

**QUANTITATIVE PERSPECTIVES  
ON SYNTACTIC VARIATION  
IN DUTCH DIALECTS**

*Published by:*

LOT  
Janskerkhof 13  
3512 BL Utrecht  
The Netherlands

phone: +31 30 253 6006  
fax: +31 30 253 6406  
e-mail: [lot@let.uu.nl](mailto:lot@let.uu.nl)  
<http://www.lotschool.nl>

*Cover illustration:* The SAND MDS map (shown on the right) is based on 1182 syntactic variables and visualises 485 SAND1 variables (shown on the left) and 697 SAND2 variables (shown in the middle) in the aggregate based on a Hamming distance measure. See Figure 6-12 on page 135 for details.

ISBN 978-90-78328-48-3

NUR 616

Copyright © 2008: Marco René Spruit. All rights reserved.

**QUANTITATIVE PERSPECTIVES  
ON SYNTACTIC VARIATION  
IN DUTCH DIALECTS**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van de Rector Magnificus prof.dr. D.C. van den Boom  
ter overstaan van een door het College voor Promoties ingestelde commissie,  
in het openbaar te verdedigen in de Aula der Universiteit

op  
woensdag 26 maart 2008, te 10:00 uur

door  
MARCO RENÉ SPRUIT  
geboren te Ermelo

## Promotiecommissie

Promotores:        prof.dr. H.J. Bennis  
                          prof.dr.ir. J. Nerbonne

Co-promotor:      prof.dr. L.C.J. Barbiers

Overige leden:     prof.dr. J.B. den Besten  
                          dr. M.J. Dunn  
                          prof.dr. P.C. Hengeveld  
                          prof.dr. R.W.N.M. van Hout  
                          mevr.prof.dr. F.M.G. de Jong  
                          prof.dr.ir. R.J.H. Scha  
                          prof.dr. F.P. Weerman

Faculteit:            Faculteit der Geesteswetenschappen

# Contents

Foreword.....	vii
1. Introduction.....	11
1.1. Motivation.....	11
1.2. Dialectological context.....	15
1.2.1. Dialect cartography.....	15
1.2.2. Dialectometry.....	19
1.2.3. Syntactic microvariation.....	22
1.3. Research dimensions.....	27
1.4. Research questions.....	29
1.5. Chapter overview.....	29
2. Dutch dialect area classifications based on aggregate syntactic differences.....	33
2.1. Introducing the dialect classification problem.....	33
2.2. Classifying Dutch dialects using subjective judgements.....	34
2.3. Combining dialectometry and syntactic variation.....	35
2.4. Measuring syntactic variation using Hamming distance.....	36
2.5. Analysing dialect distances using multidimensional scaling.....	38
2.6. Classifying Dutch dialects using a syntactic measure.....	39
2.7. Comparing the computational and perceptual dialect classifications.....	42
2.8. References.....	43
3. Measures of syntactic distance and the role of geography.....	45
3.1. Introduction.....	45
3.2. Syntactic Atlas of the Dutch Dialects.....	47
3.3. Hamming Distance Measure.....	49
3.4. Multidimensional Scaling Analysis.....	50
3.5. Map of the Dutch Dialects.....	51
3.6. Syntactic variation in context.....	52
3.6.1. Syntax versus Perception.....	52
3.6.2. Syntax versus Pronunciation.....	53
3.6.3. Syntax versus Geography.....	54
3.7. Feature Variables.....	56
3.8. Atomic Variables versus Feature Variables.....	58
3.9. Conclusions.....	61
3.10. Future Research.....	61
3.11. References.....	62
3.12. Appendix.....	63
4. Associations among linguistic levels.....	65
4.1. Introduction.....	65
4.2. Research questions.....	67
4.3. Data sources.....	68
4.4. Distance measures.....	70

4.5. Dutch dialect area perspectives .....	73
4.6. Consistency .....	78
4.7. Correlations between linguistic levels.....	79
4.8. Linguistic levels correlated with geography .....	81
4.9. Linguistic correlations without the influence of geography .....	83
4.10. Conclusions.....	85
4.11. Discussion and future research .....	86
4.12. References .....	88
5. Discovery of association rules between syntactic variables.....	91
5.1. Introduction.....	91
5.2. Syntactic variation database .....	93
5.3. Sample data illustration and diagram .....	95
5.4. Association rule mining based on proportional overlap .....	95
5.5. Evaluating the quality of a rule.....	98
5.6. Discovery of association rules between syntactic variables.....	100
5.7. Data mining the Syntactic Atlas of the Dutch Dialects .....	101
5.8. Conclusions.....	106
5.9. Discussion .....	107
5.10. References .....	108
6. Summary and conclusions .....	111
6.1. Chapter summary .....	111
6.2. Conclusions in questions and answers .....	115
6.3. Directions for future research .....	120
6.3.1. Alternative measures of syntactic distance.....	120
6.3.2. Incorporation of SAND2 data.....	127
References.....	137
Relevant software.....	143
List of figures.....	145
List of tables .....	147
List of terms.....	149
Nederlandse samenvatting.....	151
Curriculum vitae.....	157

## Foreword

There are many reasons why I have enjoyed this PhD research so much. Perhaps most importantly, a PhD project presents the unique opportunity to delve into uncharted intellectual territory for four consecutive years. Time to explore new ideas and to integrate them meaningfully in existing knowledge. Having mostly worked on relatively small projects as an independent software developer for a number of years, it has been a true pleasure to finally be able to investigate a problem in depth without being limited by various mundane restrictions—i.e. time and money. The global nature of scientific research constitutes another attractive aspect of performing a PhD research. I have presented this work in the context of conferences, extended visits, seminars, workshops and student courses in Canada, Italy, France, Belgium and throughout the Netherlands. Additionally, I have been able to attend various summer and winter schools, conferences and workshops in Italy, France and throughout the Netherlands. I hope that—in the context of this work—a positive correlation between my research results and travel time may be found.

There are many people who I wish to thank for their support during the last four years. First of all, I would like to gratefully acknowledge the continuous guidance, support and feedback that I have received from my promotores **Hans Bennis** and **John Nerbonne**, and my co-promotor **Sjef Barbiers**. I have always felt truly fortunate to be able to further develop my scientific research skills under the wings of your esteemed expertise. Thanks, Hans, John and Sjef!

I have performed my research in the context of the *Determinants of Dialectal Variation* (DDV) project. I would like to thank the core DDV members at the University of Groningen—**Wilbert Heeringa**, **Charlotte Gooskens**, **Reneé van Bezooijen** and **Hermann Niebaum**—for their many valuable insights from which this work has benefited significantly. I especially wish to thank Wilbert Heeringa for his enduring and accurate support and feedback in all dialectometrical aspects of my research. I also want to thank **Peter Kleiweg** for e-sharing his cartographic expertise with me, in relation to his excellent but unpronounceable dialectometrical software package *RuG/L04*. Thanks, Wilbert, Charlotte, Reneé, Hermann and Peter!

During spring 2006, I was fortunate to be able to visit the Linguistics department of the University of Trieste. This period has certainly been one of the highlights of my research. I wish to thank **Giuseppe Longobardi** for his generous hospitality and support to make my stay as worthwhile and pleasant as possible. I also thank **Gabriele Rigon** for our many walks and talks in and around downtown Trieste, and his skilled help in battling Italian bureaucracy. Similarly, I wish to thank **Cristina Guardiano** for warmly welcoming and guid-

ing me through the cities and universities of Modena and Reggio Emilia. I also would like to thank **Chiara Gianollo** for our professional and personal discussions. Thanks, Pino, Gabriele, Christina and Chiara!

At the Meertens Instituut, I shared room 1.76 with **Margreet van der Ham** and **Alies MacLean**. No secret was ever safe in this social centre of the institute! It has been a lot of fun, getting to know you. Furthermore, Margreet regularly helped me out during my struggles to better understand the SAND data. Of course, I would like to thank all other Meertens colleagues as well. You have made me feel at home from day one. I have always especially enjoyed our daily lunchtime discussions. Thanks, Margreet, Alies, Mathilde, Reina, Hilje, Jan-Pieter, and everyone else I should mention here!

For the sake of completeness, I also wish to thank everyone else who somehow contributed to the current work in any way. I obviously haven't credited you appropriately in these acknowledgements. Truthfully thanks, I'm sloppily sorry!

Oddly enough, this dissertation would most likely not exist today if I hadn't been unorthodoxly manoeuvred towards this position by **Edwin Brinkhuis**. Back in spring 2003, Edwin was the software development coordinator at the Meertens Instituut. He arranged an introductory meeting for me with Sjef Barbiers. I was an independent software entrepreneur at the time and always interested in challenging projects. Edwin advised me to introduce myself at the Meertens Instituut to assess the possibilities of being contracted temporarily as an external developer for one of their upcoming linguistic projects. However, to my surprise, Sjef started our meeting with the words "*So... Marco, what do you already know about this PhD project?*", to which I open-mindedly replied "*Not much, apparently*". One hour later I left the Meertens Instituut with a revived desire to further develop my scientific research skills. The scholarly flame has been burning ever since. Thanks, Edwin!

My loving thanks are due to **my family** for always encouraging, supporting and believing in me. I realise that I owe you much of what I have become. Thank you, dear Mom, Dad, Karin, Kiki and all other relatives of our small but precious family!

Finally, and above all, my most cherished thanks are due to my wife **Jet Haasbroek** for her passionate and wholehearted love, encouragement, understanding, patience, guidance, support, and everything else. Thank you, Jet! You are the best thing that ever happened to me. I proudly dedicate this work to you, to honour the love we have found during these years.



*for Jet*



# 1. Introduction



Figure 1-1: “Toon wast \_\_\_”.

## 1.1. Motivation

In standard Dutch people are expected to complete the sentence depicted in Figure 1-1 as shown in example (1a).<sup>1</sup>

- (1) a. Toon wast *zich*.  
      ‘Toon washes REFL’
- b. Toon wast *hem*.  
      ‘Toon washes him’
- c. Toon wast *zijn eigen*.  
      ‘Toon washes his own’
- “Toon washes himself.”

The example describes a *washing* relation between the subject *Toon* and the object *himself*. Standard Dutch grammar prescribes that in this case the reflexive pronoun *zich* ‘REFL’ is to be used.<sup>2,3</sup> However, it is a well-known fact that

---

<sup>1</sup> The picture was presented to 259 Dutch dialect speakers with the instruction to complete the sentence in their local dialect. The geographical distributions of the attested variation in the depicted syntactic context are shown on map 68b in SAND1 which is discussed in Section 1.2.3.

<sup>2</sup> The Dutch reflexive pronoun *zich* cannot be literally translated to English. It is normally annotated with REFL or SIG in word-by-word translations.

<sup>3</sup> The introduction does not mention relevant linguistic properties (such as number, person and gender) in the context of this example for explanatory purposes.

Dutch shows variation in the choice of reflexive pronouns (Bennis and Barbiers, 2003). For example, dialect speakers along the coast line of the central-northern Frisian and south-western Flemish regions in the Dutch language area (see Figure 1-8 on page 25) prefer the personal pronoun *hem* ‘him’ instead. This form is similar to the English pronoun *him* and is shown in example (1b). Furthermore, in the centre of the Dutch language area the alternative form *zijn eigen* ‘his own’ frequently occurs. This form is listed in example (1c). The standard Dutch object pronoun *zich* ‘REFL’ appears most frequently near the eastern Dutch language border and, perhaps not surprisingly, highly resembles the German reflexive pronoun *sich*.

The language situation above illustrates one type of syntactic variation. This includes language variation with respect to word order, morphosyntax and doubling phenomena. Morphosyntactic variation investigates the patterns of word formation which depend on the syntactic context (such as inflection), whereas syntactic variation studies the ways in which linguistic elements (such as words and clitics) are put together to form constituents (such as phrases or clauses).<sup>4</sup> Examples (1a-c) show three different syntactic forms to express the same meaning as depicted in Figure 1-1. Although the prescriptive grammar of Dutch dictates that in standard Dutch the objective pronouns *hem* ‘him’ and *zijn eigen* ‘his own’) cannot refer to the subject *Toon* in the same clause, examples (1b-c) illustrate that this rule does not hold in dialects of Dutch. However, the different ways in which dialect speakers linguistically express the meaning of the picture in Figure 1-1 also form a coherent grammatical system. Only three types of constructions occur in this particular syntactic context.

- (2) a. ‘t Lijkt wel *of* er iemand in de tuin staat.  
 ‘it looks AFFIRM if there someone in the garden stands’
- b. ‘t Lijkt wel *dat* er iemand in de tuin staat.  
 ‘it looks AFFIRM that there someone in the garden stands’
- c. ‘t Lijkt wel *of dat* er iemand in de tuin staat.  
 ‘it looks AFFIRM if that there someone in the garden stands’
- d. ‘t Lijkt wel *of* er *staat* iemand in de tuin.  
 ‘it looks AFFIRM if there stands someone in the garden’

“It looks as if there is someone in the garden.”

---

<sup>4</sup> The definitions of morphosyntactic and syntactic variation are based on explanations in Merriam-Webster's Collegiate Dictionary and the Random House Unabridged Dictionary.

Another example of syntactic variation is observable when Dutch dialect speakers translate the sentence *'t Lijkt wel of (er) iemand in de tuin staat* ('it looks [affirmative] if (there) someone in the garden stands') into their local dialect. Examples (2a-d) show a selection of the attested variation in Dutch dialects with respect to the introduction of the subject clause (*er) iemand in de tuin staat* ('(there) someone in the garden stands') in this particular syntactic context. It turns out that nearly all dialects in the Netherlands share the standard Dutch realisation using the complementiser *of* 'if' to introduce the subject clause. Example (2a) shows the standard Dutch form. Exceptions are the Frisian area where the complementiser *dat* or *at* 'that' predominantly occurs, and the central southern (Brabant) region where people frequently combine the two complementisers into *of dat* 'if that'. The alternative forms are shown in examples (2b) and (2c), respectively. In Belgium the latter 'complementiser doubling' configuration is the most frequently occurring expression in this syntactic context, although the *dat* 'that' pattern also regularly appears. Finally, there are a few areas in the Frisian and Flemish provinces where the verb *staan* 'to stand' is in the second position in the subject clause. Example (2d) also illustrates another type of syntactic variation by showing that different word orders may express the same semantic content. Figure 1-9 shows a geographical map to visualise the attested variation in this particular syntactic context. Section 1.2.3 discusses syntactic variation in more depth and provides various other examples.

Several linguistically relevant observations may be extrapolated from the two language situations discussed above. First, various types of syntactic variation in Dutch dialects exist which often differ from the grammatical rules of the standard language. This observation indicates that dialectal variation research enriches the empirical domain of syntactic research. Also, analyses of dialectal variation patterns may result in more fine-grained linguistic theories. Empirical dialect data may also help improve the validation process of linguistic theories. Therefore, dialectal variation research may contribute to a better understanding of the inner workings of the human language system.

Second, there is a system behind the patterns of syntactic variation. Different variants do not occur randomly and geographical patterns of variation are quite easily distinguishable for an individual syntactic form. In other words, the geographical distribution of an individual syntactic phenomenon is often geographically coherent to a certain extent. This observation indicates that there might be a relationship between syntactic variation and geographical distance. This work assumes that investigations of language variation in geographical space not only illustrate patterns of variation at a certain point in time, but may also reflect residues of linguistic and cultural changes over time. Section 3.6.2 describes the case of the Frisian city dialect islands to illustrate how settlement history might still be reflected in geographical variation patterns in present-day dialects.

Third, the geographical distributions of the syntactic variation patterns in examples (1a-c) and (2a-d) do not overlap perfectly. Frisian and Flemish regions are discernable in both language situations. The reflexive pronoun *hem* ‘him’ in the context shown in example (1b) regularly occurs in both dialect areas. However, Frisian and Flemish dialect speakers use different syntactic expressions in the complementiser context shown in example (2). In Frisian dialects *(d)at* ‘that’ frequently occurs, whereas Flemish dialects often choose *of dat* ‘if that’. Examples (2b) and (2c) show the Frisian and Flemish realisations, respectively. Furthermore, the region near the eastern language border in which the reflexive pronoun shown in example (1a) appears—i.e. the area where most dialects share the *zich* ‘REFL’ pronoun—does not exist at all in the complementiser distribution. This observation demonstrates that interpretational problems may promptly arise when several distribution patterns of syntactic phenomena are combined for joint analysis at higher levels of abstraction to study more general characteristics of syntactic variation. Interpretability of the geographical distributions decreases as more variables are added for joint comparative analysis (of the type described above).

The current research presents several ways to solve this type of uninterpretability. It demonstrates various methods to objectively and verifiably analyse syntactic variation for any given degree of detail. The techniques are quantitative by nature, which means that the linguistic data are represented and compared numerically using a ‘linguistic ruler’. This is a computational instrument comparable to a geometrical ruler used to measure the distance between two points on a piece of paper in centimetre units. With such a ruler the linguistic distances between any pair of dialects can be measured in an objective and verifiable manner. Another type of computational ruler is introduced to measure the degrees of correspondence between any combination of syntactic variables. The syntactic measurements are also compared with measurement results based on pronunciation and lexical variation to put the syntactic variation patterns in a broader language variation context.

To summarise, this dissertation investigates how to adequately measure syntactic variation in Dutch dialects. It analyses Dutch syntax from a number of quantitative perspectives to study more general characteristics of syntactic variation. The motivation for this research is threefold. First, this work aims to contribute to a better understanding of syntactic variation in the Dutch language area. Second, this work aims to contribute to a better understanding of the relationship between syntactic variation and variation at other linguistic levels. Research into the associations among linguistic levels may help determine whether there might be structural, typological constraints linking variation at the linguistic levels. These two aspects might, ultimately, also provide new insights in the human language system in general. Third, and finally, this work aims to contribute to a better understanding of the relationship between syn-

tactic variation and geographical patterns of variation outside the realm of linguistics. As stated before, geographical patterns of syntactic variation may reflect residues of political, social and cultural changes over time. This work discusses a correlation between dialect borders and the political history of Friesland in Section 3.6.2. Furthermore, Section 6.3.2 uncovers a correlation between a syntactic dialect border and a social-cultural, Catholic-Protestant boundary. These two examples merely serve to provide a glimpse into the relatively uncharted expanse of potentially relevant interdisciplinary relationships.

## 1.2. Dialectological context

The research described in this dissertation is of a multidisciplinary nature. It most notably combines and extends scientific work from the research areas of dialectology, dialectometry, syntactic microvariation, data analysis and data mining. The current section provides a historically-oriented overview of the most closely related research areas to position this work in the scholarly field.

