Gebied Geesteswetenschappen
Subsidieaanvraag Programma

Dossier **VPR-02-46**

1. **Applicant:**
prof.dr.ir. J.Nerbonne
hoogleraar
Informatiekunde,
Faculteit der Letteren
Rijksuniversiteit Groningen
P.O. Box 716
9700 AS
Groningen

Tel. +31 (0)50 363 58 15
Tel. +31 (0)50 526 14 39 (priv.)
FAX +31 (0)50 363 68 55
email: nerbonne@let.rug.nl

2. **Title** The Determinants of Dialectal Variation

3. **Summary**
Techniques for assessing the linguistic DISTANCE between the pronunciation of Dutch dialectal varieties have been shown to be consistent and valid in adducing classifications for which expert consensus exists. The current project wishes to extend these results to Dutch lexis (vocabulary) and syntax and to German pronunciation and finally, to examine quantitative models that seek to account for the variation through dialect area (tribal history), geography and/or settlement size.

4. **Institution** Center for Language & Cognition, Groningen

5. **Program Structure**

Subproject	**Type**	**Institution**	**Supervisor**	**Researcher**
Determinants	Postdoc	CLCG, Groningen	J. Nerbonne	W. Heeringa
Syntax	OiO	Meertens, Amsterdam	H. Bennis	I. van Gemert
German	OiO	CLCG, Groningen	H. Niebaum	not yet known

In addition the proposal budgets 0.5 fte (total, i.e., 0.125fte/year) to allow the principal investigator to supervise and prepare the monograph synthesizing the results of the project.

6. Research Group

Name	Specialty	Institute	Commitment
Co-applicants: a.			
prof.dr. H.Bennis Hans.Bennis@meertens.knaw.nl Tel. +31 (0)20 4628 523	Syntactic Variation	Meertens	1 hr./week
prof.dr. H.Niebaum niebaum@let.rug.nl Tel. +31 (0)50 363 59 63	Lower Saxon	RuG	2 hr./week
Others: b.			
dr. Sjef Barbiers Sjef.Barbiers@meertens.knaw.nl	Dutch Syntax	Meertens	2 hr./week
dr. Leonie Cornips Leonie.Cornips@meertens.knaw.nl	Dutch Syntax	Meertens	1 hr./month
dr. Ton Goeman Ton.Goeman@meertens.knaw.nl	Dutch Dialectology	Meertens	1 hr./month
dr. Jan-Wouter Zwart zwart@let.rug.nl	Dutch Syntax	RuG	1 hr./month

We have asked the last three mentioned in "others" above to serve on an advisory board for the project, in particular, to attend three meetings per year. We shall want to expand this later, including at least one of the Marburg group.

7. **Period of Subsidy** Ca. 9/2003 - 8/2007

8. **Results to be Achieved (per Subproject)**

Principal Investigator The P.I. will undertake a monograph in collaboration with the postdoctoral researcher on *Determinants*. The focus will be on dialectometric techniques that can be used to characterize linguistic varieties in the aggregate. Reference will be made to the software and data, both of which will be made available publicly.

Determinants This postdoctoral researcher will collaborate in the monograph above and aim also at three journal publications on the following topics:

(a) "Lexically Based Dutch Dialect Classification", e.g., to be submitted to *Nederlandse Taalkunde*, or *Taal en Tongval*

(b) "Determinants of Dutch Dialect Variation", e.g., to be submitted to *Language Variation and Change*.

(c) "Variation in the Aggregate", together with the P.I. on the linguistic significance of this work, suitable for submission to a general linguistics journal such as *Linguistics* or *Language*.

Syntax This project should aim at producing a Ph.D. dissertation, of which two chapters should be suitable for independent journal publication:

(a) "Measuring Syntactic Distance", suitable e.g. for submission to *Language Variation and Change*.

(b) "Syntactically Based Dutch Dialect Classification", suitable for submission to *Nederlandse Taalkunde*, or *Taal en Tongval*.

German This project should aim at producing a Ph.D. dissertation, of which two chapters should be suitable for independent journal publication:

(a) "Applying Levenshtein Distance to German Dialect Pronunciations", suitable, e.g., for submission to *Dialectologia et Geolinguistica*.

(b) "Pronunciation-Based German Dialect Classification", suitable for submission to *Zeitschrift für Dialektologie und Linguistik*, or *Germanistische Linguistik*.

Occasional conference contributions, partially overlapping with these papers, may also be expected of all the participants.

Appendix. Executive Summary

Our overarching research question is simple to formulate:

What determines dialectal variation?

We aim to provide an answer to this question using a *quantitative* methodology and large amounts of dialectal data from Dutch and German. A more concrete formulation of the research goal is that we aim to determine the proportion of linguistic variation which can be attributed to various candidate factors, such as dialect area (normally associated with tribal history), geography, or the opportunity for social contact. We propose to examine several levels of linguistic structure, including in particular pronunciation, word choice and syntax. A significant second question then is the degree to which the various linguistic levels correlate.

The study of language variation has always been an important aspect of linguistic research. It provides insights into historical, social and geographical factors of language use in our society. In recent years it has also become increasingly clear that the study of language varieties such as dialects also bears upon the study into theoretical aspects of the language system. Variationist studies promise to demarcate the possible range of human language in more detail. The proposed study shares these motivations and seeks to broaden the empirical base of theoretically inspired study of variation as well to introduce quantitative methods to it.

Earlier work aimed at explanation in linguistic variation has had to focus on a small number of linguistic features, often only one feature, e.g., the Dutch pronunciation of the final unstressed <-en> syllable in words such as *Leiden, leven* and *teken* ('Leiden', 'life' and 'sign', respectively). In an effort to attain higher levels of generality, attention has been paid to the vowel systems of languages (Martinet 1952, Moulton 1962, Labov 1994), but that is the highest level of aggregation that has been possible. While this has led to insights in the linguistic organization of variation, and also to insights in the relation between language and social identity, such approaches cannot adduce the determinants of linguistic variation quantitatively.

The proposed work builds on a series of studies since 1997 in which a technique for measuring pronunciation difference has been developed, refined and validated. The section 'Background' (below) summarizes this. The key benefit of the techniques we build on is that they permit measurements in large amounts of language material to be aggregated, allowing—for the first time—a characterization of entire varieties, e.g., the pronunciation of the Dutch of Schiermonnikoog island or the Dutch of teenagers in Kerkrade.

