaart

Het onderzoeksprogramma ‘Grammatica op de kaart’ wil met gebruikmaking van nieuwe geografische database-technieken laten zien hoe de geografische verspreiding van taalverschijnselen ons uiteindelijk iets kan leren over hoe grammatica functioneert in het menselijk brein. We doen dit door deskundigen op verschillende gebieden in onderlinge samenwerking moderne theorievorming te laten combineren met geavanceerde empirische technieken.

Al zeker sinds de negentiende eeuw heeft het gebruik van geografische kaarten een belangrijke rol gespeeld in de ontwikkeling van de taalkunde als een serieuze wetenschappelijke discipline. Met name gebieden als de dialectologie, maar ook de taaltypologie hebben traditioneel een serieuze geografische belangstelling. Ongeveer vijftig jaar geleden maakten grote delen van de taalwetenschap echter een draai waardoor op het eerste gezicht geografisch materiaal minder belangrijk werd. Dit kwam doordat de primaire belangstelling uitging naar de manier waarop de taal functioneert in de geest (of het brein) van het individu, waarmee de taalkunde dus wordt gezien als een deeldiscipline van de psychologie. De aandacht ging daarbij eerder naar het abstracte taalsysteem dat mensen gemeen hebben dan naar de taalvariatie die hen onderscheidt.

Op het eerste gezicht is er weinig reden om te denken dat geografische patronen ons iets kunnen leren over de menselijke geest. Desalniettemin is de afgelopen tien jaar de belangstelling voor taalvariatie onder grammaticatheoretici nationaal sterk toegenomen. Dit komt onder andere doordat men is gaan inzien dat juist de subtiele, al dan niet samenhangende verschillen tussen dialecten ons veel kunnen leren over wat er wel en niet mogelijk is menselijke taal. Er zijn zowel in Europa als in Amerika grote databestanden aangelegd met de resultaten van volgens nieuwe methodologische inzichten verworven dialectgegevens. Oudere verzamelingen zijn bovendien digitaal ontsloten.

In dit eerdere werk bleef de geografische dimensie tot nu toe echter nagenoeg onbelicht. Ook de atlassen die de afgelopen decennia geproduceerd zijn, worden nog benaderd alsof ze collecties van individuele dialecten betreffen, zonder dat het er veel toe doet waar welk dialect gesproken wordt. Er worden wel kaarten getekend, maar de regio’s die daarop zijn afgebeeld spelen nauwelijks een rol.

Met vier samenhangende deelprojecten wil ‘Grammatica op de kaart’ laten zien dat juist ook geografische informatie, op een goede manier statistisch bewerkt en gevisualiseerd, ons inzicht kan geven in het taalsysteem.

In het eerste deelproject gaat het om inclusierelaties: de observatie dat we sommige taalverschijnselen A alleen vinden in regio’s waar ook taalverschijnselen B zich voordoen, maar niet andersom. Vooral als er verschillende onafhankelijke regio’s zijn waarvoor we dit kunnen vaststellen, is dit een sterke aanwijzing dat de verschijnselen correleren. Interessanter nog is het dat het verschijnsel ook gecorreleerd is aan een psychologisch fenomeen: ook sprekers van dialecten die A noch B hebben, zullen een voorkeur hebben voor alleen B boven de combinatie van A en B, terwijl ze alleen A geheel en al zullen uitsluiten.

In het tweede deelproject gaat het om overgangsgebieden die geografisch tussen een gebied liggen waar we verschijnsel A vinden en een gebied met verschijnsel B. Dialecten in regio's waarin verschijnselen zo op elkaar botsen, vertonen soms gedrag dat elders onbekend is. De leidende hypothese is dat sprekers hier geconfronteerd worden met conflicterende informatie over hoe de grammatica van hun taal in elkaar moet zitten.

Het derde deelproject bestudeert de grammaticometrie, de gedachte dat het inzichtelijker is om geformaliseerde grammaticale beschrijvingen op de kaart te zetten dan om hetzelfde te doen met losse woorden of grammaticale constructies, zoals tot nu toe gebruikelijk is. Door gebruik te maken van moderne technische middelen wordt het mogelijk deze (al in de jaren vijftig geopperde) gedachte te implementeren. Daarmee wordt het meteen mogelijk om taaltheorieën die gebaseerd zijn op zulke formele grammatica's één op één te vergelijken met theorieën die gebaseerd zijn op constructies.

De eerste drie deelprojecten zijn bedoeld voor promovendi. Het vierde deelproject wordt uitgevoerd door een postdoc en ontwikkelt technisch instrumentarium dat nuttig is voor het uitvoeren van het nieuwe type onderzoek dat ons voor ogen staat. Het gaat hierbij enerzijds om het verfijnen van statistische instrumenten om te bepalen of we een aantal punten op een dialectkaart samen tot een 'gebied' mogen rekenen, en anderzijds om het ontwikkelen van visualisatiemethoden om de uitermate complexe, meerdimensionale informatie ook voor de menselijke beschouwer doorgrondelijk te maken.