### 1.2.1. Dialect cartography

The research field of dialectology studies the linguistic properties of dialects—i.e. geographically bound (informal) language varieties. In other words, there is an inherent relation between language and geography. Geographical maps are often used to visualise the geographical occurrences of linguistic phenomena.

Jellinghaus (1892) is the first to provide a geographical dialect map of the Dutch language area. Until then, only regional maps had been published. The dialect map uses red boundary lines to divide the Dutch language area into a Frisian (central North), Saxon (north-eastern) and Franconian (western and southern) main region. The red and green boundaries are based on various word collections and on the series of dialect translations of the parable of *De Verloren Zoon* ('the lost son') published in Winkler (1874). However, the lines should not be interpreted as isoglosses—i.e. geographical boundary lines delimiting the area in which a given linguistic feature occurs. Although Jellinghaus describes a large number of linguistic properties for numerous dialects, the text does not specify a one-to-one correspondence between boundary lines and linguistic features. Nevertheless, it seems plausible that Jellinghaus' observations are in line with the boundary lines on the dialect map reprinted in Figure 1-2.

Te Winkel (1901) contains a more detailed dialect map of the Dutch dialect varieties (Figure 1-3). The map is based on two linguistic questionnaires sent out by the *Aardrijkskundig Genootschap* ('geographical society') in 1879 and 1895. The questionnaires brought about 284 answers for 212 dialects and 209 answers for 194 dialects, respectively. Te Winkel's dialect map contains various refinements and differences when compared to Jellinghaus' map. An example of a refinement can be found in the boundary line delimiting the central-eastern



Figure 1-2: Jellinghaus' (1892) map based on an interpretation of various word and parable translations.



Figure 1-3: Te Winkel's (1901) map based on an interpretation of two linguistic questionnaires.

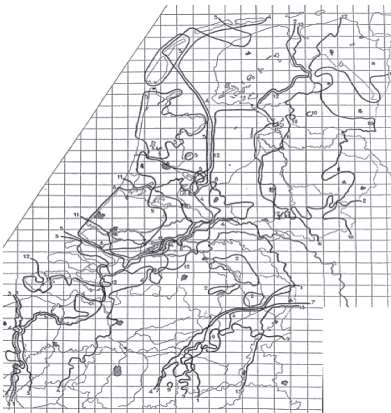


Figure 1-4: Weijnen's (1958) map based on 18 isophones and isomorphemes.

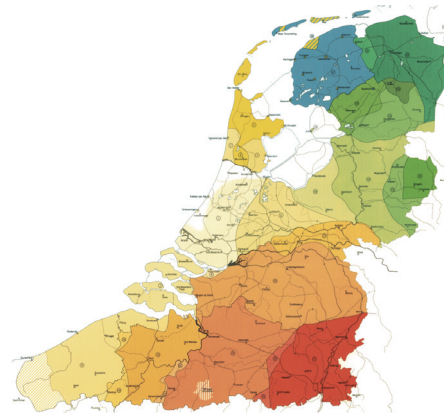


Figure 1-5: Daan and Blok's (1969) map based on subjective judgements.

Saxon region on Jellinghaus' map. Te Winkel's map subdivides this region in two separate Saxon areas in shades of blue. An example of a difference between the two maps can be found around the southern Saxon boundary line on Jellinghaus' map. The single boundary line on Jellinghaus' map is shown as two significantly-sized Saxon-Franconian transitional areas in grey-purple on Te Winkel's map. The dialect areas on Te Winkel's map seem primarily based on a number of isophones of /â/ and /î/ sounds. An isophone is a geographical boundary line which separates areas with identical sounds in certain words. However, the dialect map presents a methodological problem through its lack



of documentation with respect to the underlying classification process. The map uses colour distinctions to indicate the amount of difference between neighbouring dialects.

Van Ginneken (1913) published a map of the Dutch dialects which highly resembles Te Winkel's (1901) map. The map is not reprinted in the current work because of its similarity to Te Winkel's map. Van Ginneken's map essentially subdivides some regions differently and occasionally uses different dialect names. Unfortunately, Van Ginneken does not provide documentation of the underlying classification process either. Although it seems plausible that both authors base their classification of the Dutch dialect area on certain isophones and isoglosses, the results cannot be verified. Goeman (1989) observes that the second edition of Van Ginneken's map remarkably coincides with unpublished survey data in Willems (1886).<sup>5</sup> Weijnen (1966) categorises the maps discussed above as being based on the intuitive method precisely because of this unverifiability.

The intuitive methodology is in sharp contrast to the technique underlying the Dutch dialect map in Weijnen (1958) shown in Figure 1-4. This historically notable map classifies the Dutch dialect area based on 18 isophones and isomorphemes—i.e. geographical boundary lines delimiting areas with identical word forms. For example, isogloss one on Weijnen's map separates the Limburg district in the south-east from the other regions in the Dutch language area based on the existence of the opposition between *falling tones* and *level high tones* in Limburg dialects.<sup>6</sup> Isogloss one overlaps to a large extent with isogloss seven (the *Uerdingen Line*), which separates the Limburg region based on the German-like realisation of the first person, singular pronoun. In Limburg dialects the *ich* form predominantly occurs instead of the standard Dutch pronoun *ik* (Weijnen 1966:424).<sup>7</sup> A geographical map based on the isogloss method automatically shows the importance of individual isoglosses since overlapping isoglosses—i.e. isogloss bundles—result in thicker boundary lines on the map and, therefore, represent important area borders. A single isogloss constitutes a less important and less certain linguistic boundary. Isogloss maps are methodologically preferable over intuitive maps because the cartographic process is verifiable. A fundamental methodological problem with isogloss maps, however, lies in the arbi-

---

<sup>5</sup> The results of the 4000-items survey in Willems (1886), which contains 347 questionnaires from 337 different localities amounting to 19,060 answer pages, remained unpublished because Willems died in 1898 before he could complete his work. Van Ginneken received Willems' data and feature occurrence tables on loan in 1914 from the archives of the Royal Academy of Dutch Language and Culture in Ghent (Goeman 1989).

<sup>6</sup> Van Oostendorp (2006) notes that these distinctive tonal contours in Limburg dialects of Dutch are traditionally called *stoottoon* ('bumping tone') and *sleeptoon* ('dragging tone').

<sup>7</sup> Map 38a in the first volume of the Syntactic atlas of the Dutch dialects (introduced in Section 1.2.3) convincingly confirms the opposition between the *ik* and *ich* forms which differentiates the Limburg region from the other regions in the Dutch language area.

trariness of the included linguistic phenomena underlying the dialect area classification. Each selection of linguistic phenomena results in different isoglosses and produces different isogloss bundles. Kessler (1995) and Heeringa (2004) mention three additional problems regarding dialect area classifications based on isoglosses. First, isoglosses do not always coincide. This results in blurry isogloss bundles with parallel or crossed boundary lines. Second, many isoglosses do not straightforwardly bisect the language area. Two variants of a linguistic phenomenon often cannot be separated geographically by a single boundary line. Instead, the geographical distributions of the linguistic variants are intermingled to a certain degree. Third, it seems inappropriate to define dialect boundaries in a language area which is often described as a dialect continuum with very gradual changes (cf. Daan and Blok, 1969; Heeringa, 2004; Spruit, 2005).

The Daan and Blok (1969) map shown in Figure 1-5 offers an authoritative perceptual perspective on language variation in the Dutch language area. Perceptual classifications of dialect areas are based on the idea that subjectivity is required to adequately judge the relevance of isoglosses. The Daan and Blok map is based on the following two questions contained in a survey which was sent out by the Dialectencommissie ('dialect committee') in Amsterdam to about 1500 respondents in 1939:

- I. In which place(s) in your area does one speak the same or about the same dialect as you do?
- II. In which place(s) in your area does one speak a definitely different dialect than you do? Can you mention any specific differences?

Daan and Blok (1969) process the survey results using the arrow method which was introduced in Weijnen (1946).<sup>8</sup> Rensink (1955) and Weijnen (1966) previously applied the arrow method with respect to Dutch dialect regions. The method uses arrows to connect neighbouring dialects which local dialect speakers judge to be similar. The procedure results in arrow-bound clusters of localities which are separated by empty spaces that form perceptual dialect area boundaries based on the language awareness of the dialect speakers. Section 2.2 discusses a number of methodological and practical problems from which the perceptual dialect area classification in Daan and Blok (1969) suffers, such as the use of different methods and informant profiles for the Netherlandic and Belgian parts of the map. For example, in Belgium the arrow method was not

---

<sup>8</sup> The Daan and Blok (1969) map actually consists of multiple geographical maps. The main, central map is the result of work by Jo Daan. This is the dialect map under discussion. The additional maps surrounding Daan's perceptual map visualise the onomastic research by D.P. Blok. In this work the Daan (1969) map is consistently referred to as the Daan and Blok (1969) map to avoid citational confusion.

applicable and the informants were local dialect experts. Furthermore, the map colours were chosen rather intuitively and the map designers ‘corrected’ some survey data. To conclude, the map may certainly be considered a historical landmark in Dutch dialectology, but it also manifests the need for a uniform, verifiable and objective method to analyse and visualise the relation between language variation and geography more accurately.

### 1.2.2. Dialectometry

The research field of dialectometry—i.e. the measurement of dialect differences—studies differences between dialects from a quantitative perspective. This is in contrast with the research methodologies discussed in the previous subsection, which are all of a qualitative nature. Qualitative linguistic research focuses on a restricted number of linguistic phenomena simultaneously which are investigated in high detail using a small but focused data set. Quantitative linguistic research investigates many linguistic phenomena simultaneously in lesser detail using large data sets. The key step in the type of quantitative research described in this work is the step from measuring individual linguistic variables to aggregated differences between language varieties. This step requires that numerical values are assigned meaningfully to linguistic variables using a measure of linguistic distance. The latter is a method to measure the linguistic distance between two language varieties, analogous to a geometrical ruler used to measure the distance in centimetres between two geometrical points on a piece of paper. Once a suitable measure of linguistic distance has been defined, individual variables can be added up to arrive at more general descriptions of language varieties. Imagine what would happen if many linguistic differences were accumulated on one geographical map without using a numerical representation of some kind for the linguistic variables: the geographical map would become an uninterpretable set of overlapping bundles of isoglosses. Therefore, a quantitative research perspective can augment more traditional, qualitative linguistic research because the linguistic data is examined from different, more general perspectives.

Séguy (1973) introduces dialectometrical methods to measure dialect distances in a successful attempt to analyse the geographical maps in the six-volume series of the *Linguistic and ethnographic atlas of Gascony* (ALG; Séguy, 1954-1973) more objectively than was possible with traditional methods. The method divides the number of linguistic items in which each pair of dialects differs by the total number of linguistic items. The numeric result is expressed as a percentage and is interpreted as the linguistic distance between any pair of dialects. The dialectometrical data contains 170 lexical, 67 pronunciational, 75 phonetic or phonological, 45 morphological, and 68 syntactic variables. Each of the five linguistic levels under investigation is weighted equally by calculating percentages for each linguistic level rather than for each linguistic item. Therefore, the

final linguistic distance is calculated as the mean of the five percentages. The linguistic distances are plotted on geographical maps after grouping the linguistic distance percentages into several distance percentage classes and by representing them with different line types (Chambers and Trudgill, 1998:138-140; Heeringa, 2004:14).

Goebel (1982) marks the beginning of a series of major contributions to dialectometrical research. Goebel designs a number of dialectometrical methods and visualisations using a selection of 696 geographical maps regarding 251 dialect varieties in the *Speech atlas of Italian and southern Swiss* (AIS; Jaberg and Jud, 1928-1940). The data set contains 569 lexical variation maps and 127 morphosyntactic variation maps. Goebel's methodologies resemble Séguy's techniques considerably, although their measurement strategies differ with respect to the research focus. Whereas Séguy calculates dialect distances, Goebel determines dialect similarities. Nevertheless, the measurement results are comparable, because dialect similarity values may be converted to dialect distances by subtracting the similarity percentages from one hundred. For example, a relative similarity of 80 percent between two dialects translates into a difference between the dialects of  $(100 - 80 =) 20$  percent. Goebel's methodological contributions to the field of dialectometry include the introduction of standard cluster analysis procedures to help interpret the data and the development of numerous dialectometrical visualisation methods, among many other improvements. Since then, Goebel's work and its empirical foundation has expanded significantly and currently includes dialectometrical investigations of the *Linguistic atlas of France* (ALF; Gilliéron and Edmont, 1902-1920) and the *Linguistic atlas of Dolomitic Ladinian and neighbouring dialects I and II* (ALD I/II; Goebel and Böhmer, 1985-2011). Goebel (2006) extensively describes the current state of Goebel's dialectometrical work.

Hoppenbrouwers and Hoppenbrouwers (1988; 2001) introduce several methods to measure linguistic distances between Dutch dialects based on the *Series of Dutch Dialect atlases* (RND; Blancquaert and Peé, 1925-1982; see Section 4.3), most notably the feature frequency method. This procedure counts the number of occurrences of 21 phonological features in the transcriptions of the same set of 139 sentences for each of the 156 dialect varieties under investigation. The Hoppenbrouwers brothers customised the set of phonological features in the *Sound Pattern of English* (SPE) by Chomsky and Halle (1968) for optimal use with the Dutch dialectal data in the RND. For example, the feature *front* indicates that a vowel is pronounced in the front of the oral cavity and not in the middle or in the back. Similarly, the feature *low* indicates that a vowel is pronounced with the tongue low and not central or high. The feature frequency method can be characterised as a corpus-based approach. Dialect distances are determined by comparing histograms of feature frequencies which are expressed in percentages. A major disadvantage of the method is that it does not incorporate

the order in which speech sounds occur. Furthermore, words are not recognised as meaningful language units. The method ignores word order, which implies that the method cannot be used to quantify syntactic variation. The feature frequency method merely focuses on variation with respect to phonetic and phonological usage patterns in the RND sentence transcriptions (Heeringa 2004).

Kessler (1995) introduces the Levenshtein distance in language variation research to measure the linguistic distances between Irish Gaelic dialects using data from the first volume of the *Linguistic Atlas and Survey of Irish Dialects* (LASID; Wagner, 1958-1969) consisting of 51 words in 95 dialect varieties. Sankoff and Kruskal (1999) discuss a broad range of applications of this generic string-edit distance algorithm. Section 4.4 provides an overview of the Levenshtein distance measure. Nerbonne et al. (1996) and Nerbonne and Heeringa (1998) describe the first applications of the Levenshtein algorithm to classify the Dutch dialect areas based on a representative selection of 100 word transcriptions in the RND. The former is a pilot study based on a relatively small set of 20 Dutch dialects, whereas the latter already takes into account pronunciation variation in 104 Dutch dialects. Heeringa (2004) most notably extends and refines this line of research. Accomplishments of the dialectometrical work by Nerbonne and Heeringa include improvements over the original Levenshtein algorithm, investigations of various statistical analysis techniques as well as experimentation with alternative visualisation methods to more accurately interpret the results. Also, the RND data selection was further expanded to include 125 word pronunciations in 360 Dutch dialects. Apart from dialectometrical investigations of pronunciation and lexical variation in Dutch dialects, also German (Nerbonne and Siedle, 2005), American English (Nerbonne and Kleiweg, 2007), Sardinian (Bolognesi and Heeringa, 2002) and Norwegian (Heeringa and Gooskens, 2003) dialects have been examined, among others.<sup>9</sup> To conclude, the Levenshtein distance measure is a powerful tool to quantify linguistic variation because it is a numerical measure—it allows differentiation between linguistic item pairs in terms of degrees of similarity, which means the algorithm can take into account levels of affinity between two linguistic items that are not equal but are nevertheless related to a quantifiable extent. This is in contrast to the nominal distance measures developed by Séguy, Goebel, the Hoppenbrouwers brothers, and others.<sup>10</sup> Unfortunately, the Levenshtein distance algorithm also has a fundamental shortcoming as a tool to accurately measure linguistic distances. Heeringa (2004:25) notes that:

---

<sup>9</sup> The data sets used in these investigations consisted of 201 words in 186 German dialects, 151 words in 483 American English dialects, 200 words in 60 Sardinian dialects, and 58 words in 15 Norwegian dialects, respectively.

<sup>10</sup> Although Goebel's GIW method (see Section 4.4) employs item frequency to incorporate gradual differences between linguistic items, the method remains nominal at a fundamental level since a comparison between two items returns either equal or unequal.

“[...] lexical, phonological and morphological differences need not be explicitly distinguished, but can be processed with the same algorithm. However, since the algorithm compares word pronunciations, syntactic differences are not processed”.

Taking into account that the Levenshtein algorithm as a tool to measure pronunciational variation in Dutch dialects in the RND will be discussed in detail in Chapter 4, only one topic remains to be discussed before the main topic of the current work can be presented meaningfully: syntactic variation in Dutch dialects.

### 1.2.3. Syntactic microvariation

The research area of syntactic microvariation—i.e. dialectal variation in the realm of syntax, also known as dialect syntax—has until recently been a vastly ignored field in linguistics. This type of research conceptually combines and extends two active specialisations in language variation research: comparative syntax and dialectology. Comparative syntactic research investigates differences between languages with respect to their syntactic properties such as word order and morphosyntactic variation. It is sometimes also referred to as syntactic macrovariation. This specialisation within the field of syntactic variation research has mainly focused on explaining the differences between standard languages—such as Dutch and English—in terms of the setting of abstract linguistic parameters within the leading linguistic paradigms of generative grammar and language typology. Section 5.1 further introduces these linguistic frameworks. As Section 1.2.1 already points out, dialectological research specialises in documenting and analysing dialectal variation, but, over the last century, examinations of the collected data have mostly been limited to the linguistic domains of the lexicon and pronunciation, and to a lesser extent, phonology and morphology.

In the recent past, however, dialect syntax has become a much more prominent topic in linguistics and syntactic properties of dialects are now being studied in a more systematic way. Barbiers and Cornips (2001:2) state that:

“[...] the study of syntactic microvariation has various goals. The goal of traditional dialect syntax is to explore the geographical distribution of syntactic variables. The geographically determined syntactic variation thus established can be used for other types of research, such as the investigation of language change and external language history. Recently, the aim of syntactic microvariation research has been extended to studying the universal properties of the human language, since it contributes to our understanding of the [1] patterns, [2] loci and [3] limits of syntactic variation within that system”.