By statistically exploring patterns of variation as functions of dialect area (tribal association), geography, or geography plus settlement size, we can determine how much linguistic variation can be explained by these factors. This will answer the primary research question.

We should like to anticipate one scruple which we have encountered in presenting these ideas. In saying we shall attempt to quantify the influence on dialectal variation, we do not imagine that our work, even if it succeeds completely, will close the book on dialect variation. When a population geneticist claims that the height of children correlates with that of their parents at the level $r = 0.63$, and therefore that parental height accounts for $r^2 \approx 40\%$ of the variation in height, this is not taken as a final answer, but rather as indication

of where detailed causal mechanisms must be sought. We can easily imagine that the successful demonstration that a larger component of linguistic variation may be attributed to social contact would, e.g., also stimulate research into the mechanisms of accommodation.

The proposed project's ambition is to harness computational power to analyze the masses of data relevant to linguistic variation. The map shown in § 9 of the grant proposal is based on 4.9×10^8 segmental comparisons, each of which involve approx. 10 phonetic features. These techniques accord with expert consensus, provide an objective basis for the notion 'dialect continuum', and allow an aggregate view of dialectal difference. We propose to explore them further in order to investigate how much linguistic variation, focusing on Dutch and German, can be accounted for by extralinguistic factors, such as area (tribal history), geographical distance, or geography plus settlement size (Trudgill's "gravitational model"). If successful, this would represent a major scientific step in the direction of an explanatory model of the sort Trudgill (1983, Chap. 3) calls for.

In addition the project promises the contributions noted in the section "Concrete Objectives", especially:

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

The background project is already making data and programs available to the community of dialectologists (see `http://www.let.rug.nl/~kleiweg/dialects/` for programs available thus far, and `http://www.let.rug.nl/~heeringa/dialectology/rnd/` for data), and the proposed project is therefore also poised to contribute practically to the general technical infrastructure of this research community.

Introduction and Goals

Please see the Appendix "Executive Summary", pp.4-5 for a general introduction.

Background

In traditional dialectology, maps are divided into dialect areas on the basis of isoglosses (divisions showing where variants are located). The dialect area maps suggest that dialect areas are the most important determinant of variation, a suggestion that cannot be substantiated using traditional methodology, however. Longstanding criticisms complain about the subjectivity of the choice of isoglosses, its "atomicity" (restriction to individual sounds or words without a way to characterize relations between entire varieties), and its coarseness—variants are alike or not, even though degrees of similarity seem so sensible that dialectologists often speak informally of a "continuum" (Bloomfield 1933, Coseriu [1]1956, 1975). Dialectometry has sought to remedy these faults by applying objective measures of differences to large amounts of dialectal material (Séguy 1971, Goebl 1982, Hoppenbrouwers & Hoppenbrouwers 1988, Herrgen & Schmidt 1989, Hummel 1991, Schmitt 1992).

The present proposal seeks in particular to continue work which has applied edit distance (also known as Levenshtein distance or string distance) to phonetic transcriptions (Nerbonne & Heeringa 1998, Nerbonne, Heeringa & Kleiweg 1999, Heeringa, Nerbonne, Niebaum & Nieuweboer 2000, Heeringa & Nerbonne 2000), measuring pronunciation differences in (now) hundreds of Dutch varieties. Because the sample is large, the method can be more general (it is limited now only by the choices made by atlas compilers). Because the technique yields a numerical measure of difference, these differences can be added, which allows one to relate entire varieties, aggregating the atomic differences. After comparing dialects on the basis of their distances the dialects can be classified by clustering or multidimensional scaling. Using clustering we get a sharp classification in the form of a tree, where the dialects are the leaves. See Figure 1 for an example. Using multidimensional scaling we get a plot on which like dialects are plotted nearby and unlike dialects are distant. When scaling to three dimensions, a map can be colored, where the dimensions are represented by the respective intensities of red, green and blue, while the areas between the sample sites are colored by interpolating from the dimensions of the dialects. In that way, dialect variation is visualized as a continuum. See Figure 2 for an example of the final output of these programs. Heeringa, Nerbonne & Kleiweg (2002) submits the dialect classification work to a rigorous statistical analysis, showing *inter alia* that the analysis is highly reliable when 100 words are involved (Cronbach's $\alpha \approx 0.95$), and that results are validated by the consensual expert opinion among dialectologists (Fowlkes-Mallows index ≈ 1.75). Current, still unpublished work analyzes Sardinian dialect variation (Bolognesi & Heeringa Accepted to appear, 2002) and the English of the American East coast—the publicly available data from the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) (Kretzschmar 1993, Nerbonne & Kleiweg 2003).

Scientific Questions and Hypotheses

Several central questions and hypotheses are to be addressed.