An explanation of the three issues formulated above can help illustrate the type of research in dialect syntax. The arguments with respect to the relevance of

studying the patterns, loci and limits of syntactic variation should be interpreted as follows. First, regarding the contribution to our understanding of language patterns, a major recent contribution of dialect syntax research to syntactic theory in general is its observation that dialects exhibit various syntactic phenomena that are generally not part of the standard language. A relevant example is the typical occurrence of doubling phenomena (cf. Barbiers, Koenenman and Lekakou, t.a. 2007). Generally speaking, dialects may be considered ‘more natural’ language systems due to the limited influence of prescriptive standard norms in comparison with standard languages.

The second argument states the contribution of dialect syntax research to our understanding of language loci—i.e. the originating centres of language variability. It argues that the source of syntactic variation may be better understood when the language system is examined in more detail from a complementary research perspective using minimally different language systems. The areal distributions of syntactic variables often reflect the spread of innovations. Bucheli and Glaser (2002) argue that a theory of language change including grammatical change should take into account that systems of neighbouring dialects may provide data concerning the direction and the stages of a certain development. In this context it should be noted that the widely known Universal Grammar hypothesis in its strongest form (cf. Chomsky, 1995) claims that syntactic variation does not exist at all. It postulates that grammar principles exist which are shared by all languages and which are innate to humans. Under this view, syntactic variation should be reducible to parameterisation of morphosyntactic features, and to different ways to realise one and the same syntactic structure phonologically (cf. Barbiers, Cornips and Kunst, 2007).

Finally, the third argument states that dialect syntax research may help determine which syntactic properties are universal by examining the limits of syntactic microvariation patterns. Although dialect varieties typically allow many more variants in language situations than standard languages, certain logically conceivable variants never occur. Barbiers (2005) examines the apparent impossibility of the 2-1-3 word order in verbal clusters such as *\*Ik vind dat iedereen kunnen<sub>2</sub> moet<sub>1</sub> zwemmen<sub>3</sub>* (\*‘I think that everyone can<sub>2</sub> must<sub>1</sub> swim<sub>3</sub>’), and Barbiers and Bennis (2003) investigate why certain logically conceivable strong reflexives in Dutch such as *hem-eigen* ‘him-own’ in *\*Jan herinnert hem-eigen dat verhaal wel* (\*‘John remembers him-own that story [affirmative]’) have never been attested. Such limits on syntactic variation demonstrate that the research field of dialect syntax can contribute to the uncovering of possible versus impossible properties of natural language, thus enhancing the empirical basis and the theoretical foundation of syntactic theory and language research in general.

However, this type of dialect syntax research would not have been possible without the recent completion of several large-scale, syntactic microvariation data collection projects. Within Europe alone, the list of recent, successful dia-

lect syntax projects includes the *Syntactic Atlas of Northern Italy* (ASIS; Poletto et al., 1992-2002), the *Freiburg Corpus of English Dialects* (FRED; Kortman et al., 2000-2005), the *Dialect Syntax of Swiss German* (SADS; Glaser et al., 2000-2002), the *Syntactically Annotated Corpus of Portuguese Dialects* (Cordial-SIN; Martins et al., 1999-2003), the *Scandinavian Dialect Syntax* (ScanDiaSyn; Vangsnes et al., 2005-2007) pilot project, and most notably, the *Syntactic Atlas of the Dutch Dialects* (SAND; Barbiers et al., 2000-2008).<sup>11</sup>

The current research represents the first large-scale, quantitative investigation of purely syntactic variation phenomena in the Dutch language area. The data source underlying this work has been entirely drawn from the *Syntactische atlas van de Nederlandse dialecten*, henceforth the SAND. The first volume (SAND1; Barbiers et al., 2005) of this unique syntactic variation database contains 145 geographical distribution maps of individual syntactic variables in 267 Dutch dialects in the Netherlands, the Northern part of Belgium and a small north-western part of France. Figure 1-6, Figure 1-7 and Figure 1-8 show the 267 dialect locations and the relevant province names. SAND1 covers syntactic variation related to the left periphery of the clause and pronominal reference. It includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. Table 1-1 provides informal examples of syntactic variables in each of these syntactic domains. Table 5-1 to Table 5-4 list a number of variable examples in more detail. The work described in the dissertation is almost entirely based on SAND1 data (and not on SAND2). The second and final volume of the SAND will appear in 2008. Section 6.3.2 concludes this dissertation with a preliminary review of the SAND2 data including several examples of SAND2 variables in context.

From a quantitative research perspective SAND1 also represents a syntactic microvariation database containing 106 syntactic contexts and 485 syntactic variables among varieties of a single language. This work defines a syntactic variable as a form or word order in a syntactic context in which two dialects can differ (Spruit, 2006).<sup>12</sup> Figure 1-9 is a near copy of SAND1 map B on page 14 to illustrate what is meant by a syntactic context and syntactic variables. It shows the geographical distribution of the attested syntactic variables in the syntactic context of a *complementiser of a comparative if-clause*. Simplified, this map interprets the different realisations of the complementiser position in comparative

---

<sup>11</sup> The *European Dialect Syntax* Project (Edisyn; Barbiers et al., 2005-2010) and the *Scandinavian Dialect Syntax* (ScanDiaSyn; Vangsnes et al., 2005-2010) project umbrella are notable examples of large-scale dialect syntax projects currently in progress.

<sup>12</sup> Note that this interpretation does not analyse the data in Figure 1-9 as a single categorical variable with seven values, but rather as seven binary variables, which may be considered a theory-neutral approach. A variable either occurs or it does not occur in a dialect. However, some linguistic structure may be lost this way (mutual exclusiveness, multiple responses).





Figure 1-6: The 267 dialect locations in the Dutch language area under investigation.



Figure 1-7: Distribution of the 267 Dutch dialects in the syntactic atlas.



Figure 1-8: The provinces in the Dutch language area under investigation.

Table 1-1: Examples of syntactic variables in context for each syntactic domain/ chapter in SAND1. Please refer to Table 5-1 to Table 5-4 for more detailed variable examples.

Chapter 1: Complementisers (map 14b):

't lijkt wel dat er iemand in de tuin staat.  
 'it looks [affirmative] that there someone in the garden stands'

Chapter 2: Subject pronouns (map 38b):

Ze gelooft dat du eerder thuis bent dan ik.  
 'she believes that you earlier home are than me'

Chapter 3: Subject doubling and subject cliticisation following yes/no (map 54a):

As- ge gij gezond leeft, leef- de gij langer.  
 'if you<sub>weak</sub> you<sub>strong</sub> healthily live, live- you<sub>weak</sub> you<sub>strong</sub> longer'

Chapter 4: Reflexive and reciprocal pronouns (map 68a):

Jan herinnert zijn eigen dat verhaal wel.  
 'John remembers himself that story [affirmative]'

Chapter 5: Fronting (map 93b):

Die rare jongen ben ik mee naar de markt geweest.  
 'that strange guy am I with to the market been'

if-clauses—such as *of*, *dat*, *of dat*, et cetera—as syntactic variables in the syntactic context 't lijkt wel \_\_\_ er iemand in de tuin staat ('it looks [affirmative] \_\_\_ there someone in the garden stands'). In standard Dutch people say 't lijkt wel *of* er iemand in de tuin staat ('it looks [affirmative] if there someone in the garden stands'), but in colloquial Dutch the following form also frequently occurs in the southern provinces: 't lijkt wel *of dat* er iemand in de tuin staat ('it looks [affirmative] if that there someone in the garden stands'). The standard language realisation *of* in this context occurs in 155 dialects and is visualised with medium brown square symbols on the map, whereas the mostly southern realisation *of dat* was recorded in 66 dialects and is shown using light brown square symbols. There are even a few northern and southern regions within the Dutch language area where the verb occurs in the second position of the if-clause: 't lijkt wel *of* er *staat* iemand in de tuin ('it looks [affirmative] if there stands someone in the garden'). The latter example also illustrates that both word form and word order may vary within a syntactic context. Finally, the 'block of four coloured squares' below the syntactic variables in Figure 1-9 indicates which syntactic variables may occur simultaneously. The colour configuration on this map helps show that only in the dialect of Zoutleeuw—in the province of Flemish Brabant near the Flemish Limburg border—do people have three different ways to express a complementiser in a comparative if-clause. The map

### 1.3.1.2 Voegwoord van vergelijkende of-zin Complementiser of comparative if-clause

't lijkt wel of er iemand in de tuin staat.  
it looks AFF if there someone in the garden stands

'It looks as if there is someone in the garden.'

■ of	155
■ of dat	66
■ dat	25
■ as/of + V2	11
■ at	5
■ as	4
■ et	3

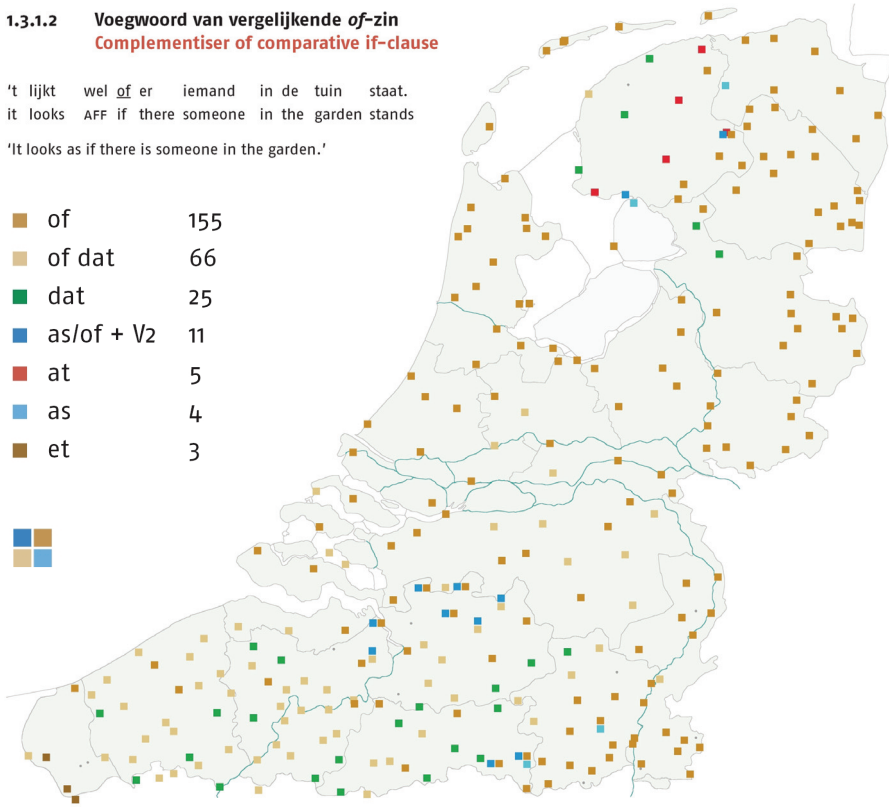


Figure 1-9: SAND1 map 14b shows seven syntactic variables in the context of a complementiser of a comparative if-clause.

distinctively marks the geographical location of Zoutleeuw with a block of dark blue, dark brown and light blue squares.

## 1.3. Research dimensions

The research in this dissertation positions itself in the scientific field of language variation research along the following four main dimensions:

- I. Quantitative instead of qualitative methodology.
- II. Syntactic instead of phonetic, phonological, morphological, pronunciational, lexical, prosodic or semantic variation.
- III. Micro instead of macro level.
- IV. Space instead of time dimension.

First, this dissertation studies language variation from a *quantitative* research perspective. This is in contrast to more traditional qualitative dialect research. The crucial step in quantitative linguistic research is to measure aggregated differences between language varieties instead of merely recording the differences between individual linguistic variables. This procedure requires numerical values to be assigned to linguistic variables which, then, can be added up using a measure of linguistic distance to arrive at more general descriptions of language varieties. Therefore, quantitative research perspectives are able to augment qualitative research because the linguistic data can be examined from different perspectives.

Second, this dissertation mainly inspects language variation at the *syntactic* level. This is in contrast with studies which investigate language variation at the lexical level to examine the vocabulary of language varieties, or focus on language variation at the pronunciation level to analyse the range of sounds occurring within a language, among others. Syntactic variation research focuses on differences among language varieties with respect to word order, morphosyntactic variation and doubling phenomena, among other aspects. As a rule of thumb, syntactic variants of a syntactic phenomenon express (nearly) the same semantic content.

Third, language variation can be studied from different levels of detail. Historically, most attention has been given to the examination of language differences at the macro level. This type of research focuses on differences between standard language varieties such as Dutch and English. However, this dissertation investigates language variation at the *micro* level which includes non-standard varieties. This work examines Dutch dialect varieties in the Netherlands, Belgium and France. Barbiers and Cornips (2001:2) formulate the relevance of syntactic variation research at the micro level as follows:

“This does not only enhance the empirical basis of syntactic theory, but it also reduces the influence of prescriptive rules and makes it possible to test potential correlations between syntactic variables while keeping other, possibly interfering factors constant”.

Fourth, and finally, this work investigates language variation in *space* instead of time. Therefore, it concentrates on linguistic differences from a synchronic perspective instead of a diachronic point of view. Language varieties are compared using data samples which are collected in or around the same time period. The differences and similarities between the varieties are analysed based on their geographical locations instead of their time of recording, which was basically stable in the early years of the 21<sup>st</sup> century.

## 1.4. Research questions

This dissertation investigates the following four research questions:

- I. How can syntactic variation be measured adequately? (*Model*)
- II. What are the syntactic distances among the Dutch dialects? (*Application*)
- III. To what extent are the linguistic levels of syntax, lexis and pronunciation associated with each other? (*Context*)
- IV. What are relevant dependencies between syntactic variables? (*Associations*)

Research questions I and II jointly address the relation between syntactic and geographical distance. The first question focuses on how to *model* syntactic differences between language varieties so that syntactic variation can be examined reliably in the aggregate to provide more general perspectives on syntactic variation. The second research question concentrates on the *application* of the measurement model to the first compendium of purely syntactic Dutch dialect data and analyses the results. These two research questions are answered in Chapters 2 and 3: “Dutch dialect area classifications based on aggregate syntactic differences” and “Measures of syntactic distance and the role of geography”, respectively.

Research question III addresses the degree to which geographical distributions of syntactic distances correlate with distributions of pronunciational and lexical distances. The question helps to put the syntactic measurement results into a broader linguistic *context* by calculating the extent to which syntactic variation correlates with pronunciational and lexical variation. This research question is the topic of Chapter 4: “Associations among linguistic levels”.

Research question IV addresses the discovery of relevant *associations* between syntactic variables. It contributes to the global linguistic research effort of parameterisation of the structural diversity of language varieties by identifying which syntactic variables nearly always co-occur. This research question is investigated in Chapter 5: “Discovery of association rules between syntactic variables”.

## 1.5. Chapter overview

This dissertation is centred around four chronologically ordered, peer-reviewed publications. Chapters 2, 3, and 5 have been published in *Linguistics in the Netherlands* (Spruit, 2005), *Literary and Linguistic Computing* (Spruit, 2006) and *Computational Linguistics in the Netherlands* (Spruit, 2007), respectively. Chapter 4 has been accepted for publication in *Lingua* (Spruit, Heeringa and Nerbonne, t.a. 2008). However, one potentially confusing remnant of this approach remains notice-

able in the terminology used in Chapter 2, in which syntactic variables in a syntactic context are referred to as feature variants of a syntactic feature. Chapter 3 documents this change in terminology. The remainder of this section introduces the research topics which are investigated in the following chapters.

Chapter 2 introduces the dialect classification problem and discusses the traditional dialect map based on subjective judgements. After introducing the research areas of dialectometry and syntactic variation, the syntactic measurement method and the analysis technique are described and the resulting Dutch dialect maps based on a syntactic measure—including geographical distribution maps for each syntactic subdomain—are discussed. The chapter concludes with a comparison of the computational dialect map based on syntactic variation with the perceptual dialect map based on subjective judgements.

Chapter 3 briefly recapitulates the work described in the previous chapter and extends it in several ways. The chapter refines the review of the Dutch syntactic variation database under investigation and revisits the syntactic measurement procedure and the analysis technique. Then, the resulting geographical colour map of the Dutch dialect area based on syntactic differences is related to dialect maps based on subjective judgements and pronunciation differences. An analysis of the correlation between syntactic and geographical distances follows. The chapter concludes with a presentation of an alternative measure of syntactic distance based on feature variables to incorporate linguistic information and compares its measurement results with the results based on atomic variables.

Chapter 4 contributes to linguistic research through a joint analysis of aggregate pronunciation, lexical and syntactic differences and in its attention to potential, mutually structuring elements among the linguistic levels. The chapter describes the two data sources under investigation and explains the two measurement procedures used to quantify linguistic differences. Colour maps of the Dutch dialect areas based on pronunciation, lexical and syntactic differences are shown to visually indicate the degrees of association. The distance measurements are also analysed with respect to consistency to ensure that the results are reliable before the exact degrees of association between pronunciation, lexis and syntax are presented. Then, the chapter lists the degrees of association between geography and the linguistic levels under investigation. The chapter concludes with refined calculations of the associations among the linguistic levels by accounting for the influence of geography as an underlying, third factor.

Chapter 5 introduces a data mining technique in linguistic research to discover associations between syntactic variables in Dutch dialects. A sample data subset is introduced to illustrate the association rule mining procedure based on proportional overlap. The chapter reviews the evaluation factors used to accurately measure the quality of the association rules and explores the most interesting rules discovered in the sample data. The chapter concludes with an exploratory

review of the data mining technique to the entire syntactic variation database, which highlights several highly ranked variable associations and discusses various directions for future research.

Chapter 6 summarises the previous chapters and provides its main conclusions in a question-answer format. The chapter ends with a general discussion of the overall results and several points of interest for future research.





## 2. Dutch dialect area classifications based on aggregate syntactic differences

“Revisiting the perceptual Daan and Blok dialect map”\*

*Spruit, M.R., 2005. Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited. In: Doetjes, J., Weijer, J. van de (eds), Linguistics in the Netherlands, 2005, John Benjamins, Amsterdam, 179-190.*

In this dialectometrical research a quantitative measure of syntactic distance is developed and applied to Dutch dialects. It will be shown that a quantitative perspective on syntactic variation provides new insights in the degree of geographical coherence in syntactic variation, using the perceptual Daan and Blok map of the Dutch dialects from a comparative perspective.

### 2.1. Introducing the dialect classification problem

Dialect speakers are aware of the existence of borders in the dialect landscape. The Daan and Blok (1969) map shown in Figure 2-1 classifies the Dutch dialects using subjective judgements from local dialect speakers to reflect this fact. However, dialects also seem to be organised in a continuum with gradual transitions which are sometimes larger and sometimes smaller. Although the existence of dialect borders does not necessarily exclude the presence of dialect continua, a measure of dialect differences is required to objectively differentiate them (Heeringa, 2004). This article describes a computational method to objectively classify the Dutch dialects using a syntactic measure.

First, the Daan and Blok dialect map based on subjective judgements is discussed in Section 2. Then, after introducing the research area in Section 3, the measurement method and the analysis technique are described in Sections 4 and 5. The resulting Dutch dialect maps are discussed in Section 6. Section 7 concludes with a comparison of the computational dialect map based on syntactic variation with the perceptual dialect map based on subjective judgements.

---

\* This research is being carried out in the context of the NWO project *The Determinants of Dialectal Variation*, number 360-70-120, P.I. J. Nerbonne. Please visit <http://dialectometry.net> for more information and relevant software. I would like to thank Hans Bennis, Sjeff Barbiers, John Nerbonne and an anonymous reviewer for their helpful comments on an earlier version of this paper.



Figure 2-1: *The Daan and Blok dialect map (reprinted from Daan and Blok, 1969)*

## 2.2. Classifying Dutch dialects using subjective judgements

The Daan and Blok dialect map uses subjective judgements from about 1500 local dialect speakers in the Netherlands, collected in 1939, to establish a classification of dialect areas in the Dutch language area. Dialect borders in the Netherlandic part of this map are found using the arrow method. In this method neighbouring dialects which speakers judge to be similar are connected by arrows. This results in clusters of localities bound by arrows and separated by empty spaces that form perceptual dialect area boundaries.

The arrow method could not be applied in Flanders because the Belgian dialectologists did not have a sufficiently large group of correspondents at their disposal. Therefore, Belgian language geographers, who often belonged to dialect-speaking groups themselves, were consulted. Also, some of the results were corrected afterwards in case of a very low response of correspondents for an area or contradictory responses, leading to consulting expert opinion rather than subjective judgements (Heeringa, 2004).

Furthermore, the colours used in the Daan and Blok dialect map were chosen more or less intuitively, although corresponding to a gradually increasing divergence from Standard Dutch. “This rank order does not follow from the judgements themselves, but was imposed by Daan on the speakers’ classification on the basis of expert knowledge of internal linguistic dialect structure” (Goeman, 2000:139).

To summarise, the classification of the Dutch dialects in the Daan and Blok map is the result of subjective judgements from local speakers, local experts and the map designers. Also, there is no differentiation within dialect areas, which contradicts the intuition that dialects are also organised in a continuum without sharp boundaries. The remainder of this article provides a computational method to objectively classify the Dutch dialects using a syntactic measure.

### **2.3. Combining dialectometry and syntactic variation**

This research combines and extends work from two different research areas: dialectometry and syntactic variation. “Dialectometry is the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography” (Nerbonne and Kretzschmar, 2003:245). The key step in dialectometry is from the measurement of individual linguistic variables to the measurement of aggregate differences of varieties. Dialectometrical methods were first described in Séguy (1971) and further investigated in Goebel (1982) and Heeringa (2004), among others. However, until recently no extensive collection of syntactic data was available, limiting dialectometrical research mainly to lexical and phonological data.

With the arrival of the first part of the *Syntactic Atlas of the Dutch Dialects* (SAND1, Barbiere et al., 2005), the first compendium of Dutch syntactic variation has become available. It is also one of the earliest syntactic atlases anywhere. SAND1 contains 145 maps showing the geographical distribution of syntactic phenomena in 267 Dutch dialects with respect to the following domains related to the left periphery of the clause and/or pronominal reference: complementisers, subject pronouns, expletives, subject doubling, subject cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting.

The SAND data were collected using various elicitation techniques (Cornips and Jongenburger, 2001), including the use of questions such as “Does this sentence occur in your dialect?” and “How common is this sentence in your dialect?”. Therefore, multiple variants may occur for an elicited syntactic feature at a given dialect location. To illustrate the syntactic variation data and the feature/variant terminology used throughout this article, an example of an elicited syntactic feature and its recorded feature variants is given in Table 2-1.<sup>1</sup>

To summarise, the feature-oriented SAND project has provided a database of observed variants per syntactic feature per geographical location. For this location-oriented dialectometrical research, these lists of locations per feature have

---

<sup>1</sup> 135 out of 145 maps in SAND1 contain unique geographical distributions of syntactic phenomena. Each of these 135 maps represents one syntactic feature and each map symbol represents one feature variant in the context of this work.

been transformed into lists of occurring feature variants per location. Using this representation the number of variant differences between pairs of locations can be measured.

*Table 2-1: Example of a syntactic feature and its recorded variants. Map 68a in SAND1 shows the geographical distribution of the syntactic feature weak reflexive pronoun as object of an inherent reflexive verb. Five feature variants have been recorded for this phenomenon throughout the Dutch language area: *zich*, *hem*, *zijn eigen*, *zichzelf*, *hemzelf*.*

Feature:	Weak reflexive pronoun as object of inherent reflexive verb					
Variants:	{ <i>zich</i> , <i>hem</i> , <i>zijn eigen</i> , <i>zichzelf</i> , <i>hemzelf</i> }					
Example:	Jan	herinnert	<i>zich</i>	dat	verhaal	wel.
	John	remembers	himself	that	story	[affirmative]'
	"John certainly remembers that story."					

## 2.4. Measuring syntactic variation using Hamming distance

The Hamming distance is calculated between each pair of dialect locations to obtain a measurement based on binary comparisons between feature variants. In this straightforward procedure the distance between dialect A and dialect B is increased by 1 for each variant that is observed in dialect A but not in dialect B, and vice versa. An outline of the Hamming distance algorithm is shown in Table 2-2.

*Table 2-2: Hamming distance algorithm applied to measure syntactic variation in dialects.*

for each pair of dialects A and B;	(level 1)
for each variant of all syntactic features;	(level 2)
if it does occur in dialect A, but does not occur in dialect B	(level 3)
or if it does not occur in dialect A, but does occur in dialect B;	
increment the distance between dialect A and B by 1.	(level 4)

Calculating the Hamming distances between all dialect pairs results in a table of differences. In this distance matrix each distance value represents the total number of different feature variant realisations between one pair of dialects. Note that a distance matrix is always symmetric because the distance from dialect A to dialect B is always identical to the distance from dialect B to dialect A. A small fragment of the SAND1 distance matrix is shown in Table 2-3.

To illustrate the measurement procedure described in Table 2-2, consider the dialects Lunteren and Veldhoven from Table 2-3 and the feature *weak reflexive pronoun as object of inherent reflexive verb* with associated variants as listed in Table 2-1. The variants *zich* and *zijn eigen* were recorded in Lunteren and the variant

*zich* was registered in Veldhoven. During the calculation of the Hamming distance between this pair of dialects (level 1), the number of differences for the feature *weak reflexive pronoun as object of inherent reflexive verb* needs to be determined (level 2). The variant *zich* is available in both dialects, therefore the dialect distance is not increased. Also, since the variants *hem*, *zichzelf* and *hemzelf* in the context of this feature do not occur in either of these two dialects, they have no effect on the distance value either. The variant *zijn eigen*, however, occurs in Lunteren but not in Veldhoven (level 3). Therefore, the dialect distance between Lunteren and Veldhoven is incremented by 1 (level 4). Thus, after this series of comparisons 5 out of 510 feature variants have been measured in order to determine the Hamming distance between this pair of dialects. This procedure is executed for all  $(267 * 266) / 2 = 35511$  dialect pairs and results in the distance matrix a part of which is shown in Table 2-3.

Table 2-3: Fragment of the SAND1 Hamming distance matrix. Each dialect pair distance is an integer between 0 and 510 which represents the total number of different feature variant realisations.

	Lunteren	Bellingwolde	Hollum	Doel	Sint-Truiden	Veldhoven	Houthalen
Lunteren		69	54	122	79	49	75
Bellingwolde	69		57	137	82	52	70
Hollum	54	57		118	63	59	75
Doel	122	137	118		117	113	123
Sint-Truiden	79	82	63	117		72	74
Veldhoven	49	52	59	113	72		58
Houthalen	75	70	75	123	74	58	

Finally, note that this measuring method does not yet take syntactic information into account. For example, the measurement could assign a distance value smaller than 1 when the reflexive feature variants *zich* and *zichzelf* are compared and a distance value greater than 1 when the distance between the variants *zich* and *zijn eigen* is determined. An even greater distance value might be assigned when one of the two dialects under comparison is lacking reflexive feature variants altogether. In its current form the distance value is incremented by 1 for all differing variant pairs. This is a generally applicable method that measures the number of differences between two sets of syntactic variants. Therefore, it is also useful as a reference measure for more advanced measurements that do

take into account syntactic properties. In addition, the measurement could be refined by taking into account statistical information such as the number of variant occurrences and the number of alternative variants per feature.

## **2.5. Analysing dialect distances using multidimensional scaling**

Multidimensional scaling (MDS) is applied to analyse the dialect relationships in the distance matrix. The goal of this procedure in this context is to optimally represent the most differentiating feature variants for each dialect in relation to all other dialects. The results are visualised with dialect colour maps.

First described in Torgerson (1952), MDS is a statistical technique for producing a lower-dimensional data set suitable for visualisation from a high-dimensional data set, while preserving the distance relationships of the high-dimensional data set as faithfully as possible. Applied to the visualisation of the syntactic distance matrix in Table 2-3, the set of 510 variant dimensions for each dialect is first scaled down to a coordinate in a three-dimensional space which represents an optimal interpolation of the most differentiating dialect variants. The coordinates do not directly correspond to actual variant values.

Then, the three-dimensional coordinates are used as values between light and dark of the three colour components red, green and blue to give each dialect location a unique composite colour. Neighbouring dialect locations will have corresponding colours if there is a correlation between geographical distance and syntactic distance. In other words, a perfect correlation will result in a colour continuum, whereas a low correlation will result in a mosaic-like map.<sup>2</sup>

Note that in this application of MDS only the relations among the colour components are fixed. The assignment of the colour components to the variant dimensions is arbitrary in itself. Therefore, swapping colour components may have a substantial effect on the visual result, especially for people with red/green colour blindness. Also, the three colour components contribute differently to the brightness of a map when viewed on a computer screen than when viewed on paper. Therefore, MDS map regions might deviate to an extent depending on viewer perception and communication medium.

Finally, although several MDS methods are available for reducing the set of 510 feature variant dimensions, only the Classical MDS procedure is used in this work. This method is known as a metric MDS procedure because it uses the actual distance values. In non-metric procedures like Kruskal's Non-metric

---

<sup>2</sup> The space between dialect locations on the MDS maps is partitioned by using the Delaunay triangulation to obtain a pattern of polygons known as Voronoi polygons or Dirichlet tessellation. This technique for determining dialect areas is also used in Goebel (1982) and Heeringa (2004). Alternatively, an interpolation procedure could be applied to colour the space between dialect locations.

MDS and Sammon's Non-linear Mapping, the ranks of the distance values are used instead.

## 2.6. Classifying Dutch dialects using a syntactic measure

In this section, the results of the application of the MDS procedure to the syntactic distance matrix are presented. First, an overview of the results is given. Then, the results for each of the seven SAND1 domains are reviewed. Finally, the aggregate SAND1 MDS dialect map is presented in Figure 2-6.

The correlation between the original set of feature variants for each syntactic domain and the corresponding representation after reducing each set to three dimensions via MDS is shown in Table 2-4. In most applications correlations below 0.8 tend to be too inaccurate to be interpreted meaningfully, whereas results between 0.9 and 1 are generally considered to be high. Therefore, based on the values in Table 2-4, the MDS dialect maps can be expected to visualise the actual dialect classification quite accurately. A general impression of the effect of each syntactic domain on the aggregate SAND1 correlation value in the last row of Table 2-4 can be obtained by combining the correlation value with the relative number of feature variants that were included in the measurement in relation to the total number of variants in SAND1 as listed in column 3.

Table 2-4: Correlation between the original sets of SAND1 feature variants and the corresponding representation after reducing each set to three dimensions via MDS.

<i>Syntactic domain</i>	<i># variants</i>	<i>% variants</i>	<i>Correlation (r)</i>
Complementisers	101	19.8	0.94660937
Subject pronouns	172	33.7	0.88065714
Expletives	13	2.5	0.87393870
Subject doubling	54	10.6	0.95438211
Subject cliticisation following yes/no	30	5.9	0.99025193
Reflexive and reciprocal pronouns	78	15.3	0.93453301
Fronting	62	12.2	0.77975377
<i>SAND1</i>	<i>510</i>	<i>100.0</i>	<i>0.95905712</i>

Figure 2-2 visualises the syntactic distances between the Dutch dialects with respect to complementisers based on 101 variant comparisons for each dialect pair. This is almost 20% of 510, the total number of available SAND1 variants. The correlation value of 0.94 means that this map visualises the geographical distribution of complementisers quite accurately. Figure 2-2 shows a distinct correlation between geographical distance and variation with respect to comple-

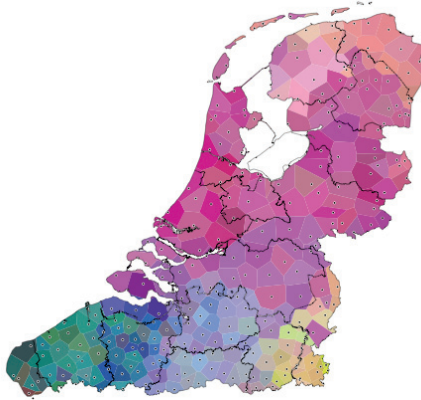


Figure 2-2: MDS map visualising syntactic distances with respect to complementisers.

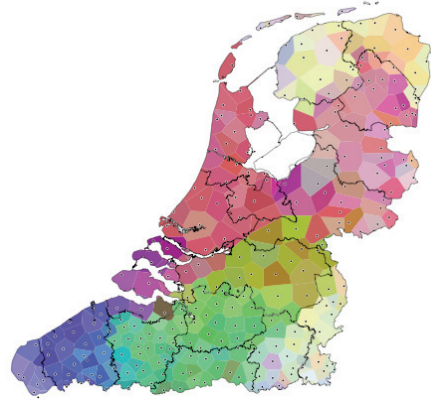


Figure 2-3: MDS map visualising syntactic distances with respect to subject pronouns.

mentisers, since neighbouring dialect locations have corresponding colours. The result is a colour continuum with more or less clustered dialect areas.

Figure 2-3 visualises the syntactic distances with respect to subject pronouns, based on 172 variant comparisons per dialect pair. This syntactic domain comprises about one-third of the total number of available variants in SAND1. Therefore, it has a substantial effect on the aggregate SAND1 dialect map. The correlation value of 0.88 between the original data and the dimension-reduced data is rather high, meaning that this map visualises the geographical distribution of subject pronouns quite well. Furthermore, note that most borders of the colour-clustered areas in Figure 2-3 are almost identical to the discernable regions in Figure 2-2.

Only a description of the MDS dialect map is provided for the data with respect to expletives, which is based on merely 13 variant comparisons per dialect pair. This is only 2.5 percent of the total number of available variants in SAND1. The resulting map is a mosaic of dialect colours which indicates a weak correlation between geographical distance and syntactic distance, since neighbouring dialect locations do not have corresponding colours. But, even though the map does not show a colour continuum, the correlation value of 0.87 is still quite high. However, this can be explained by the fact that only 13 feature variant dimensions were used, which is not enough data for the MDS procedure to be reliably represented in three dimensions.

Figure 2-4 visualises the syntactic distances with respect to reflexive and reciprocal pronouns based on 78 variants. This is 15 percent of the total number of available SAND1 feature variants. Again, the correlation value of 0.93 is quite high. Interestingly, the map in Figure 2-4 significantly resembles the descriptive Dutch dialect area classification with respect to reflexives in Barbiers and Bennis



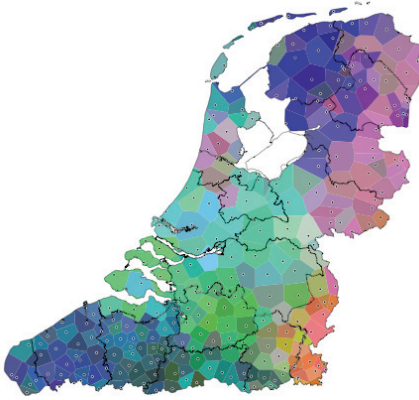


Figure 2-4: MDS map visualising syntactic distances with respect to reflexive and reciprocal pronouns.



Figure 2-5: MDS map visualising syntactic distances with respect to fronting.

(2004). In this description, which is also based on SAND1, five main dialect areas are distinguished: an eastern group, a Frisian area, a West- and East-Flemish region, a Flemish Limburg group and an Antwerp and south-west and central Dutch area. Contours of these generalisations can also be found on the map in Figure 2-4.

Figure 2-5 shows the correlation between geographical and syntactic distance with respect to fronting, based on 62 variants per dialect pair. This is about 12 percent of the total number of SAND1 variants. This mosaic-like map clearly illustrates that there is little significant correlation between geographical distance and syntactic distance because many neighbouring dialect locations do not have corresponding colours. This may indicate that the SAND1 fronting data is actually made up of several fronting subdomains which do not have corresponding geographical distributions. This analysis would explain the low correlation value of 0.78 as an indication that the fronting data is of a too heterogeneous nature to be accurately displayed in one three-dimensional MDS map. In other words, at least four dimensions would be required in order to adequately represent the fronting data. This observation makes the aggregate SAND1 dialect map even more interesting.

The SAND1 MDS dialect map is shown in Figure 2-6. This map visualises the correlation between geographical distance and syntactic variation in Dutch dialects. As can be seen in Figure 2-6, aggregating all these different distribution patterns in the SAND1 domains, including the heterogeneous fronting data, results in a remarkably homogeneous colour continuum with easily discernable dialect regions. Also note the strikingly high correlation value of 0.96, considering the diversity of the SAND1 data domains. This means that only few of the most differentiating distance relationships were lost during the MDS procedure.



Figure 2-6: The SAND1 MDS dialect map based on a syntactic Hamming distance measure.