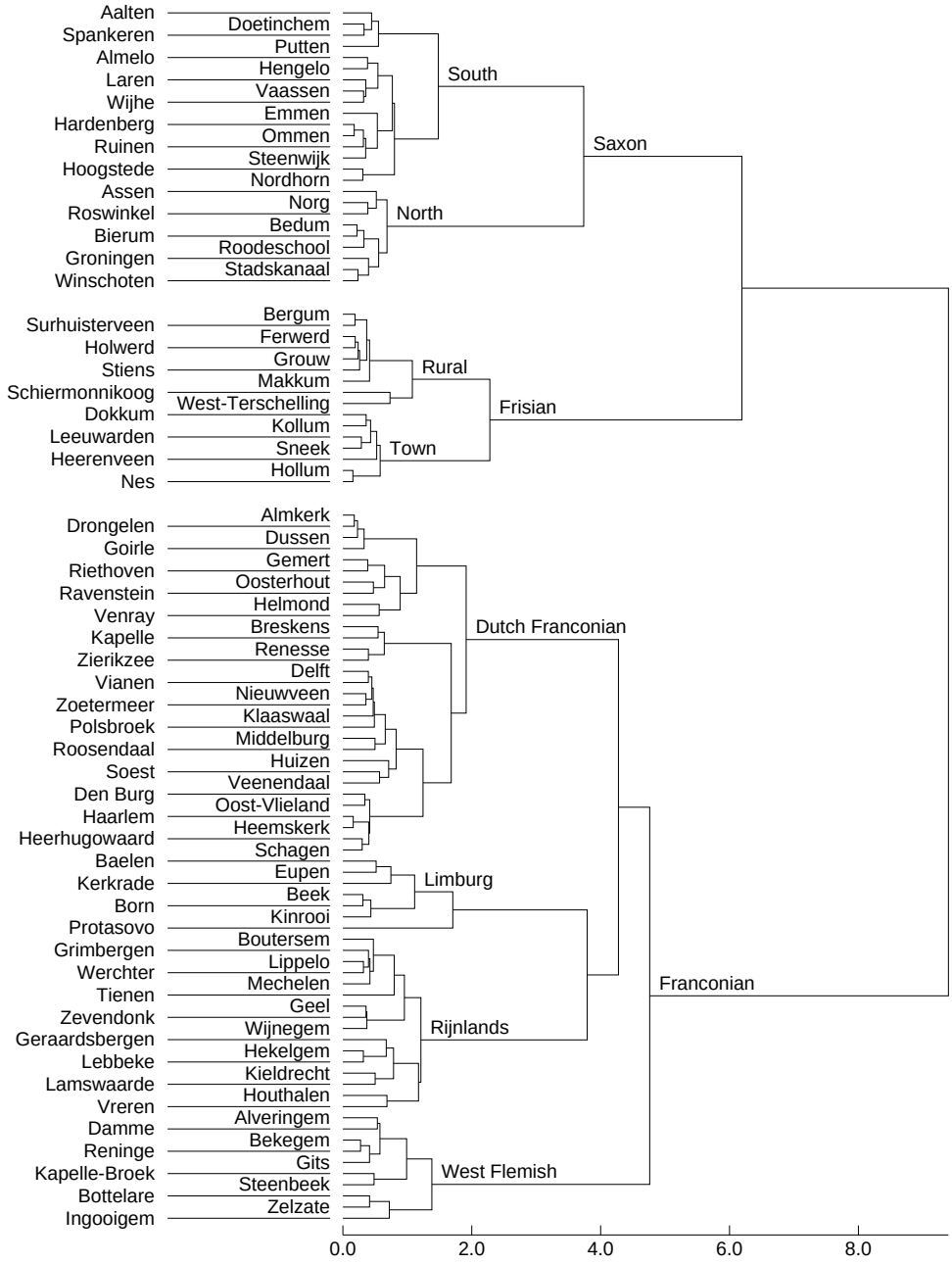

Figure 1: The result of clustering the matrix of Levenshtein distances. The traditional dialect areas (labeled) emerge distinctly. Within the Limburg cluster one finds Protosavo, an emigrant Low German variety, spoken by Siberian Mennonites and investigated by Nieuweboer (1998) (included out of curiosity).

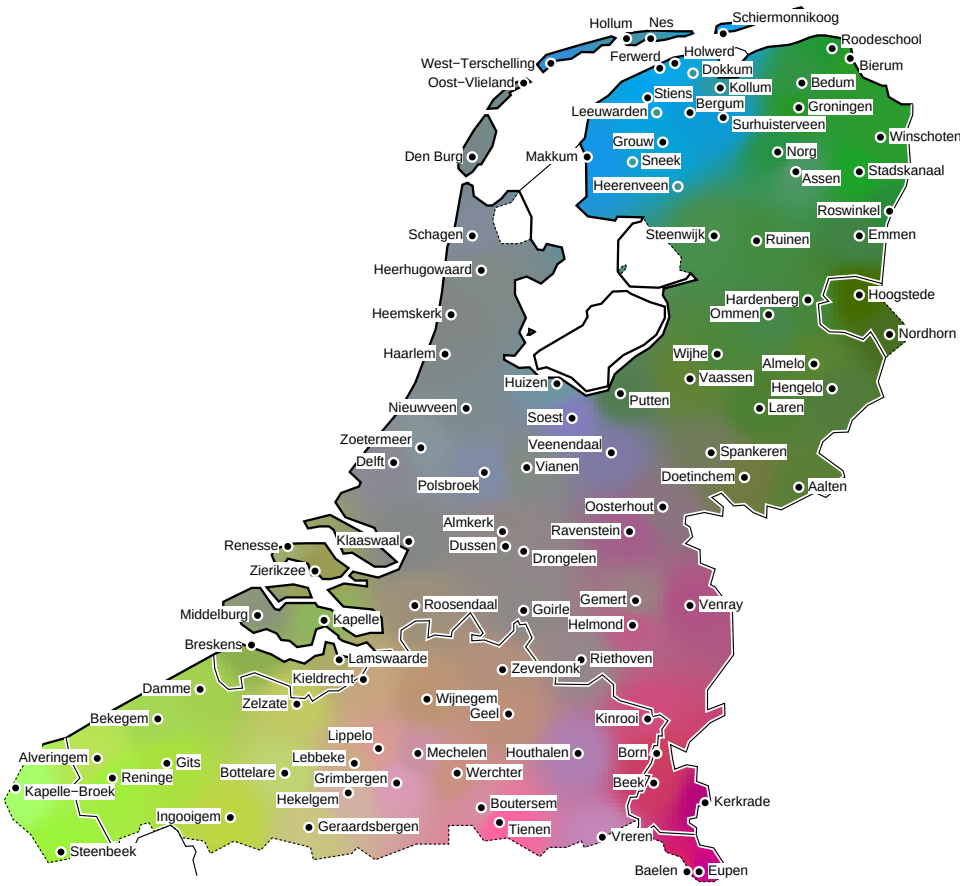

Figure 2: The most significant dimensions in average Levenshtein distance, as identified by multi-dimensional scaling, are colored red, green and blue. The map gives form to the dialectologist's intuition that dialects exist "on a continuum," within which, however significant differences emerges. The traditional distinctions among the Frisian dialects (blue), Saxon (dark green), Limburg (red), and Flemish (yellow-green) emerge when dimensions are displayed. The blue dots in white circles represent Town Frisian, a geographically distributed variety.

The focus of the proposed project is to ask what determines linguistic variation. The traditional view that seems to underlie dialect maps which differentiate areas categorically is that these areas (or the tribal history they are identified with) determine a good deal of variation. We shall test, e.g., whether geographical distance *simpliciter* is not the better predictor. In small-scale studies we have determined, e.g., that an oversimplified geographical model correlates with Dutch pronunciation variation at the level $0.65 \leq r \leq 0.70$, which means that at least $42\% - 49\%$ of pronunciation variation is accounted for by a very simplified geographical model. Finally, we shall examine Trudgill's gravity model (Trudgill 1983, pp.73-78), which incorporates distance and population size, since this gives us an estimate of the chance for social contact, which we suspect must underly dialect proximity.