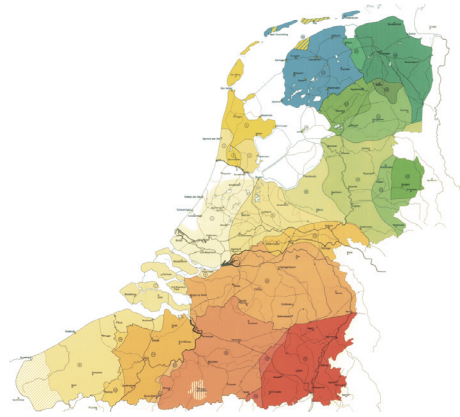


Figure 2-7: The Daan and Blok dialect map based on subjective judgements (see also Figure 2-1).

Therefore, the SAND1 MDS dialect map in Figure 2-6 can be considered a reliable visualisation of syntactic variation in Dutch dialects.

## 2.7. Comparing the computational and perceptual dialect classifications

In Figure 2-6 and Figure 2-7 the computational MDS dialect map based on a syntactic measure is shown next to the perceptual Daan and Blok dialect map based on subjective judgements. The correspondence between the objective and the subjective classification of Dutch dialect varieties is quite remarkable. The classification of the Dutch dialects in the bottom half of both maps is nearly identical, although significant differences are visible as well in the central eastern and central western regions. The MDS dialect map only reveals a few relatively subtle dialect area borders in the top half of the map, whereas the Daan and Blok dialect map shows many dialect area borders within this region.

These discrepancies might indicate that these distinct dialect borders do not exist on a syntactic level or that these borders have been fading during the last century. However, considering the resemblance between the Flemish area on the Daan and Blok dialect map as classified by Belgian dialectologists and the Flemish region on the MDS dialect map, it seems that local dialect speakers' prejudice might also play a significant differentiating role in perception of syntactic variation between neighbouring dialects in the Netherlandic part of the Daan and Blok dialect map. Furthermore, non-expert dialect speakers tend to be more sensitive to lexical and phonological differences than to variation on a syntactic level.

The correspondence between the Frisian area and the Limburg region with respect to subject pronouns in Figure 2-3 is still visible in Figure 2-6 as shades of purple. Although this might indicate a SAND1 data bias with respect to subject pronouns, it also shows a non-local dialect area relation that could never have been derived using Daan and Blok's arrow method.

To conclude, a few notable highlights of this dialectometrical perspective on syntactic variation are provided. First, the objective classification of Dutch dialect varieties based on a syntactic measure highly resembles the classification based on subjective judgements on the Daan and Blok dialect map. Second, the Belgian dialect classification on the Daan and Blok map based on more objective expert judgements corresponds to a higher degree with the classification based on the objective syntactic measure than with the Netherlandic dialect classification based on intuitive judgements. These two points confirm and validate the syntactic measurement method. Third, although syntactic variation appears in many feature dimensions, its aggregate geographical distributions can be represented accurately in merely three dimensions after reduction via MDS. This is a computational confirmation of the intuition that syntactic variation is organised in groups of related patterns. Additional research will include refinements of the syntactic measure and analysis of feature dependencies for further exploration.

## 2.8. References

- Barbiers, S., Bennis, H., Devos, M., Vogelaar, G. de, Ham, M. van der (eds), 2005. *Syntactic Atlas of the Netherlandic Dialects*, Volume 1. Amsterdam University Press, Amsterdam.
- Barbiers, S., Bennis, H., 2004. *Reflexieven in dialecten van het Nederlands. Chaos of structuur?*. In: Caluwe, J. de, Schutter, G. de, Devos, M. et al. (eds), *Schatbewaarder van de taal*. Johan Taeldeman. Liber Amicorum. Academia Press Gent en Vakgroep Nederlandse Taalkunde Universiteit Gent, Gent, 43–58.
- Cornips, L., Jongenburger, W., 2001. *Elicitation techniques in a Dutch syntactic dialect atlas project*. In: Broekhuizen, H., Wouden, T. van der (eds), *Linguistics in the Netherlands*, 2001, John Benjamins, 53–63.
- Daan, J., Blok, D., 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Volume XXXVII, Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.
- Goebel, H., 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Volume 157, Philosophisch-Historische Klasse Denkschriften. Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of Rase, W., Pudlatz, H..
- Goeman, A., 2000. *Perception of Dialect Distance: Standard and Dialect in Relation to New Data on Dutch Varieties*. In: Long, D., Preston, D. (eds), *Handbook of perceptual dialectology*, Volume II, 2000, John Benjamins, 137–151.

- Hearing, W., 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis Rijksuniversiteit Groningen, Groningen.
- Nerbonne, J., Kretzschmar, W., 2003. *Introducing Computational Methods in Dialectometry*. In: Nerbonne, J., Kretzschmar, W. (eds), *Computational Methods in Dialectometry*, Special issue of *Computers and the Humanities*, Volume 37(3), 245–255.
- Séguy, J., 1971. *La relation entre la distance spatiale et la distance lexicale*, *Revue de Linguistique Romane*, Volume 35, 335–357.
- Torgerson, W., 1952. *Multidimensional scaling: I. Theory and method*. *Psychometrika*, Volume 17, 401–419.

### 3. Measures of syntactic distance and the role of geography

“Incorporating regression analyses and feature variables”\*

*Spruit, M.R., 2006. Measuring Syntactic Variation in Dutch Dialects. In: Nerbonne, J., Kretzschmar, W. (eds), Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation, Volume 21, Oxford University Press, Oxford, 493–506.*

This research applies dialectometrical methods to purely syntactic dialect data. It will be shown that there is geographical cohesion in syntactic variation when viewed in the aggregate. The amount of syntactic variation which can be accounted for by geography will be determined. Dialectometrical techniques will be used to develop an additive measure of syntactic differences. Multidimensional scaling will be applied to visualise the geographical distribution of the Dutch dialects with respect to syntactic variation in the aggregate. The Dutch dialect map based on a syntactic measure will be compared with a dialect map based on subjective judgements and a dialect map based on pronunciation differences to put the syntactic measurement results into perspective. An alternative way to measure syntactic distance will be presented and will provide indications for future research to more accurately quantify syntactic variation.

#### 3.1. Introduction

This research combines and extends work from the research fields of dialectometry and syntactic variation to answer the question whether there is geographical cohesion in syntactic variation when dialectal differences are viewed in the aggregate. Dialectometrical techniques are used to develop an additive measure of syntactic differences. These techniques can also provide an answer to the question of how much of the recorded syntactic variation can be accounted for by geography.

The Daan and Blok (1969) map of the Dutch dialects shown in Figure 3-1 can be seen as an early attempt to represent dialectal differences in the aggregate. The classification of the Dutch dialects on this map is derived using subjective judgements of local speakers, local experts and Daan and Blok themselves. However, Spruit (2005) notes a number of practical and methodological prob-

---

\* This research is being carried out in the context of the NWO project *The Determinants of Dialectal Variation*, number 360-70-120, P.I. J. Nerbonne. Please visit <http://dialectometry.net> for more information and relevant software. I would like to thank Sjeff Barbiers, John Nerbonne and two anonymous reviewers for their valuable comments on earlier versions of this paper.

lems which may have significantly influenced the outcome of this classification of Dutch dialect areas based on perceptual differences. Therefore, objective methods are required to assign numerical values to linguistic phenomena to aggregate individual dialect differences. These dialectometrical methods were first described in Séguy (1971) and further investigated in Goebel (1984) and Heeringa and Nerbonne (2001), among others. However, these dialectometrical studies were mainly limited to lexical and phonological data. Most notable in this context is the application of the Levenshtein method to aggregate differences in dialect pronunciation in Heeringa (2004).

The first application of dialectometrical methods to purely syntactic dialect data is described in Spruit (2005). This work first reviews the results of the application of a measure based on binary comparisons between syntactic variables for each of the seven available syntactic subdomains. Then, all dialect differences are aggregated and the resulting map of the Dutch dialects with respect to syntactic variation is compared with the Daan and Blok map based on subjective judgements.

The present paper extends the work described in Spruit (2005) in several ways. First, the geographical distribution with respect to syntactic variation in Dutch dialects is also compared with the map of the Dutch dialects based on a measure of pronunciation differences in Heeringa (2004). Second, geographical distances are correlated with syntactic distances using regression analyses to investigate how much of the recorded syntactic variation can be accounted for by geography. Finally, syntactic variables are annotated with abstract features to obtain a set of underlying feature variables. These underlying variables are used to measure the differences between the Dutch dialects. The results are compared with the measurement results based on atomic variables.

The term *variable* is central to this work. Generally speaking, a variable may be defined as a linguistic unit in which two language varieties can vary. In the context of this work a syntactic variable is defined as a form or word order in a syntactic context in which two dialects can differ. Several types of variables can be distinguished. First, the main part of this paper uses syntactic variables as they have been recorded, without interpretations. These variables are referred to as atomic variables. Second, atomic variables can be combined to form composite variables. These variables are not used in this paper. Third, the final part of this paper introduces feature variables which are formulated by manually annotating syntactic variables with linguistic feature information. These variables can be defined using insights from the research field of syntactic theory.

This paper is structured as follows. The data with respect to syntactic variation in Dutch dialects are introduced in Section 2. The syntactic measurement procedure and the analysis technique are described in Sections 3 and 4 respectively.

The resulting geographical colour map of the Dutch dialect area based on a syntactic measure is presented in Section 5 and is related to distributions based on perception and pronunciation in Section 6. The latter section also includes an analysis of the correlation between syntactic variation and geographical distance. An alternative measure of syntactic distance based on feature variables is presented in Section 7. The measurement results based on feature variables are compared with the results based on atomic variables in Section 8. The paper concludes with a recapitulation of the most significant results and directions for future research in Sections 9 and 10.

### **3.2. Syntactic Atlas of the Dutch Dialects**

Until recently dialectometrical research was mainly limited to lexical and phonological data because no extensive collection of purely syntactic data was available. This situation has changed with the arrival of the first part of the Syntactic Atlas of the Dutch Dialects (SAND1, Barbiers et al., 2005). It contains 145 maps showing the geographical distribution of syntactic variables in 267 Dutch dialects. Geographical distributions of individual syntactic variables are shown in 134 maps.<sup>1</sup> The other 11 maps display correlations between syntactic variables. The second volume of the SAND will appear in 2008 and will contain data with respect to syntactic variation in verbal clusters and negation.

The SAND data were collected using a wide range of both written and oral syntactic elicitation techniques (Cornips and Jongenburger, 2001). First, a literature study was conducted to prepare a written questionnaire containing 424 questions. This was sent out to 850 informants to optimally design the interviews with local dialect speakers. The written questionnaire included indirect grammaticality judgements, translation tasks and completion (fill-in-the-blank) tasks. Then, seven pilot interviews were conducted to evaluate the validity of the elicitation tests. The oral elicitation tasks included translations, completion tasks, meaning questions and repetition tasks.

At each measuring point in the Netherlands the interview was not carried out by the field workers themselves but by local dialect speaking assistants, since most field workers did not speak the local dialect. The field worker would first instruct the assistant. Then, the assistant conducted the interview with the informant in the local dialect to avoid accommodation effects. The field worker's main role was to ensure adherence to the interview protocol. In Belgium no separate interview assistants were employed because the Belgian field workers were regional dialect speakers themselves. All in all, it may be safely assumed

---

<sup>1</sup> Spruit (2005) mentions 135 maps. However, this included SAND1 map 73b which does not contain unique data. It has been left out of the measurement procedures reported on in this work.

that the extensive SAND methodology provides a solid foundation for the results presented in this paper.

SAND1 covers syntactic domains related to the left periphery of the clause and pronominal reference. It contains data with respect to complementisers, subject pronouns, expletives, subject doubling, subject cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. In the context of this work SAND1 contains 507 syntactic variables distributed over 134 maps. Each map represents one syntactic context and each map symbol represents one syntactic variable.<sup>2</sup> Therefore, the 507 syntactic variables average to slightly less than four variables per syntactic context.

*Table 3-1: Map 68a in SAND1 shows the five syntactic variables in the context of weak reflexive pronoun as object of inherent reflexive verb.*

Context:	Weak reflexive pronoun as object of inherent reflexive verb				
Variables:	{ zich, hem, zijn eigen, zichzelf, hemzelf }				
Example:	Jan	herinnert	<u>zich</u>	dat	verhaal wel.
	Jan	remembers	himself	that	story AFFIRM
	“John certainly remembers that story.”				

*Table 3-2: Map 82b in SAND1 shows the six syntactic variables in the context of short object relative.*

Context:	Short object relative				
Variables:	{ die, dat, wie, der, den/dem, as }				
Example:	Dat	is	de	man	<u>die</u> ze geroepen hebben.
	That	is	the	man	who they called have
	“That is the man who they have called.”				

Table 3-1 illustrates the mapping from SAND1 maps to syntactic variables with an example of variables in one syntactic context in the reflexives subdomain. Map 68a in SAND1 shows the geographical distribution of five syntactic variables in the context of *weak reflexive pronoun as object of inherent reflexive verb*. The variables *zich*, *hem*, *zijn eigen*, *zichzelf* and *hemzelf* have been recorded in this context throughout the Dutch language area. In this paper this map represents one of the 134 syntactic contexts and five of the 507 syntactic variables. Table 3-2 further illustrates this mapping with an example of variables in a syntactic context in the fronting subdomain. Map 82b in SAND1 shows the geographical distribution of six syntactic variables in the context of *short object relative*. In this

---

<sup>2</sup> Syntactic variables are referred to as syntactic features in Spruit (2005).



context the variables *die*, *dat*, *wie*, *der*, *den/dem* and *as* were observed. Therefore, this map represents six of the 507 syntactic variables in this paper.

To summarise, the variable-oriented SAND contains a wealth of purely syntactic data suitable for dialect-geographical research. Dialectometrical methods can be applied after the lists of dialect locations per syntactic variable are transformed into sets of occurring syntactic variables per dialect location.

### 3.3. Hamming Distance Measure

The results presented in this work are based on Hamming distance measurements between syntactic variables. The syntactic distance between a pair of dialects is calculated by comparing the occurrences of all syntactic variables between each dialect pair. If a variable is observed in dialect A but not in dialect B, or if a variable is not recorded in dialect A but does occur in dialect B, then the distance between dialects A and B is incremented by 1. Most results in this paper are based on atomic variables as described in the introduction.

Table 3-3: Fragment of the distance measurement between two dialects using five syntactic variables.

	Lunteren	Veldhoven	distance
[sand1,68a]: <i>zich</i>	+	+	0
[sand1,68a]: <i>hem</i>	-	-	0
[sand1,68a]: <i>zijn eigen</i>	+	-	1
[sand1,68a]: <i>zichzelf</i>	-	-	0
[sand1,68a]: <i>hemzelf</i>	-	-	0
		total:	<hr/> 1

Table 3-3 illustrates a fragment of the procedure to measure the syntactic distance between the dialects of Lunteren and Veldhoven using atomic variables. It lists the occurring variables in the syntactic context *weak reflexive pronoun as object of inherent reflexive verb* as shown in Table 3-1. The variables *zich* and *zijn eigen* were recorded in Lunteren and the variable *zich* was observed in Veldhoven. Since the variable *zich* is available in both dialects, the dialect distance is not increased. The variables *hem*, *zichzelf* and *hemzelf* do not occur in either of these two dialects and have no effect on the distance value either. However, the variable *zijn eigen* occurs in Lunteren but not in Veldhoven. This increases the dialect distance between Lunteren and Veldhoven by 1.

This measurement based on binary comparisons of syntactic variables is carried out for all 507 variables, and the procedure is repeated for all  $(267 * 266) / 2 =$

35511 unique dialect pairs.<sup>3</sup> The final result is a Hamming distance matrix a part of which is shown in Table 3-4. In this matrix each distance value represents the total number of different syntactic variable realisations between one pair of dialects. For example, the matrix shows that 47 different variable realisations were recorded between the dialects of Lunteren and Veldhoven after comparing all 507 syntactic variables.

Table 3-4: Fragment of the SAND1 Hamming distance matrix.

	Lunteren	Bellingwolde	Hollum	Doel	Sint-Truiden	Veldhoven
Lunteren		66	52	122	77	47
Bellingwolde	66		56	134	81	51
Hollum	52	56		116	63	59
Doel	122	134	116		115	111
Sint-Truiden	77	81	63	115		72
Veldhoven	47	51	59	111	72	

### 3.4. Multidimensional Scaling Analysis

Multidimensional scaling (MDS) is applied to analyse the relationships in the dialect distance matrix. The MDS procedure was first described in Torgerson (1952) and displays the structure of distance data as a geometrical picture. In the context of this work, MDS is used to represent the matrix of differences between dialect locations in as low-dimensional a space as possible. The results are visualised with dialect colour maps.

When the MDS technique is applied to the syntactic distance matrix, the set of 267 dialect dimensions for each dialect is scaled down to a coordinate in a three-dimensional space. This coordinate is the minimisation of changes in the distance matrix. The coordinates do not directly correspond to actual dialect distances anymore.

---

<sup>3</sup> A distance matrix is always symmetric because the distance from dialect A to dialect B is always identical to the distance from dialect B to dialect A. Therefore, only the distances in either the lower left part or the upper right part need to be included in the measurement. Also, all distances from a dialect to itself are excluded from the procedure.

The three-dimensional coordinates are then used as values between light and dark of the three colour components red, green and blue. This results in a unique composite colour for each dialect location. Then, the dialect points on the maps are blown up to small areas until they border each other and there is no uncoloured space left.<sup>4</sup> Neighbouring dialect areas will have corresponding colours if there is a correlation between geographical distance and syntactic distance. Therefore, a perfect correlation will result in a colour continuum, whereas a low correlation will result in a mosaic-like map.

All MDS results presented in this paper are based on the Classical MDS procedure. This method is known as a metric MDS procedure because it uses the actual distance values to reduce the set of 267 dialect dimensions. A non-metric procedure like Kruskal's Non-metric MDS uses the ranks of the distance values instead. In general, results are comparable.:

### **3.5. Map of the Dutch Dialects**

Figure 3-2 shows the SAND1 MDS dialect map derived from the Hamming distance matrix. The map visualises the correlation between geographical distance and syntactic variation in Dutch dialects and incorporates all 507 syntactic variables in the seven SAND1 subdomains. The dialect maps for the SAND1 subdomains are presented and discussed extensively in Spruit (2005). The SAND1 MDS dialect map can be characterised as a continuum of gradually changing dialect areas. This typology not only supports the view that dialect varieties are organised in areas but also the view that these areas form a continuum without sharp boundaries (Heeringa and Nerbonne, 2001).