A serious related question is the degree to which various linguistic levels correlate, e.g. lexical choice and pronunciation. Are the varieties which are close in pronunciation the same as those which are close lexically? There are claims that this is so (Kurath & McDavid 1961, Chambers & Trudgill 1980), but there are likewise linguistic reasons to suspect that lexical choice is more volatile than other linguistic levels, which would lead to greater variety. Nerbonne & Kleiweg (2003), quantifies the correlation at $r = 0.65$ for the American LAMSAS. We aim to quantify the degree to which the various levels correlate for Dutch, and to include syntax for the first time.

Given a large range of phenomena, we expect to see strong uniform associations among those phenomena which are linguistically dependent on one another. While strong associations which are geographically limited indicate varieties which result from social or geographical forces, uniform and near-uniform associations with no dependence on geography suggest systematic linguistic structure whose explanation should be sought in linguistic theory. This motivates the interest of theoreticians in large scale studies of (micro-) variation which promise to open a new window on the organization of language, but its exploration is likewise essential to the enterprise here. The sorts of statistical analyses we have in mind benefit from representative samples, and sample elements which are dependent on one another must be identified (and eliminated from the statistical analysis). The fact that syntax is claimed to show phenomena with such strong associations also motivated our choice of syntax as opposed to morphology, a field where data is also available, but where we do not expect the same analytical problems.

Concrete Objectives

We translate the scientific goals above into the following more tangible objectives:

- The *Determinants* subproject is charged with answering the following questions:
  (a) What are the lexical distances between the Dutch dialect varieties? How much of this variation may be attributed to the various factors under investigation?
  (b) How much of the Dutch pronunciation variation may be attributed to geography, settlement size and tribal history?
  (c) To what degree do pronunciation, vocabulary and syntax correlate?

  We foresee incidentally that this subproject will involve an application of existing techniques and software to recent Dutch pronunciation data (Goeman-Taeldeman data, see below).

- We look to the *Syntax* subproject to answer these questions:
  (a) What is a reasonable, implementable measure of the syntactic distance (and inversely, similarity) among Dutch dialect varieties? How does it guard against overweighing linguistically related phenomena?
  (b) What are the syntactic distances among the Dutch dialect varieties? (In collaboration with the **Determinants** subproject) how much of this variation may be attributed to the various factors under investigation?

  This subproject will take the form of a statistical and cartographic inquiry into the results of the SAND project (Bennis, den Besten, Devos, Rooryck & Van der Auwera 1999, Cornips & Jongenburger 2001, Barbiers, Cornips & van de Kleij 2002). It will develop software to facilitate the application of appropriate statistical analyses as well as the search for significant associations among (categorical) syntactic data.
- The *German* subproject will answer the following questions:
  (a) Does the measure of pronunciation difference developed in Groningen to compare and classify Dutch dialectal varieties function well on the unanalysed, but related language German?
  (b) What are the pronunciation distances among the German varieties?
  (c) (In collaboration with the **Determinants** subproject) how much of this variation may be attributed to geography, tribal history, and settlement size?
- Together, the subprojects should realize a set of implemented techniques for exploring dialectal data, including methods for manipulating data abstractly, for visualizing the results of these abstractions in high quality maps, and for normalizing them with respect to some sorts of individual variation. Ideally the software should provide a flexible means with which to map the geographic distribution of linguistic variables, including potentially abstract variables. The cartography should be supported by a geographic information system, such as ArcInfo or public-domain GRASS (`www.baylor.edu/grass/`), but it needs to be supplemented by statistical programs and programs for string manipulations. The subprojects will contribute to this set of techniques, and the postdoc will be charged with coordinating the work.

Added Value of Programmatic Attack

The project is naturally divided into three subprojects—a central project focused on theoretical and technical issues, and two applications involving different kinds of dialectal material, German pronunciation material on the one hand and Dutch syntactic material on the other. The added value of the having three projects work on this problem simultaneously is clear: without the two applications, the central theoretical project would be in danger of working in a vacuum, i.e., with too little attention to the real questions of the dialectologist. The applications on the other hand require a substantial theoretical and technical base which the central project will focus on. As we shall note in the third research question below, the syntax project, because of the different nature of its data, is motivated methodologically as well.

Subproject *Determinants*

This postdoc subproject, to be supervised by J.Nerbonne, will have two distinct major tasks, first to develop, implement and analyze a distance measure for lexical (vocabulary) differences, making use of the digitized pronunciation data available at Groningen and to derive and analyze lexical distances among Dutch dialect varieties; and second, to develop three geographic models and test them for the degree to which they can explain various linguistic differences quantitatively. Finally, we shall ask this project to be responsible for obtaining measures of the degree to which pronunciation, lexis and syntax correlate in linguistic variation.

The measure of lexical differences was the basis for the first work in dialectometry (Séguy 1971), and is intuitively simple: a reasonable measure of dialect distance is the fraction of vocabulary which is different. For example, one checks, for one hundred concepts, what words are used for these at various sites. If two varieties use the same words for 75 concepts, they lie then 0.25 lexical units apart. For dialect atlases which systematically record lexical differences, this measure is straightforward to apply (Nerbonne & Kleiweg 2003). There is unfortunately no systematic lexical atlas of Netherlandic dialects, no "LAND", so to speak,[1] so we shall estimate lexical distance indirectly.

When we examine the RND, we see that lexical and phonological records are mixed. There is a large number of lexical alternatives such as *kippen, hoenderen, tuutn*, for 'chickens', *vriend, kameraard* for 'friend', *timmerman, schrijnwerker* for 'carpenter', *knuppel, stok* for 'stick', *pet, muts, klak, kap* for 'cap', or *ragebol, kopstubber, koppejager, spinnekopborstel, roversbol, halfmaan, spinnejager, spinnekop* for 'mop for clearing spider webs'. When we measure the pronunciation differences among corresponding words in the RND, this mixture of levels reveals itself in an uneven, actually bi-modal distribution of phonetic distance among the distances (per linguistic item). There are many relatively close pronunciations and other, much more distant ones — where these latter result when two different lexical items are compared. For the purpose of this project, the assessment of lexical distance will proceed in this way. Naturally we shall analyze the precision of this technique by evaluating it against data.