A correlation coefficient of nearly 0.96 is achieved using the Classical MDS method. This value indicates how much of the syntactic variance is represented in the first three dimensions of the MDS solution, which, in this context, quantifies the amount of syntactic variance represented in the map colours. Correlation values between 0.9 and 1.0 are quite high, indicating that the MDS result faithfully represents the information in the original distance matrix. Thus, the claim can be made that the SAND1 MDS dialect map visualises the actual dialect relationships accurately.<sup>5</sup>

---

<sup>4</sup> The space between dialect locations on the MDS maps is partitioned by using the Delaunay triangulation to obtain a pattern of polygons known as Voronoi polygons or Dirichlet tessellation. This technique for determining dialect areas is also used in Goebl (1982) and Heeringa (2004). Alternatively, an interpolation procedure could be applied to colour the space between dialect locations.

<sup>5</sup> Application of Kruskal's Non-metric MDS method results in a nearly identical dialect map. This can be interpreted as a confirmation of the reliability of the SAND1 MDS map shown in Figure 3-2.

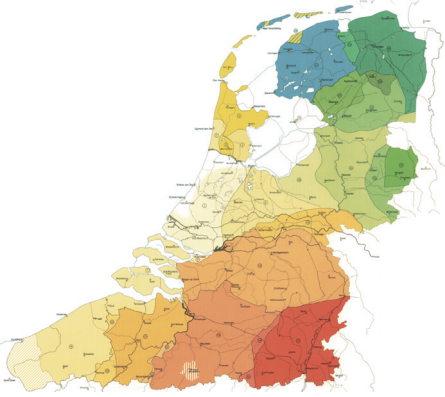


Figure 3-1: The Daan and Blok map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok, 1969).



Figure 3-2: The SAND1 map of the Dutch dialects based on a syntactic measure after application of the Classical MDS procedure.



Figure 3-3: Map of the Dutch dialects based on pronunciation differences after application of Kruskal's Non-metric MDS procedure (reprinted from Heeringa, 2004).



Figure 3-4: The selection of 21 dialect locations used in the regression analyses.

### 3.6. Syntactic variation in context

#### 3.6.1. Syntax versus Perception

The SAND1 MDS map in Figure 3-2 is shown next to the Daan and Blok dialect map in Figure 3-1. This view puts the geographical distribution of syntactic variation into a perceptual perspective. The objective SAND1 dialect area classification based on a syntactic measure looks quite similar to Daan and Blok's

subjective dialect area classification based on subjective judgements. The similarities are even more remarkable when taken into account the fact that the colours used in the Daan and Blok dialect map were chosen more or less intuitively, although corresponding to a gradually increasing divergence from Standard Dutch (Goeman, 2000).

However, there are some notable differences between these two maps as well. For example, the Daan and Blok dialect map shows no differentiation within dialect areas. This contradicts the intuition that dialects are also organised in a continuum without sharp boundaries. Another significant difference can be found in the north-eastern part of the Netherlands. The Daan and Blok map shows a number of clearly distinguishable dialect areas in shades of green in this region, but the SAND1 MDS map reveals only a few relatively subtle dialect areas in the north-eastern area. The Frisian area, in distinctive blue on the Daan and Blok map, is also much less pronounced on the SAND1 map. It could be that these perceived dialect borders simply do not exist on a syntactic level. After all, it is often assumed that non-expert dialect speakers tend to be more sensitive to lexical and phonological differences than to variation on a syntactic level. A comparison of the SAND1 MDS dialect map with Heeringa's MDS dialect map based on pronunciation differences may support this argument.

### **3.6.2. Syntax versus Pronunciation**

The SAND1 MDS map in Figure 3-2 is shown above the Heeringa MDS dialect map based on pronunciation differences in Figure 3-3. This view illustrates the geographical distribution of syntactic variation in comparison to pronunciation. The pronunciation dialect map shows a smooth dialect continuum except for the Frisian city dialect islands in the blue Frisian area. These varieties are symbolised with diamonds to indicate that they do not belong to the group in which they are found geographically. Apart from the general observation that the SAND1 MDS map shows a less smooth colour continuum overall, the most interesting discrepancy between these two maps is arguably the complete absence of the Frisian city dialect islands in the SAND1 MDS map. Upon closer examination, however, only three out of thirteen Frisian dialect islands on the map in Figure 3-3 also occur as dialect locations in the SAND.<sup>6</sup> This mismatch of locations already explains most of the discrepancy between Figure 3-2 and Figure 3-3, since city dialect islands are by definition of a local and isolated nature.

Furthermore, Van Bree (1994) shows that “[...] in the sixteenth century in the wake of a major political upheaval [...] Town Frisian emerged as Dutch spoken

---

<sup>6</sup> The three Frisian city dialect islands in Figure 3 which also occur in the SAND are Midsland, Heerenveen and Kollum.

by Frisians”. It is “[...] the result of a second language acquisition process which was broken off at a certain point, after which conventionalisation took place.” (Van Bree, 1994:80-81). Van Bree concludes that Town Frisian leans especially towards Standard Dutch at the lexical and lexico-phonological levels because these linguistic levels are known to have a low stability gradient. These linguistic levels can be acquired quickly and go hand in hand with a much higher degree of awareness. Syntax, on the other hand, is known to have a high stability gradient which makes it very linguistically stable. Once it is acquired, slowly, it becomes very hard to unlearn. Moreover, most language users are scarcely aware, if at all, of syntactic elements (Van Bree, 1992). Therefore, the interrupted second language acquisition process has caused Town Frisian to resemble Standard Dutch on the pronunciation level but remained Frisian-like at the syntactic level. This historical background of the Frisian city dialects completes the explanation of the main discrepancy between the syntax-based and pronunciation-based dialect maps in Figure 3-2 and Figure 3-3.

Finally, there is no visual correspondence at the pronunciation level in Figure 3-3 between the central-northern Frisian area and the south-western Flemish region. Figure 3-2, on the other hand, does indicate some correspondence between these areas in shades of purple at the syntactic level. Apart from these observations the SAND1 MDS map seems to correlate with the pronunciation-based MDS map to a reasonable extent. However, statistical analyses will have to be performed to more precisely address the extent of the correlation between these linguistic levels.

### 3.6.3. Syntax versus Geography

Regression analyses were performed to determine how much of the syntactic variance can be explained with geographical distance. A selection of 21 dialects was used.<sup>7</sup> This amounts to  $(21 * 20) / 2 = 210$  dialect pair comparisons. Figure 3-4 shows that the dialects were chosen in such a way that a cross section of dialect varieties throughout the Dutch language area was obtained. A similar approach based on pronunciation differences is presented in Heeringa and Nerbonne (2001). The regression analysis shown in Figure 3-5 results in a correlation value of nearly 0.75, which means that about  $(0.75)^2 = 56$  percent of syntactic distance can be explained with geographical distance in a linear relationship. Interestingly, using a logarithmic function to describe the relationship between syntactic and geographical distance results in a somewhat lower correlation of 0.69. This is different from the results at the pronunciation level in

---

<sup>7</sup> The following 21 dialects were used in the regression analyses, listed from the north-east to the south-west of the Dutch language area: Nieuw-Scheemda, Spijkerboor, Rolde, Hooghalen, Diever, Staphorst, Wezep, Epe, Hoog Soeren, Lunteren, Geldermalsen, Waspik, Zundert, Ossendrecht, Doel, Koewacht, Zaffelare, Gent, Deinze, Waregem and Kortrijk.

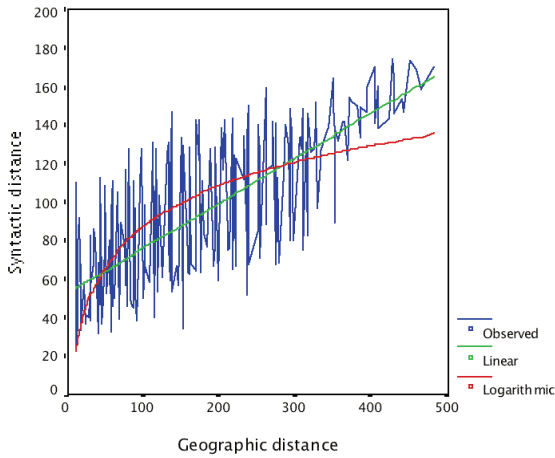


Figure 3-5: Geographical distances versus syntactic distances using the subset of 21 dialect locations shown in Figure 3-4.

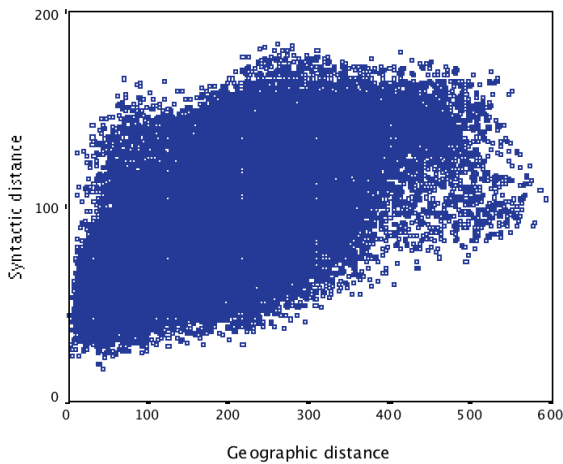


Figure 3-6: Geographical distances versus syntactic distances using all 267 dialect locations.

Heeringa and Nerbonne (2001) where a logarithmic function best describes the relationship between geographical distance and pronunciation differences.

Another regression analysis was performed to determine the correlation between syntactic variance and geographical distance using all  $(267 * 266) / 2 = 35511$  dialect pairs. This analysis is visualised in Figure 3-6 and results in a correlation value of nearly 0.55, which means that around 30 percent of syntactic distance can be explained with geographical distance when all available SAND1 data are taken into account.

### 3.7. Feature Variables

All results presented in the previous sections have been derived from a syntactic measure based on binary comparisons between atomic variables as described in the introduction. In this section the first results are presented using a syntactic measure based on binary comparisons between feature variables.

Feature variables have been formulated to abstract away from the atomic variables as they occur. The idea is to measure differences between dialects at a more structural level which may only be obtained after syntactic analysis. Feature variables can help capture the notion that some variables are less different from each other than other variables. Using feature variables the syntactic distance between the atomic variables *zich* and *zijn eigen* can be assigned a higher value than the distance between the atomic variables *zich* and *zichzelf*. This strategy combines syntactic research from both empirical and theoretical areas. A part of the mapping from atomic variables to feature variables with respect to reflexive pronouns is presented in Table 3-5.<sup>8</sup>

Table 3-5: Mapping from atomic variables (first column) to feature variables (first row) with respect to reflexive pronouns.

	personal <i>hem</i>	reflexive <i>zich</i>	possessive <i>zijn</i>	ownness <i>eigen</i>	focus <i>zelf</i>
hem	+				
hemzelf	+				+
zich		+			
zichzelf		+			+
zijn			+		
zijn zelf			+		+
zijn eigen			+	+	
zijn eigen zelf			+	+	+

The column headers in Table 3-5 show the core set of feature variables such as *personal* and *focus* in the reflexives subdomain. The most relevant atomic variables are listed in the row headers. A plus sign in a given cell indicates that the feature variable in the column header is represented by the atomic variable in the row. For completeness, feature variables in syntactic contexts related to reciprocals and one-pronominalisation are listed in the appendix in Table 3-7 and Table 3-8. These features carry less weight during the dialect distance

<sup>8</sup> Helke (1970), Postma (1997) and Barbiers and Bennis (2004), among others, argue that reflexives commonly have possessive structures.



Table 3-6: Fragment of the distance measurement between two dialects using five feature variables (first column).

	Lunteren { zich, zijn eigen }	Veldhoven { zich }	distance
personal	-	-	0
reflexive	+	+	0
possessive	+	-	1
ownness	+	-	1
focus	-	-	0
			total: 2

measurements because they only describe the variation with respect to the syntactic contexts *reciprocal pronouns* and *one pronominalisation*.<sup>9</sup>

The syntactic measure determines the distance between a pair of dialects by comparing all occurring feature variables between two dialects. If a feature variable is represented in dialect A but not in dialect B, or if a feature variable does not manifest itself in dialect A but does occur in dialect B, then the distance between dialects A and B is incremented by 1.

Table 3-6 illustrates a fragment of the measurement procedure using feature variables for the dialect pair Lunteren and Veldhoven. It lists the feature variables represented by the atomic variables in the syntactic context *weak reflexive pronoun as object of inherent reflexive verb* as shown in Table 3-1. The features *reflexive*, *possessive* and *ownness* are represented in the atomic variables *zich* and *zijn eigen* as recorded in Lunteren. In Veldhoven only the feature variable *reflexive* is reflected in the atomic variable *zich*. Since the feature variable *reflexive* is available in both dialects, the dialect distance is not increased. The features *personal* and *focus* are not represented in either of these two dialects and have no effect on the distance value either. However, the features *possessive* and *ownness* are both reflected in Lunteren but not in Veldhoven. Therefore, the dialect distance between Lunteren and Veldhoven is increased by two.

Abstracting away from occurring atomic variables to represented feature variables has several advantages when measuring dialect distances. For example, a measure based on atomic variables cannot differentiate between the variables *zich* and *zichzelf* on the one hand and *zich* and *zijn eigen* on the other hand. Both are assigned a distance value of one because in both cases the two variables are different. A measure based on feature variables also assigns a distance value of

---

<sup>9</sup> Furthermore, the feature *nominative* is used in the reflexives subdomain to help describe the variation with respect to the syntactic context *reflexive pronouns in adverbial middle constructions* as shown in SAND1 map 77a.

one between the variables *zich* and *zichzelf* because they share the *reflexive* feature variable but differ with respect to the *focus* feature variable as shown in Table 3-6. However, the distance between the variables *zich* and *zijn eigen* is assigned a distance value of three because the three underlying features for these variables do not match at all. The atomic variable *zich* reflects the *reflexive* feature variable and the atomic variable *zijn eigen* represents the *possessive* and *onwness* feature variables.

Differentiation between dissimilar variable pairs is possible by virtue of the abstract nature of feature variables. There is no one-to-one mapping from atomic variables to feature variables as can be seen in Table 3-5. This property can be used to develop a more refined syntactic measure to further increase accuracy. For example, a syntactic distance measure could take into account both the number of similarities as well as the number of differences in a so-called similarity-difference distance coefficient. Such a distance coefficient would allow for a differentiation between three variable comparison states. First, a variable can occur in dialect A but not in B. Second, a variable can occur in both dialects. Third, a variable can occur in neither dialect. This is in contrast with a measure using distance values which does not enable differentiation between the second and third comparison states. Results using a measure based on distance coefficients will be reported on in future research.

An obvious downside of using feature variables is the requirement of feature formulation and annotation of all data. All atomic variables in all syntactic contexts need to be assigned syntactic features. This task requires consultation with syntactic theorists to formulate meaningful feature variables which also allow for a partitioning of the available data which differentiates the atomic variables from each other.

### **3.8. Atomic Variables versus Feature Variables**

The measurement results using either atomic variables or feature variables have been compared with respect to the SAND1 data in the reflexives subdomain. The geographical distributions turn out to be nearly identical after application of the MDS procedure. The measure using atomic variables consisted of 75 comparisons between each pair of dialects, and application of the MDS procedure results in a three-dimensional solution which correlates highly with the original distance matrix ( $r = 0.93$ ).<sup>10</sup> The measure using feature variables included 61 comparisons between each pair of dialects and results in a correlation of 0.94. These correlations indicate that both atomic variables as well as feature variables can be used to faithfully illustrate syntactic variation in three dimensions. Furthermore, both maps correspond to a reasonable extent to the de-

---

<sup>10</sup> The MDS map visualising syntactic distances with respect to reflexive and reciprocal pronouns is printed in Spruit (2005:186).

scriptive Dutch area classification with respect to reflexives in Barbiers and Bennis (2004). This description distinguishes 5 main dialect areas in the geographical distribution of variation with respect to reflexives. Contours of these generalisations can also be found on the MDS maps.

The fact that the syntactic measure using feature variables does not yield more differentiating results with respect to the reflexives subdomain is not unexpected. Using SAND1 synthesis map 76a and the descriptive classification in Barbiers and Bennis (2004) as references, the application of the syntactic measure using atomic variables already results in a quite adequate geographical distribution of variation with respect to reflexives. A more promising syntactic subdomain where a measure using feature variables should outperform the measure using atomic variables is the more complex and more heterogeneous fronting subdomain. Spruit (2005) shows that measurements using atomic variables in the SAND fronting subdomain do not result in interpretable areas. A measure using feature variables may lead to a more homogeneous geographical distribution. This work will be reported on in future research.

Regression analyses were performed to correlate the syntactic measure using atomic variables with the measure using feature variables with respect to reflexives. A regression analysis using the same selection of 21 dialect locations as shown in Figure 3-4 results in a correlation coefficient of 0.93. A regression analysis using all 266 dialects leads to a correlation value of 0.92.<sup>11</sup> This means that there is a strong correlation between the syntactic measure using atomic variables and the measure using feature variables with respect to reflexives.

Figure 3-7 and Figure 3-8 show the geographical distances versus syntactic distances with respect to reflexives using atomic variables and feature variables, respectively. Using a linear function to describe the relation between geographical distance and syntactic variation with respect to reflexives results in relatively low correlation values of 0.47 and 0.38 using atomic variables and feature variables, respectively. A logarithmic function better describes the correlation between geographical and syntactic distance in both cases. However, the resulting correlation values of 0.53 and 0.48 are not much higher when atomic variables and feature variables are used, respectively. Furthermore, the measure using feature variables also results in a somewhat higher standard error value.

---

<sup>11</sup> No data is available with respect to reflexives for the dialect of Morbecque.

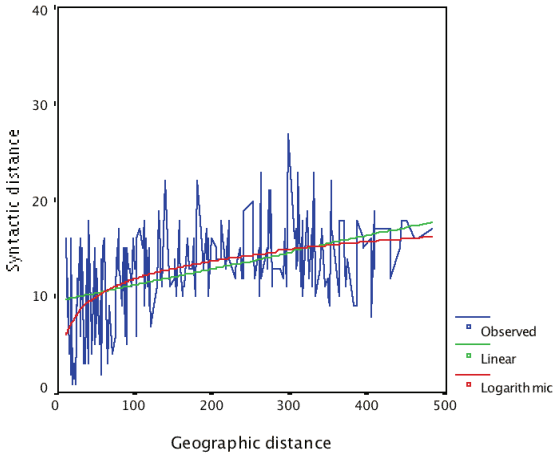


Figure 3-7: *Geographical distances versus syntactic distances with respect to reflexives using atomic variables.*

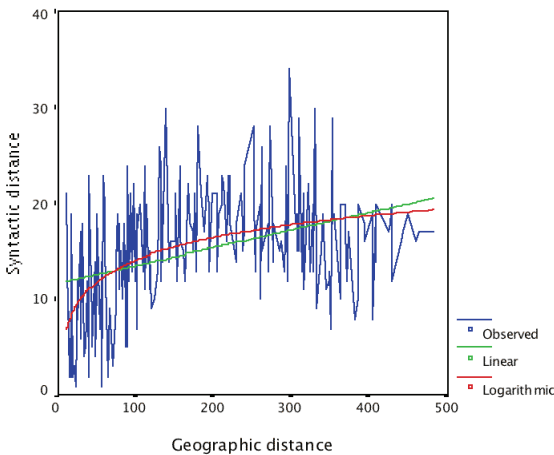


Figure 3-8: *Geographical distances versus syntactic distances with respect to reflexives using feature variables.*

All in all, the results based on a measure using either atomic variables or feature variables are quite similar. An explanation may be found in the shape of the regression curves shown in Figure 3-7 and Figure 3-8. Both regressions start from a relatively steep angle until the syntactic distance levels off to a fairly flat level in relation to the geographical distance, suggesting that measuring syntactic distances between distant dialect locations no longer reliably reflects linguistic dissimilarity. This assumption may be confirmed using the *local incoherence* validation method described in Nerbonne and Kleiweg (2007). Local incoher-

ence is a numerical probe to compare distance matrices with respect to the degree to which they reflect local geography faithfully. Lower local incoherence scores indicate that a given distance matrix better reflects local conditioning of dialect differences. Application of this method to the distance matrix based on atomic variables results in a local incoherence value of 10.3. The matrix based on feature variables results in a local incoherence value of 10.7. This means that the measure using atomic variables brings about slightly better results than the measure using feature variables, which confirms the results of the regression analysis.

### **3.9. Conclusions**

This first application of dialectometrical methods to purely syntactic data includes several notable highlights and directions for future research. Most significantly, this quantitative perspective on syntactic variation demonstrates that there is, in fact, geographical cohesion in syntactic variation. Furthermore, the classification of Dutch dialect varieties based on a syntactic measure using atomic variables highly resembles the classification based on subjective judgements on the Daan and Blok dialect map. This can be interpreted as a confirmation and validation of the syntactic measurement method. There also seem to be good overlaps between the objective classifications of Dutch dialect varieties based on syntactic and pronunciation differences, but more precise analysis is required. Finally, a measure using feature variables yields highly similar results with respect to syntactic variation in the reflexives domain. Even though these first results using feature variables do not directly increase accuracy of the syntactic measure, they do provide new and promising pathways to more accurately quantify syntactic variation. This includes differentiation between dissimilar variable pairs and the inclusion of the number of similarities as well as differences in the syntactic measure.

### **3.10. Future Research**

Future research will continue and extend the current work. First, feature variables will be formulated and annotated with respect to the remaining SAND domains, starting with the fronting subdomain. Second, statistical information such as variable frequency will be included for use in weighted similarity and dissimilarity measures. Third, the second and final part of the SAND will become available in 2006. The application of dialectometrical methods to the purely syntactic domains in SAND2 may lead to new insights as well. Fourth, statistical techniques will be applied to explore dependencies among syntactic variables. Finally, correlations between linguistic levels will be analysed in more detail.

### 3.11. References

- Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M. van der (eds), 2005. *Syntactic Atlas of the Netherlandic Dialects*, Volume 1. Amsterdam University Press, Amsterdam.
- Barbiers, S., Bennis, H., 2004. *Reflexieven in dialecten van het Nederlands. Chaos of structuur?*. In: Caluwe, J. de, Schutter, G. de, Devos, M. et al. (eds), *Schatbewaarder van de taal. Johan Taeldeman. Liber Amicorum*. Academia Press Gent en Vakgroep Nederlandse Taalkunde Universiteit Gent, Gent, 43–58.
- Bree, C. van, 1992. *The stability of language elements, in present-day eastern Standard-Dutch and eastern Dutch dialects*. In: Leuvensteijn, J. van en Berns, J., *Dialect and Standard language [...] in the English, Dutch, German and Norwegian language areas*, Amsterdam, 178–203.
- Bree, C. van, 1994. *The development of so-called Town Frisian*. In: Bakker, P. and Mous, M. (eds), *Mixed Languages. 15 Case Studies in Language Intertwining*, Studies in Language and Language Use, Volume 13, IFOTT Amsterdam, 69–82.
- Cornips, L., Jongenburger, W., 2001. *Elicitation techniques in a Dutch syntactic dialect atlas project*. In: Broekhuizen, H., Wouden, T. van der (eds), *Linguistics in the Netherlands*, 2001, John Benjamins, 53–63.
- Daan, J., Blok, D., 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Volume XXXVII, Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandische Uitgevers Maatschappij, Amsterdam.
- Goebel, H., 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Volume 157, Philosophisch-Historische Klasse Denkschriften. Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of Rase, W., Pudlatz, H..
- Heeringa, W., Nerbonne, J., 2001. *Dialect Areas and Dialect Continua*. In: Sankoff, D., Labov, W., Kroch, A. (eds), *Language Variation and Change*, Volume 13, 2001, 375–400.
- Heeringa, W., 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis Rijksuniversiteit Groningen, Groningen.
- Helke, M., 1970. *The Grammar of English Reflexives*. Doctoral Dissertation, MIT, Cambridge.
- Nerbonne, J., Kleiweg, P., 2007. *Toward a Dialectological Yardstick*. In: *Journal of Quantitative Linguistics*, Volume 14(2), Routledge, New York, 148–167.
- Postma, G., 1997. *Logical entailment and the possessive nature of reflexive pronouns*. In: Bennis, H., Pica, P., Rooryck, J. (eds), *Perspectives on Binding and Atomism*, Foris, Dordrecht, 295–322.
- Séguy, J., 1971. *La relation entre la distance spatiale et la distance lexicale*, *Revue de Linguistique Romane*, Volume 35, 335–357.
- Spruit, M., 2005. *Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited*. In: Doetjes, J., Weijer, J. van de (eds), *Linguistics in the Netherlands*, 2005, John Benjamins, Amsterdam, 179–190.
- Torgerson, W., 1952. *Multidimensional scaling: I. Theory and method*. *Psychometrika*, Volume 17, 401–419.

### 3.12. Appendix

The following two tables show the mapping from atomic variables to feature variables related to reciprocals and one-pronominalisation in the reflexives sub-domain as described in Section 3.7.

Table 3-7: Mapping from atomic variables (first column) to feature variables (first row) with respect to reciprocal pronouns.

	contrast <i>ander</i>	quantifier <i>me/malle</i>	quantifier <i>elk/enk/alle</i>	finite <i>één/een</i>	suffix –e <i>e(n)</i>	suffix –s <i>s</i>	composite <i>een-ander</i>
deendander	+			+			+
één			+	+			
eenaar				+			
eenander	+			+			
elkaar			+				
elkander	+		+				
enkander	+		+	+			
mallekaar		+	+				
mekaar		+					
mekaars		+				+	
mekander	+	+					
mekandere(n)	+	+			+		
mekanders	+	+				+	
mekare		+			+		

Table 3-8: Mapping from atomic variables (first column) to feature variables (first row) with respect to one-pronominalisation.

	animate	ellipsis	deletion
zo'n rare vrouw één	+	+	
zo'n ding één		+	
'n rare één			+





## 4. Associations among linguistic levels

“Correlating geographical versus pronunciational, lexical and syntactic distances”<sup>\*</sup>

*Spruit, M.R., Heeringa, W.J., Nerbonne, J., t.a. 2008. Associations among Linguistic Levels. Lingua, Special issue on Syntactic databases. Selected papers presented in the special session Comparing Aggregate Syntaxes, Digital Humanities conference, Paris, July 6, 2006.*

In this paper we measure the degrees of association among aggregate pronunciational, lexical and syntactic differences in 70 Dutch dialect varieties. First, we show that pronunciation is marginally more strongly associated with syntax than it is with lexis and that syntax and lexis are only weakly associated. Then, we check for the influence of geography as an underlying factor because geography is known to strongly correlate with each of the linguistic levels under investigation. We find that pronunciation and syntax are more strongly associated with geography than lexis is. Finally, we refine the results by accounting for the influence of geography as an underlying factor and show that the association between pronunciation and syntax turns out to be largely based on geography. Some influence between pronunciation and syntax remains but the association between pronunciation and lexis is stronger. There is virtually no association between syntax and lexis.

### 4.1. Introduction

The goal of this paper is to contribute to the understanding of the associations among linguistic levels by examining geographical distributions of linguistic microvariation. Investigations of linguistic variation in geographical space can not only illustrate patterns of variation at a certain point in time, but may also reflect residues of linguistic and cultural changes over historical time. This argument effectively interprets synchronic distributions as evidence of diachronic patterns of diffusion (Nerbonne and Heeringa, t.a. 2007). We study distributions of linguistic variation in the aggregate to compensate for the noisiness of individual distributions and to examine the data from more general perspectives in which we aggregate over many variables. We conduct this investigation at an

---

<sup>\*</sup> This paper was presented in the special session Comparing Aggregate Syntaxes at the Digital Humanities conference in Paris on July 6, 2006. It is based on joint research by the University of Groningen and the Meertens Instituut in Amsterdam. The Meertens Instituut is the national institute for research and documentation of Dutch language and culture. The Computational Linguistics department at the University of Groningen is known for its attention to quantitative linguistics and dialectometry. For three years now, these two research groups have been collaborating in the Determinants of Dialectal Variation project, NWO number 360-70-120, P.I. J. Nerbonne. More information is available on our project's website at <http://dialectometry.net>.

aggregate level in order to avoid the choice of a single individual variable, such as the pronunciation of /r/, which risks biasing its results based on the selection. The current study necessarily examines the linguistic levels under investigation on the basis of large collections of numerically interpreted data, because a robust, empirical foundation is required to analyse data from a more general perspective. The adopted quantitative methodology focuses on more general characteristics within and among linguistic levels because individual variables are only taken into account through their relationships with other variables. Metaphorically speaking, the current approach quantifies associations among “the linguistic forests behind the variable trees”.

We are now in a position to assess the dialectometrical distances among fairly many sites at three different linguistic levels: pronunciation, lexicon (or vocabulary) and syntax. Pronunciational differences mainly arise from linguistic variation at the phonetic level, but may also include variation at the phonological and morphological levels. We quantify lexical and syntactic differences at a nominal level using a frequency-weighted similarity measure introduced by Goebel (1982) and we measure pronunciational differences numerically using Levenshtein distance (Nerbonne et al., 1999; Heeringa, 2004). The novelty of this paper consists first in the opportunity to include syntax among the linguistic levels we analyse, and second, in its attention to potential, mutually structuring elements among the linguistic levels.

We suggest that the associations we attempt to detect are interesting first from a typological point of view, and second, from the point of view of identifying what influences linguistic variation. Addressing the second point first, we note that, although there are many candidate influences which might be affecting how languages vary, including e.g. settlement size, social class, sex, and educational level, only geography has proven its value in large-scale, quantitative studies (Nerbonne and Heeringa, t.a. 2007). We proceed here from the common assumption that there are no structural ties between lexical and nonlexical variables. In the present context, this means that we assume there is no linguistic reason to suspect correlation either between the pronunciational and lexical levels or between the syntactic and lexical levels. If we were to demonstrate significant correlations between lexical and nonlexical levels beyond those geography can explain, we would conclude that extralinguistic, non-geographical influences were at work. This should encourage the search for extralinguistic variables, but also suggest how important it might be.

The relation between phonology and syntax is more complicated, since it is easily conceivable that there might be *structural* constraints linking variation at these levels (see below). If phonology and syntax turn out to co-vary beyond the level explained by geography, this might reflect the influence of such structural, typological constraints. Of course, it might just as well reflect the influence of the same variables which account for the correlations between lexical

and non-lexical variables, so we shall need to interpret any correlation between phonology and syntax in light of the investigation between the lexical and non-lexical levels.

This paper is structured as follows. Section 2 formulates the two research questions addressed in this work. Section 3 describes the two data sources. Section 4 explains the two measurement procedures used to quantify linguistic differences. Section 5 presents colour maps of the Dutch dialect areas based on pronunciation, lexical and syntactic differences to provide a visual indication of the degrees of association. Section 6 analyses our distance measurements with respect to consistency to ensure that the results are reliable. Section 7 lists the exact degrees of association between pronunciation, lexis and syntax. Section 8 provides the degrees of association between geography and the linguistic levels under investigation. Section 9 refines the results in Section 7 by accounting for the influence of geography as an underlying, third factor. Section 10 recapitulates the main results. The paper concludes with a discussion and directions for future research in Section 11.

## **4.2. Research questions**

While most linguists would predict that vocabulary is more volatile than pronunciation and syntax and might predict that lexical choice should show little association with other linguistic levels, there have been predictions linking pronunciation with syntactic properties (Donegan and Stampe, 1983). Both pronunciation and syntax are highly structured systems, within which a single linguistic parameter might lead to a multitude of concrete and measurable effects.

We address two research questions in the present paper, the first of which is fairly straightforward:

- I. To what degree are aggregate pronunciation, lexical and syntactic distances associated with one another when measured among varieties of a single language? Particularly, are syntax and pronunciation more strongly associated with one another than either (taken separately) is associated with lexical distance?

To answer the questions above, it is sufficient to calculate correlation coefficients among the distance measurements for the three linguistic levels. This is a reasonable measure of the degree to which the three linguistic levels are associated.

However, it would be a mistake to interpret any such correlation as influence without checking for the influence of a third factor, especially since geography has already independently been shown to strongly correlate with each of the linguistic levels under investigation (Heeringa and Nerbonne, 2001; Cavalli-

Sforza and Wang, 1986; Spruit, 2006). Therefore, it is quite plausible that geography could influence each of the levels separately, leading to the impression of structural influence between them. We suggest that this should be regarded as a null hypothesis, i.e. that there is no influence among the various linguistic levels. This leads to the second research question we address in this paper:

- II. Is there evidence for influence among the linguistic levels, even once we control for the effect of geography? Particularly, do syntax and pronunciation more strongly influence one another than either—taken separately—influences or is influenced by lexical distance?

We attack these latter questions in multiple regression designs, checking for the effects of linguistic levels on one another once geography is included as an independent variable.

### 4.3. Data sources

This research is based on two Dutch dialectal data sources: the *Reeks Nederlandse Dialectatlassen* (RND; ‘Series of Dutch Dialect atlases’; Blancquaert and Peé, 1925-1982) and the first volume of the *Syntactische Atlas van de Nederlandse Dialecten* (SAND1; ‘Syntactic Atlas of the Dutch Dialects’; Barbiers et al., 2005). Both atlases describe Dutch dialects in the Netherlands, the Northern part of Belgium and a small north-western part of France. The RND data also include the north-eastern area of the Belgian province Luik and the German county Bentheim.

The RND is a 16-volume series of Dutch dialect atlases which were edited by Blancquaert and Peé. The first volume was compiled by Blancquaert and appeared in 1925. The final volume was published in 1982 and was edited by Peé. The RND contains translations and phonetic transcriptions of 139 sentences in 1,956 Dutch dialects. The data were recorded between 1922 and 1975. We use a digitised selection of 125 words from 360 dialects. Figure 4-1 shows the geographical distribution of the RND locations. The selected words represent all vowels and consonants and are used to measure both pronunciations and lexical distances. Heeringa (2001) discusses the selection of words and dialect locations from the RND in detail.<sup>1</sup> The next section provides several examples of words and transcriptions in the RND.

SAND1 contains 145 geographical distribution maps of individual syntactic variables in 267 Dutch dialects. Figure 4-2 shows the geographical distribution of the SAND locations. It covers syntactic variation related to the left periphery of the clause and pronominal reference. This includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject

---

<sup>1</sup> The RND data are publicly available at <http://www.let.rug.nl/~heeringa/dialectology/atlas/rnd>.

Table 4-1: Map 14b in SAND1 shows seven syntactic variables in the context of complementiser of comparative if-clause.

Context:	Complementiser of comparative if-clause
Variables:	{ of, *of dat, dat, *as/of + V2, at, as, et }
Examples:	‘t lijkt wel <u>of dat</u> er iemand in de tuin staat. ‘t lijkt wel <u>of</u> er staat iemand in de tuin. ‘it looks [affirm.] if that there stands someone in the garden stands’ “‘It looks as if there is someone in the garden.’”

Table 4-2: Map 54a in SAND1 shows four syntactic variables in the context of subject doubling 2 singular.

Context:	Subject doubling 2 singular
Variables:	{ V <sub>FINITE</sub> __, *__ V <sub>FINITE</sub> __, C __, *C <sub>COMPARATIVE</sub> __ }
Examples:	<u>Ge</u> gelooft <u>gij</u> zeker niet dat hij sterker is as <u>gij</u> . <u>Ge</u> gelooft <u>gij</u> zeker niet dat hij sterker is as <u>-ge</u> <u>gij</u> . ‘you <sub>weak</sub> believe you <sub>strong</sub> certainly not that he stronger is than you <sub>weak</sub> you <sub>strong</sub> ’ “‘You do not seem to believe that he is stronger than you.’”

cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. SAND1 contains 106 syntactic contexts.<sup>2</sup> Table 4-1 and Table 4-2 provide two examples of variation in syntactic contexts as described in SAND1.<sup>3</sup> The second and final volume of the SAND is due to appear in 2007 and will describe syntactic variation in Dutch dialects with respect to verbal clusters, negation and quantification.<sup>4</sup>

As stated above, the RND data are used to measure both pronunciational and lexical distances. The SAND1 data are used to measure syntactic distances. However, we can only relate the measurements obtained from these two data sources if the results are based on exactly the same set of dialect locations. We cannot assume that two geographically close locations are also closely related on all three linguistic levels. Therefore, we only use the intersection of the 360 RND dialects and the 267 SAND1 dialects.<sup>5</sup> As shown in Figure 4-3, the result-

<sup>2</sup> The number of available syntactic contexts is lower than the number of geographical maps because SAND1 contains numerous correlation maps which show syntactic variables from different perspectives. Also, some syntactic contexts are presented using multiple maps.