Second, this project will explore at least three models for the explanation of linguistic differences. These models will be tested on pronunciation (Dutch and German), lexis (Dutch) and syntax (Dutch). The three explanatory models to be considered are first, the areas found in dialect maps, often associated with tribal settlement, e.g., the Frisian, Franconian and Lower Saxon areas standardly shown in maps of the Dutch-speaking area (but also more finely distinguished areas); second, geography (distance); and third, geography plus settlement size (Trudgill's "gravitational" model, Chambers & Trudgill (1980, pp.196ff), Trudgill (1983, pp.73-78)). Models will be developed for the Netherlands and (in collaboration with the **German** subproject) Germany. Each of the models will ultimately be tested by a regression analysis seeking to explain the linguistic variable on the basis of the extralinguistic ones (area, distance, or Trudgill's "gravity"). Heeringa & Nerbonne (2002) demonstrates the feasibility of the method, but on a limited (27-site) sample of pronunciations only and using the most primitive geographic notion (distance "as the crow flies"). We expect therefore each of these models to result in novel findings, and we

[1] The honorable, but geographically limited exceptions of the *Woordenboek van de Brabantse Dialecten* and the *Woordenboek van de Limburgse Dialecten* notwithstanding.

expect to generate special interest through this first application of Trudgill's model to a large amount of data.

The investigation of the relevant notion of "geography" for the purpose of explaining linguistic variation will need to incorporate an historic component since the purpose of examining geographic distance is ultimately to estimate the ease with which contact was likely, and this has varied over time. For this purpose, distances "as the crow flies" are less useful than travel distances, but these have varied as wetlands were drained and bridges were built. Van Gemert (2002) has conducted an initial experiment in this direction, using a geographic information system (GIS), to estimate travel time before the construction of railroads, highways or the causeway linking North Holland to Frisia. Information about wetlands still needs to be incorporated.

A final task which we shall ask of this subproject is to evaluate the degree to which three linguistic levels correlate, viz., pronunciation, lexis and syntax, all on the basis of Dutch material. We can evaluate pronunciation and lexis on the basis of RND material, digitized at Groningen (see `http://www.let.rug.nl/~heeringa/dialectology/rnd/`), and syntax on the basis of the Syntax subproject. Since this material varies with respect to the time at which it was collected and also with respect to the selection of the sites, we shall undertake a second analysis comparing the digitally available data from the Goeman-Taeldeman project (from *Fonologische Atlas van de Nederlandse Dialecten* (FAND) (Goossens, J.Taeldeman & G.Verleyen 1998)) with the SAND material (see the Section below on syntax). By design SAND uses approx. 125 of the sites in the FAND. To measure the correlation between lexis and syntax will involve comparing the RND and SAND data, and thus a loss of exact temporal commensurability, and will also require that we sometimes identify closest geographic matches where data has been gathered from different sites. Fortunately, both data sets are dense enough so that we expect no difficulty in the last step.

Subproject *Syntax*

The major objective of this graduate student subproject, to be supervised by H.Bennis, is to develop and apply a quantitative measure of syntactic difference between geographic varieties of Dutch, in itself an innovative step. Second, we shall analyze dialect syntax using the same classificatory tools developed in Groningen for pronunciation, esp. clustering and MDS. This step will not require great innovation, but it will result in a dialect classification based on syntax, one of the first of its kind. And third, in collaboration with subproject **Determinants**, this subproject will ask what the determinants are, using the models derived from area, distance and Trudgill's "gravity".

Data for this project will be obtained from the SAND project (VNC, 1/1/2000–31/12/2003, see `http://www.meertens.nl/projecten/sand/sand.html`), which describes and analyzes the syntactic variation in the Dutch dialects spoken in The Netherlands and Flanders (Cornips & Jongenburger 2001, Barbiers et al. 2002). Sjef Barbiers is the SAND project manager and member of the research group of this grant application. The SAND's annotated database contains the results of oral interviews at 250 evenly distributed sites in The Netherlands and Flanders (of which approx. 50% were also used in the FAND). The database contains over 100 sentences per interview and codifies 125 points of syntactic variation, concentrating on four empirical domains: (i) the left-periphery of the clause; (ii) the right-periphery of the clause; (iii) negation

and quantification; (iv) pronominal reference. The left-periphery includes phenomena such as complementizer agreement, subject doubling, verb third, pronominal subject drop, complementizer omission, the *Imperativus pro Infinitivo* effect. The right-periphery involves phenomena such as varying word order in verb clusters, verb cluster interruption, presence and placement of the infinitival marker *te*, the *Infinitivus pro Participio* effect, the *Participium pro Infinitivo* effect. Variation in negation includes negative concord, the presence of a negative particle, clause final negation. Finally, variation in pronominal reference includes different forms of anaphoric and pronominal elements correlating with different syntactic distributions.

We view syntax as characterized by these 125 nominal variables, and begin with the idea that we measure the difference in the 125-element sequence as the complement of overlap, corrected to reflect the chance of overlap in nominal variables which include more than two categories (thus the overlap in a 4-valued variable counts twice as much as the overlap in a 2-valued variable).

The development of a measure of syntactic difference must overcome a significant conceptual hurdle. Whereas we treat the data items in vocabulary overlap and pronunciation difference as independent, this looks inappropriate in the case of syntax. There is consensus among syntacticians that some syntactic variables are strongly associated because they reflect the same fundamental structures realized in a variety of superficial positions.[2] We might refer to variables which covary because they reflect the same fundamental linguistic structures as LINGUISTICALLY RELATED in contrast to those which covary geographically, which we might refer to as GEOGRAPHICALLY COHESIVE.

Theoretical literature on Dutch microsyntax has demonstrated that at least part of the syntactic similarity between dialects must be attributed to linguistically related features, in fact there are candidates in each of SAND focal areas.

left periphery Hoekstra & Smits (1997) claim that complementizer agreement is identical to the agreement morpheme on the finite verb in inversion, and that it only occurs in varieties in which there is no difference between agreement on the finite verb in the present and the past tense. Van Craenenbroeck and Van Koppen suggest that there may be counterexamples, however (personal communication).

right-periphery Vanden Wijngaerd (1996, and references there) claims that there is a correlation between the absence of the participial prefix *ge-*, the possibility of the word order V1 V3 V2 and the possible absence of the *Infinitivum pro Participium effect*. In Standard Dutch *Jan had kunnen komen* lit. 'John had can-inf come-inf' is the only grammatical option, with the infinitival form of *kunnen* rather than the participial form. A number of Northeastern dialects, however, have *Jan had komen kund* lit. 'John had come could', with the participial form of *kunnen* and the main verb preceding it. This participial form is claimed to be allowed only in those dialects that have no *ge-* prefix on participles and that allow the main verb to precede the second auxiliary.

[2]In fact the same point might be made about pronunciation, but is normally countered by the consideration that frequently realized variants deserve a proportionally heavier weight. This, coupled with the belief that phonological samples are representative, obviates the need for weighting schemes in pronunciation. But the syntactic material was collected partly with the goal of detecting linguistically conditioned microvariation, so there is a special danger of overrepresenting strongly associated features.

pronominal reference The possibility of ONE-pronominalization with full nouns, as in Northern Brabantish *Gij zijt ok unnen raren mens één* lit. 'you are also a strange man one' is claimed to correlate with full agreement between the numeral *één* 'one' and the rest of the noun phrase (Barbiers & Greijmans 2002).

negation It is claimed that varieties with a negation particle are always varieties that also have negative concord (Zeijlstra 2002).

To avoid weighting related variables too heavily, we must investigate their association and perhaps discount linguistically related variables, e.g., through an information-gain based weighting (see Nerbonne & Heeringa (1998) for an application of this sort of weighting to a phonological problem). It is impossible to examine *all* logically complex dependencies, such as the example from the right periphery, since there are too many possibilities, but we can examine all of the proposed dependencies. To contrast claims of linguistic relatedness, we may make use of a measure of geographic cohesion (Nerbonne & Kleiweg 2003).

This subproject is poised to contribute a first measure of syntactic distance between Dutch varieties, a first classification of varieties on the basis of syntax, and, in collaboration with the **Determinants** subproject, estimates of the factors which might explain this variation. Finally, the project is particularly valuable because it requires attention to the potentially confounding effect of deeper linguistic structure in the program of understanding the geographical distribution of variation. It is a fortunate circumstance that syntactic theoreticians are likewise interested in teasing apart those aspects of (co-)variation that may have a structural linguistic explanation. This may provide a glimpse at the division of labor between nature (theoretically determined correlation) and nurture (distance) in the domain of linguistic variation.

Subproject *German Pronunciation*

This graduate student project, to be supervised by H.Niebaum, will replicate the extensive dialectometric studies performed on Dutch varieties on an interesting database of German transcriptions of impeccable quality. In cooperation with the *Determinants* subproject, this project will likewise investigate the degree to which area, distance and Trudgill's "gravity" model may account for the pronunciation variety found.

Through an informal cooperation with the University of Marburg initiated by Hermann Niebaum (Groningen), and involving John Nerbonne and Rogier Nieuweboer (Groningen) and Angelika Braun and Hermann J. Künzel (*Deutscher Sprachatlas*, Marburg) a large database of phonetic transcriptions of German dialect transcriptions is currently being digitized in Groningen. The West German data was collected by Georg Heike at the end of the 1960's and beginning of the 1970's, and supplemented by East German data collected by Joachim Göschel after the German reunification (completed in 1992). The goal of the collection was to record and preserve German dialectal variation, with special interest in providing material comparable to the Wenker atlas, but collected roughly one century later. Each data set was recorded on tape and transcribed by professional phoneticians, who then consulted to resolve discrepancies. The entire data set consists of 178 words as they are pronounced at 180 locations evenly distributed throughout Germany. Rogier Nieuweboer is overseeing the digitization in Groningen (translation into XSAMPA). Each

data set is being digitized twice, once by a secretary and once by a student assistant. Nieuweboer will check this work, focusing on points of discrepancy between the two digitizations. We expect therefore that the researcher in this project will begin with excellent data, ready for analysis.

The first task of this researcher will therefore be to analyze the data using software available in Groningen for the measurement of dialect pronunciation differences based on phonetic transcription, and tested on Dutch and American data (see `http://www.let.rug.nl/~kleiweg/dialects/`for description of software and opportunity to download). This involves examining the natural clusters in the set of sites based on average pronunciation difference, extracting the most important dimensions of variation (Multi-Dimension Scaling), and examining which pronunciation differences characterize the clusters and dimensions most reliably. In the course of the current project, with its focus on validating techniques developed for the analysis and classification of Dutch, special attention will be paid to the treatment of data items which might be considered non-independent and which therefore could confound the statistical analysis of the data (regression analysis). Our leading hypothesis about this possibility is that including related items may result in a weighting that reflects frequency, and that this positive effect compensates for the potentially confounding effects of including non-independent items. The final step is important to relate this work to traditional dialectology.

The second task will be to create explanatory models based on dialect area (Bavarian, Alemannic, Rhineland, etc.), distance and Trudgill's "gravity" model. This is relatively novel work for a graduate student, but he or she can rely on Hermann Niebaum for expertise on German dialects and the **Determinants** subproject for expertise in statistical analysis and GIS (see that project description). Naturally, this task includes the evaluation of the contribution of the model toward the explanation of dialect variation, i.e., testing the hypotheses that area, distance and "gravity" contribute an explanation of dialectal variation. The interest analysing the geographical contribution to linguistic variation in Germany as well as in the Netherlands is potentially great, given the different geographical situation in Germany, in particular the mountainous areas of the south.

The project will also be in a position to compare the competing perspectives of *dialecticality* (Herrgen & Schmidt 1989) vs. interdialectal distance (Nerbonne et al. 1999) as basis for large-scale comparison.

References

Barbiers, Sjef, Leonie Cornips & Susanne van de Kleij, eds (2002), *Syntactic Microvariation*, Meertens Institute/NIWI, Amsterdam. elec. publication, `http://www.meertens.knaw.nl/books/synmic`.

Barbiers, Sjef & Martine Greijmans (2002), 'Microvariation in the syntax of one-insertion', *GLOW Newsletter* **48**, 15–16. Paper presented at GLOW 25, Amsterdam/Utrecht.

Bennis, Hans, Hans den Besten, Magda Devos, Johan Rooryck & Johan Van der Auwera (1999), Syntactische atlas van de nederlandse dialecten. VNC grant proposal.

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

Bolognesi, Roberto & Wilbert Heeringa (Accepted to appear, 2002), 'De invloed van dominante talen op het lexicon en de fonologie van sardische dialecten', *Gramma/TTT* .

Chambers, Jack & Peter Trudgill (1980), *Dialectology*, Cambridge University Press, Cambridge.

Cornips, Leonie & Willy Jongenburger (2001), 'Het design en de methodologie van het SAND project', *Nederlandse Taalkunde* **16**, 215–232.

Coseriu, Eugenio ([1]1956, 1975), *Die Sprachgeographie*, Gunter Narr, Tübingen.

Gemert, Ilse van (2002), Het geografisch verklaren van dialectafstanden met een geografisch informatiesysteem (GIS), Master's thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Goebl, Hans (1982), *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichischen Akademie der Wissenschaften, Wien.

Goossens, J., J.Taeldeman & G.Verleyen (1998), *Fonologische Atlas van de Nederlandse Dialecten, Deel I*, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.

Heeringa, Wilbert & John Nerbonne (2000), Change, convergence and divergence among Dutch and Frisian, *in* P.Boersma, Ph.H.Breuker, L.G.Jansma & J. van der Vaart, eds, 'Philologia Frisica Anno 1999. Lezingen fan it fyftjinde Frysk filologekongres', Fryske Akademy, Ljouwert, pp. 88–109.

Heeringa, Wilbert & John Nerbonne (2002), 'Dialect areas and dialect continua', *Language Variation and Change* **13**(2), 375–398.

Heeringa, Wilbert, John Nerbonne, Hermann Niebaum & Rogier Nieuweboer (2000), Measuring Dutch-German contact in and around Bentheim, *in* D. Gilbers, J. Nerbonne & J. Schaeken, eds, 'Languages in Contact', Rodopi, Amsterdam-Atlanta, pp. 145–156.

Heeringa, Wilbert, John Nerbonne & Peter Kleiweg (2002), Validating dialect comparison methods, *in* W. Gaul & G. Ritter, eds, 'Classification, Automation and New Media: Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation', Springer, Heidelberg, pp. 445–452.

Herrgen, Joachim & Jürgen Erich Schmidt (1989), Kontrastive dialektgeographie, *in* W. Putsche, W. Veith & P. Wiesinger, eds, 'Dialektgeographie und Dialektologie', Vol. 90 of *Deutsche Dialektgeographie*, N.G.Elwert Verlag, Marburg, pp. 304–346.

Hoekstra, Erik & Caroline Smits (1997), 'Vervoegde voegwoorden in de nederlandse dialecten', *Cahiers van het P. J. Meertens-Instituut* **9**, 6–30.

Hoppenbrouwers, Cor & Geer Hoppenbrouwers (1988), 'De featurefrequentiemethode en de classificatie van nederlandse dialecten', *TABU: Bulletin voor Taalwetenschap* **18**(2), 51–92.

Hummel, Lutz (1991), *Dialektometrische Analysen zum kleinen deutschen Sprachatlas (KDSA)*, Max Niemeyer Verlag, Tübingen.

Kretzschmar, William A. (1993), *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, The University of Chicago Press, Chicago.

Kurath, Hans & Raven I. McDavid (1961), *The Pronunciation of English in the Atlantic States*, University of Michigan Press, Ann Arbor.

Labov, William (1994), *Principles of Linguistic Change. Vol. 1: Internal Factors*, Blackwell, Oxford.

Martinet, André (1952), 'Function, structure and sound change', *Word* **8**, 1–32.

Moulton, William (1962), 'Dialect geography and the concept of phonological space', *Word* **18**, 23–32.

Nerbonne, John & Peter Kleiweg (2003), 'Lexical explorations in LAMSAS', *Computers and the Humanities* **37**, xxx–yyy. in preparation, 8/02, sketch available at `www.let.rug.nl/ kleiweg/lamsas/finland/finland.pdf`.

Nerbonne, John & Wilbert Heeringa (1998), 'Computationele vergelijking en classificatie van dialecten', *Taal en Tongval* **50**(2), 164–193.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg (1999), Edit distance and dialect proximity, *in* D. Sankoff & J. Kruskal, eds, 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.', CSLI, Stanford, CA, pp. v–xv.

Schmitt, Ernst Herbert (1992), *Interdialektale Verstehbarkeit*, Franz Steiner Verlag, Stuttgart.

Séguy, Jean (1971), 'La relation entre la distance spatiale et la distance lexicale', *Revue de Linguistique Romane* **35**, 335–357.

Trudgill, Peter (1983), *On Dialect. Social and Geographical Perspectives*, Blackwell, Oxford.

Vanden Wijngaerd, Guido (1996), Participles and bare argument structure, *in* W. Abraham, S. Epstein, H. Thrainsson & J.-W. C. Zwart, eds, 'Minimal Ideas: Syntactic Studies in the Minimalist Framework', Benjamins, Amsterdam, pp. 283–304.

Zeijlstra, Hedde (2002), What the Dutch Jespersen cycle may reveal about negative concord. Unpublished manuscript, University of Amsterdam.

9. **Work Program**

The project plans three annual meetings, alternating at Groningen and at the Meertens Institute. The researchers employed by the project, the supervisors and the rest of the research group (p.2, item 6) will attend.

Determinants

Year	Tasks to be completed
1.	Develop & apply lex. dist. measure
2.	Develop & apply three geographical models (Dutch)
3.	Prepare Goeman-Taeldeman data, investigate corr. of ling. levels, collaborate with **Syntax, German** subprojects on geography.

Syntax

Year	Tasks to be completed
1.	Familiarity with SAND, develop programs to extract features from SAND
2.	Investigate linguistic relatedness among syntactic features, develop corrections for associated features in syntactic distance measures.
3.	Produce classification of varieties based on syntax. Explore geographical models with **Determinants** subproject.
4.	Ph.D. Thesis

German Pronunciation

Year	Tasks to be completed
1.	Familiarity with German dialectology, Marburg data, Groningen dialectometrical programs.
2.	Classification, including comparison to alternative proposals, comparison to successful techniques for Dutch.
3.	Given best classification, explore geographical models with **Determinants** subproject.
4.	Ph.D. Thesis

10. **Resubmission** This is NOT a resubmission, but a proposal by the same group was rejected in the *vooraanmelding* phase in June, 2001.

11. **Other financial contributions** are not expected.

12. **Brief CV Applicant**

Education

Ph.D., Linguistics, 1984. Ohio State University. Diss: *German Temp. Logic.*
M.S., Computer and Information Science, 1984. Ohio State University.

Employment

1/99-present Director, Center for Language and Cognition, Groningen. Research dir. 53 faculty members, 30 postdocs, Ph.D. students.

2/93-present University of Groningen. University Professor (*hoogleraar B*) of Computational Linguistics and Chair of *alfa informatica* (Humanities Computing). Secondary appt. in Computer Science.

4/90-1/93 German Research Center for Artificial Intelligence

7/85-3/90 Hewlett-Packard Laboratories. Adjunct Asst. Prof., Stanford University (Symbolic Systems).

Professional Activities President, Association for Computational Linguistics (ACL), 2002; Coordinator, TMR postdoc network in *Learning Computational Grammars*, 1998-2002 (M€ 2.6); Columnist, ELSNews, 2001 (Newsletter for European Language and Speech Network of Excellence; Chair, European Chapter of ACL, 1997-98; Member of Steering Committee for Speech, Language and Logic, National Science Foundation of the Netherlands, 1994-1998; Associate Editor, *Computational Linguistics*, 1992-94.

Dissertations supervised

Formalizing the Minimalist Program Mettina Veenstra, 11 Jun 1998
The Prize of Neutrality (Comp. History) George Welling, 25 Jun 1998
Machine Learning of Phonotactics Erik Tjong Kim Sang, 19 Oct 1998
Connectionist Lexical Processing Ivilin Stoianov, 23 Mar 2001
Dialogue-Based Disambiguation Rob Koeling, 25 Jan 2002
Structured Features in Maximum-Entropy Parsing, Tony Mullen 22 Mar 2002
and several students in early stages of theses.

Other Books (see application for core dialectological pubs.)

John Nerbonne, Klaus Netter and Carl Pollard (eds.) *German in Head-Driven Phrase Structure Grammar.* Stanford: CSLI, 1994.
John Nerbonne. *Electronische Incunabula.* Inaugural lecture, Sept. 1995
Sake Jager, John Nerbonne and Arthur van Essen (eds.) *Language Technology and Language Teaching* Lisse: Swets and Zeitlinger, 1998.
Dicky Gilbers, John Nerbonne and Jos Schaeken (eds.) *Languages in Contact*, Rodopi: Amsterdam, 2000.

Core Publications Applicant

Nerbonne, John with Wilbert Heeringa and Peter Kleiweg. 1999. Edit distance and dialect proximity, *in* D. Sankoff & J. Kruskal, eds, 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison', CSLI, Stanford. pp.v-xv.

Nerbonne, John & Wilbert Heeringa. 1998, 'Computationele vergelijking en classificatie van dialecten', *Taal en Tongval* **50(2)**, pp.164-93.

Wilbert Heeringa, John Nerbonne, Hermann Niebaum, Rogier Nieuweboer, and Peter Kleiweg. Dutch-German Contact in and around Bentheim. In: Dicky Gilbers, John Nerbonne and Jos Schaeken (eds.) *Languages in Contact.* (= *Studies in Slavic and General Linguistics* 28). Amsterdam: Rodopi. 2000. pp. 145–156.

Wilbert Heeringa, John Nerbonne, and Peter Kleiweg. Validating Dialect Comparison Methods. In: Wolfgang Gaul and Gerd Ritter (eds.) *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation.* Heidelberg: Springer. 2001. pp. 445-452.

Wilbert Heeringa and John Nerbonne. Dialect Areas and Dialect Continua. In: *Language Variation and Change* 13, 2002. pp. 375-398.

13. **Motivation for Choice of Researcher**

drs. Wilbert Heeringa completed his *doctorandus cum laude* in *Alfa-informatica* in Groningen in 1997, receiving the *Scriptieprijs* from the *Nederlands Genootschap voor Informatietechnologie.* He is a solid programmer. He worked as an assistant to J.Nerbonne, developing material for statistics instruction during 1997-98, entering a five-year graduate program in Aug. 1998. During his graduate student tenure he has regularly assisted in teaching statistics, further developing knowledge which is likewise essential to the work proposed here.

In addition to the five co-authorships noted in the core publications of the applicant, Heeringa has had *eight* further papers accepted for publications, bringing to a total of six his papers in the most important journals for Dutch dialectology and for methods in dialectology and variationist linguistics. He has two more papers in preparation for submission to a special issue of *Computers and the Humanities* on computational methods in dialectometry, which, together with those on his CV, constitute the scientific contribution of his Ph.D. thesis, likewise in preparation. This is an unusually productive graduate student career.

Heeringa's work is invariably thorough in exploring many alternatives, creative in exploring techniques of analysis, and honest in reporting qualifications and shortcomings. There is no better candidate for this position.

drs. Ilse van Gemert recented completed her *doctorandus* in *Informatiekunde* (formally *Alfa-informatica*) with a course grade average of over $7,5$. She worked as an student assistant in statistics laboratories, and completed an internship at the Meertens institute making dialect databases accessible via the world-wide web. Her Master's thesis estimated travel time in the Netherlands before the time of the railroads, highways and the causeway linking North Holland to Frisia, and investigated the degree to which this correlates with pronunciation difference as measured

by Nerbonne and Heeringa, so that she is familiar with the goals and basic methods of the project proposed here.

She has been an intelligent and reliable student in Groningen and is enthusiastic about a scholarly career. As of Sept. 1, 2002 she will work part-time as a programmer at the Meertens institute and hopes to continue her dialectometry work in some capacity in Groningen. Ms. van Gemert is an excellent candidate for this position.

14. **Budget** *K€* 453. One postdoc *K€* 160, two graduate students at *K€* 130 each, one replacement for P.I. Nerbonne 0,5 fte year at *K€* 25, travel and material costs of *K€* 8.

Personnel

Year	PD-Mnth	OiO-Mnth	Total
2003	4 (1,0)	8 (1,0)	12 (1,0)
2004	12 (1,0)	24 (1,0)	36 (1,0)
2005	12 (1,0)	24 (1,0)	36 (1,0)
2006	8 (1,0)	24 (1,0)	32 (1,0)
2007		16 (1,0)	16 (1,0)
Totals	36 (1,0)	96 (1,0)	132 (1,0)

Material and Travel

Year	Material	Travel	Total
2003	€800	€400	€2.000
2004	800	1.200	2.800
2005	800	1.200	2.800
2006	800	1.200	2.800
2007		800	800
Totals	€3.200	€4.800	€8.000

15. **Budgetary Remarks**

- **Replacement Salary** is requested for P.I. Nerbonne to allow time for supervision and for preparation of monograph synthesizing project results.
- **Material Costs** 1 4-year license for ArcInfo (geographic analysis package) @ €800 = €3.200
- **Travel Costs** 2 trips/year Marburg for OiO in **German** 1 trip/year for supervisor (Niebaum or Nerbonne) @ €400/trip (travel and lodging for two days/trip). Total: €4.800,–
- **Travel** Meertens – Groningen. *p.m.*
- Conference travel will be applied for separately.