<sup>3</sup> Spruit (2006) provides more examples of syntactic contexts in SAND1.

<sup>4</sup> The SAND data are accessible from the Dynamic Syntactic Atlas of the Dutch dialects (Dyna-SAND) at <http://www.meertens.knaw.nl/sand>.

<sup>5</sup> The Dutch language area under investigation, as shown in Figure 3, borders on the North Sea in the North and in the West. Germany lies along the Eastern border. The south-western border

ing 70 common dialect varieties in the Netherlands and the Northern part of Belgium are not perfectly geographically distributed.<sup>6</sup> The north-eastern and southern areas are overrepresented and the western and central areas are somewhat underrepresented. However, these underrepresented areas are known to have relatively fewer differentiating characteristics than the overrepresented areas. Therefore, we expect the intersection of the RND and SAND1 dialects to adequately represent the language variation spectrum in the Dutch dialect area for our purposes.

#### 4.4. Distance measures

The dialect differences *within* each linguistic level need to be measured before the associations *between* the linguistic levels can be quantified. We use the *Levenshtein* distance and the *gewichteter Identitätswert* method to measure the dialect differences within each linguistic level.

The Levenshtein distance is used to measure pronunciation differences. It was first described in Levenshtein (1966). Generally speaking, it is a string edit distance measure which calculates the minimally required steps to change one sequence of symbols to another sequence of symbols. Sankoff and Kruskal (1999) discuss a broad range of applications of the Levenshtein distance. Contrary to other well-known distance measures such as the Hamming, Manhattan and Euclidean distance measures, the Levenshtein distance measure is able to quantify the differences between sequences of different lengths. The algorithm is based on the optimal alignment between two sequences of symbols and uses one of the operations *insert*, *delete* or *substitute* at each symbol comparison. Kessler (1995) first applied the Levenshtein distance to measure differences between phonetic transcriptions of word pronunciations in Irish Gaelic dialects. Heeringa (2004) refines the Levenshtein algorithm in several ways to more accurately measure pronunciation differences in Dutch dialects. It describes the enhanced version of the algorithm we use in this work in great detail on pages 79-119. The refinement uses comparisons of spectrograms of the component sounds to differentiate between dissimilar sounds acoustically.

---

of the province West-Vlaanderen lies adjacent to France. The remaining southern border follows the Dutch-French language border in Belgium.

<sup>6</sup> These are the 70 common dialect varieties in alphabetical order, as shown in Figure 3: Aalst, Aalten, Almelo, Anjum, Appelscha, Arendonk, Bakkeveen, Bellingwolde, Bergum, Beveren, Boutersem, Bree, Brugge, Coevorden, Druten, Eibergen, Emmen, Ferwerd, Fijnaart, Gemert, Gent, Geraardsbergen, Gistel, Goes, Gramsbergen, Groenlo, Groesbeek, Groningen, Haaksbergen, Heerenveen, Hindeloopen, Hollum, Houthalen, Huizen, Humbeek, Kamperhout, Kerkrade, Kollum, Kortrijk, Lauw, Lemmer, Mechelen, Midsland, Oldemarkt, Onstwedde, Oostende, Ootmarsum, Opperdoes, Ossendrecht, Overijse, Roeselare, Ronse, Roswinkel, Schiermonnikoog, Spakenburg, Staphorst, Steenbergen, Steenwijk, Tegelen, Tienen, Urk, Utrecht, Vaals, Veurne, Vriezenveen, Waregem, Warffum, West-Terschelling, Zierikzee, Zundert.

Table 4-3: String alignment and Levenshtein distance calculation between two pronunciations of the Dutch word *hart* 'heart'.

Alignment	[ <i>hart</i> ]	[ <i>ærtə</i> ]	Edit operation	Cost
1	h		delete h	1
2	a	æ	substitute æ for a	1
3	r	r		0
4	t	t		0
5		ə	insert ə	1
				—
Levenshtein distance between [ <i>hart</i> ] and [ <i>ærtə</i> ] =				$3 / 5 = 0.6$

Table 4-3 illustrates the string alignment principle and the Levenshtein distance calculation between two pronunciations of the Dutch word *hart* 'heart'. The example does not take into account the refinements mentioned above as to more clearly illustrate the general principle. The word *hart* is pronounced as [*hart*] in Haarlem, whereas in Brugge people say [*ærtə*]. First, the Levenshtein algorithm aligns the two pronunciations optimally. Then, the number of edit operations are counted which are required to change the first pronunciation into the second one. Finally, the number of operations is divided by the string alignment length to obtain the normalised Levenshtein distance between these two pronunciations of the word *hart*, which, in this case, is  $(1+1+1 / 5 = ) 0.6$ . The aggregate pronuncional distance between Haarlem and Brugge is calculated by accumulating all pronuncional distances between the two dialects and dividing the aggregate distance by the total number of pronuncional comparisons.

Lexical and syntactic distances are measured at a nominal level using the *gewichteter Identitätswert* (GIW).<sup>7</sup> This is a frequency-weighted similarity value which was introduced in dialectometry by Goebel (1984). The GIW method counts infrequent words more heavily than frequent ones. This opposes the tendency in several areas of quantitative linguistics that very infrequent words should be treated as noise, unreliable evidence of linguistic structure (Nerbonne and Kleiweg, 2007). We use the inverse of Goebel's original similarity measure to obtain GIW distance values by subtracting the similarity value from 1.

The RND data require additional preparation before they can be used to measure lexical distances. This step arises because the RND does not contain lexical identity information between word pronunciations. Therefore, we manually

<sup>7</sup> The GIW method measures differences between variable pairs at a nominal level. This means that two variables are either equal or unequal. The Levenshtein distance is a numerical measure which allows differentiation between variable pairs in terms of degrees of similarity.

determined the represented lexemes for each of the 125 word pronunciations from a layman's perspective. We did not analyse the word pronunciations from an etymological point of view. The following example of the lexical concept *zijn* 'to be' illustrates the lexeme identification procedure. The word pronunciations [bɪŋ], [bɪnt] and [bɛnə] are considered to be forms of a single lexeme, which however differs from the single lexeme instantiated in [zɛm], [zɪnt] and [zan], even though the pronunciations [bɪnt] and [zɪnt] seem very similar. Also, inflectional variants do not play a role in the context of this procedure. For example, the words [bɔːmkə] and [bɔːmpjə] are identified as two pronunciations of the lexeme *boompje* 'little tree'. Both words are morphologically derived from the root *boom* 'tree'. Heringa et al. (2007) contains more information regarding the lexeme identification procedure. In contrast to the above, SAND1 does not require additional annotation of the data. It already presents each syntactic variable within its syntactic context.<sup>8</sup>

Table 4-4 illustrates the GIW method as a measure of lexical similarity between the dialects of Middelstum and Ommen. As already noted, we employ the inverse of the original similarity measure as illustrated in Table 4-4 to obtain GIW distance values by subtracting each similarity value from 1. The distance calculations are omitted from Table 4-4 to enhance readability. The described procedure is identical when syntactic differences are measured. The example shows that the two dialects use the same lexemes for the concepts *vriend* 'friend' and *schip* 'ship': *kameraad* 'comrade' and *schip* 'ship', respectively. However, the two dialects use a different lexeme to reference the concept *duwen* 'to push'. In Middelstum people say *stoten* 'to thrust', whereas in Ommen people use *drukken* 'to press'.

Table 4-4: *Weighted similarity calculation between two dialects based on word choices for the three concepts of vriend 'friend', schip 'ship' and duwen 'to push' using the gewichteter Identitätswert (GIW) measure.*

concept	Middelstum	Ommen	matches	comparisons	GIW
vriend	kamərʊt	kamərɔ:t	1 - (140 / 354)		= 0.60
schip	sxɪp	sxɪp	1 - (353 / 360)		= 0.02
duwen	støˈŋ	drykɔ	-	-	= 0

---


$$\text{Weighted similarity between Middelstum and Ommen} = 0.62 / 3 = 0.21$$

---

<sup>8</sup> Roughly speaking, in SAND1 each geographical distribution map represents a syntactic context and each map symbol represents a syntactic variable. A map symbol, by definition, can only be shown on a map. Therefore, SAND1 variables are always presented within context.



This information is used to calculate the lexical distance between the two dialects. First, the lexeme *kameraad* ‘comrade’ references the concept *vriend* ‘friend’ in 140 dialects. In 214 dialects a different lexeme is used instead. This results in a weighted similarity of  $(1 - 140 / 354 =) 0.6$  and a complementary GIW distance value of  $(1 - 0.6 =) 0.4$ . Unfortunately, in six dialects no data was available for this concept. We ignore missing concepts because there is obviously nothing to measure.<sup>9</sup> Next, the concept *schip* ‘ship’ is nearly always referenced by the same lexeme. Therefore, the GIW method considers this information to be of little help in quantifying the linguistic variation between the two dialects. The weighted similarity of  $(1 - 353 / 360 =) 0.02$  and the corresponding GIW distance of 0.98 reflect this consideration appropriately. The different similarity weights (0.6 versus 0.02) assigned to the first two concepts in Table 4-4 demonstrate that similarity weighting in the GIW method *emphasizes* rather than ignores infrequently occurring words. Finally, the third concept *duwen* ‘to push’ in Table 4-4 is realised with different lexemes in the two dialects. The GIW method always assigns different lexemes a similarity value of 0.0 to designate the dissimilarity between the lexemes. This is equivalent to a maximum GIW distance of 1.0. The lexical GIW distance measurements between the dialects of Middelstum and Ommen based on the three concepts shown in Table 4-4 result in a weighted similarity of  $(0.62 / 3 =) 0.21$ , which this work translates into the corresponding lexical GIW distance value of  $(1 - 0.21 =) 0.79$ .

#### 4.5. Dutch dialect area perspectives

We present colour maps of the Dutch dialect areas based on pronunciation, lexical and syntactic differences in pairwise comparisons to provide a general impression of the associations between the pronunciation, lexical and syntactic levels before we calculate the exact degrees of association in Section 7. We first present the De Schutter (1994) map and the Daan and Blok (1969) map of the Dutch dialect areas in Figure 4-4 and Figure 4-5 as external points of reference. These two non-computational dialect maps are based on perception and expert opinion, respectively.

Our dialect colour maps employ Multidimensional scaling (MDS) to visualise the pronunciation, lexical and syntactic variation in the Dutch language area. This statistical technique was first described in Torgerson (1952). We apply the MDS procedure to display the general dialect relationships as faithfully as possible in one three-dimensional, full-colour picture. The procedure to visualise the distance measurements consists of the following three steps. First, each dialect’s distance relationships to all other dialects are reduced to coordinates in a three-dimensional space using the three most important dimensions arising

---

<sup>9</sup> The lexical distance measurements are based on GIW comparisons between 103 and 125 concepts, with 121 concepts on average.

from the MDS analysis. These coordinates optimally represent the original dialect distance relationships. They do not directly correspond to actual dialect distances anymore. Second, the three-dimensional coordinates are used as values between light and dark of the three colour components red, green and blue. This effectively means that a dialect's unique set of characteristics is translated into one unique composite colour. Neighbouring dialects have corresponding colours if they are also linguistically close to each other. They are progressively assigned less related colours as they are less related linguistically. Third, the dialect points on the maps are blown up to small areas until they border each other and there is no uncoloured space left. This space partitioning technique uses the well-known Delaunay triangulation to obtain a pattern of Voronoi polygons.<sup>10</sup> The final result is that a colour continuum arises if there is a perfect relation between geographical distance and linguistic distance, whereas a mosaic-like map results when this relation is not strong. Heeringa (2004:156-163) discusses the technical details of the MDS technique and the Delaunay triangulation in detail from a linguistic perspective.

We need to ensure that the MDS maps visualise the linguistic variation accurately. The MDS procedure calculates a correlation coefficient to indicate the amount of linguistic variance which is represented in the first three dimensions of the MDS solution and, therefore, in the MDS map colours. The correlation coefficients between the dialect distance relationships and the MDS coordinate distance relationships are 0.94, 0.74 and 0.89 at the pronunciation, lexical and syntactic levels, respectively. The coefficients are also shown as *r*-values in Figure 4-6 to Figure 4-8. In most applications correlations below 0.8 tend to be too inaccurate to be interpreted meaningfully, whereas coefficients between 0.9 and 1 are generally considered to be high. Norušis (1997) notably defines  $r^2 = 0.6$  (i.e.  $r = 0.77$ ) as the minimum acceptable correlation in the context of the MDS procedure. However, the exact correlation threshold values likely vary within each specific research context. All in all, we conclude that the dialect colour maps in Figure 4-6 and Figure 4-8 accurately represent the original distance measurements at the pronunciation and syntactic levels. The lexical MDS map in Figure 4-7 also represents the original lexical variance to an acceptable extent for our exploratory purposes, but it should be interpreted more cautiously.

The Daan and Blok dialect map in Figure 4-5 shows the classification of the Dutch dialect varieties based on subjective judgements from local speakers, local experts and the map designers themselves. The Netherlandic dialect area borders were derived from a written survey from 1939 among 1,500 local dialect speakers. The Belgian part was mostly based on the perception of local

---

<sup>10</sup> Goebel (1982) introduces the use of Voronoi tiling, sketched here, to illustrate the results of dialectometrical analyses. Alternatively, an interpolation procedure could be applied to colour the space between dialect locations.

dialect experts. The methodology and results of the map creation procedure are discussed in Heeringa (2004:12-13), among others. Our dialect colour maps follow Daan and Blok (1969) in the assignment of the colour blue to the Frisian area in the central north and the colour green to the north-eastern Lower Saxon region to simplify comparisons between the dialect maps.

Spruit (2005) provides a visual comparison between a syntactic MDS map based on Hamming distances and the perceptual Daan and Blok map in Figure 4-5. The syntactic MDS map in Spruit (2005) is very similar to the syntactic MDS map based on GIW distances shown in Figure 4-8. Therefore, the dialect maps in Figure 4-5 and Figure 4-8 are also remarkably similar. Interestingly, “[...] the Belgian dialect classification on the Daan and Blok map based on more objective expert judgements corresponds to a higher degree with the classification based on the objective syntactic measure than with the Netherlandic dialect classification based on intuitive judgements” (Spruit 2005:189). Among other suggestions, the work mentions the role of prejudice in perception and sensitivity to pronunciation differences as possible explanations. Heeringa (2004:230-233) discusses the similarities and differences between the perceptual Daan and Blok map in Figure 4-5 and the pronunciational MDS map shown in Figure 4-6.

The De Schutter dialect map in Figure 4-4 is a simplified expert consensus map of the Dutch dialect areas. It is heavily based on the Daan and Blok map but also relies on several other dialect maps.<sup>11</sup> The author considers this map to reflect the general opinion of traditional dialectologists at the end of the 20<sup>th</sup> century. It shows the six main Dutch dialect areas: the central-northern Frisian, north-eastern, central-western, south-western, central-southern and south-eastern dialects.

The two maps in Figure 4-6 and Figure 4-7 visualise the variation in the Dutch language area with respect to pronunciation differences and lexical differences, respectively. We can expect a substantial correlation between the two linguistic levels based on the visual correspondences between the two maps. For example, the central-northern Frisian area in blue stands out very prominently on both maps. The most prominent difference is arguably the clear-cut northern border on the lexical map of the central-south area in pink. This border can not be made out on the pronunciational map.

Figure 4-8 shows the variation in Dutch dialects with respect to syntactic variation. When we visually compare this map with the lexical map shown in Figure 4-7, we can already be quite certain that the degree of association between the syntactic and lexical levels will be lower than the correlation between the pronunciational and lexical levels, as discussed in the previous paragraph. For example,

---

<sup>11</sup> The De Schutter map based on expert consensus also takes the Dutch dialect area classifications in Weijnen (1958) and Goossens (1977) into account.



Figure 4-1: Distribution of the 360 Dutch dialects in the RND atlas.



Figure 4-2: Distribution of the 267 Dutch dialects in the SAND atlas.



Figure 4-3: Distribution of the 70 common Dutch dialects in the RND and SAND atlases with the relevant province names.



Figure 4-4: Expert consensus map of the Dutch dialects (translated from De Schutter, 1994).

the appearance of the Frisian area in blue on the syntactic map is not nearly as prominent as on the lexical map. Also, the south-east area on the syntactic map in light-blue is quite prominently present, whereas this area can hardly be made out on the lexical map.

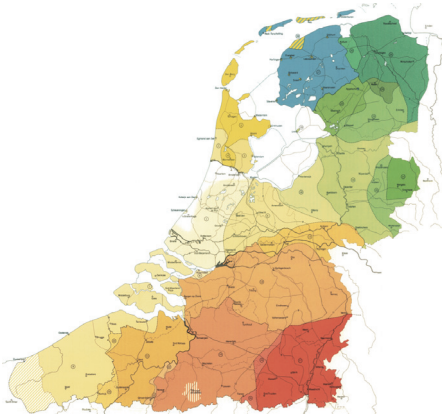


Figure 4-5: Perceptual map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok, 1969).



Figure 4-6: Pronunciational MDS map of the Dutch dialects based on Levenshtein distances ( $r = 0.94$ ).



Figure 4-7: Lexical MDS map of the Dutch dialects based on GIW distances ( $r = 0.74$ ).



Figure 4-8: Syntactic MDS map of the Dutch dialects based on GIW distances ( $r = 0.89$ ).

The final pair of maps compares pronunciational (Figure 4-6) and syntactic (Figure 4-8) differences in Dutch dialects. After a single glance at these two maps we can already speculate that the correlation between these two linguistic levels will be higher than the correlation between the lexical and syntactic levels in Figure 4-7 and Figure 4-8, respectively. However, it is uncertain whether the correlation between the pronunciational and syntactic levels is also stronger than the correlation between the pronunciational and lexical levels in Figure 4-6 and Figure 4-7, respectively. For example, these two maps differ in the degree of separation with respect to the blue Frisian area, but they correspond to a

higher degree in the southern areas than the pronunciational versus lexical maps correspond with each other. Therefore, we need to *calculate* the degree of association among these three linguistic levels to answer this question satisfactorily.

#### 4.6. Consistency

We want to ensure that our distance measurements are consistent—we want to know that our results are reliable. Therefore, we use Cronbach's alpha to measure the minimum reliability of our distance measurements when applied to our data sources. Cronbach's alpha was first described in Cronbach (1951). It is a coefficient of consistency and can be described as a function of the number of linguistic variables ( $n_{var}$ ) and the average inter-correlation value among the variables ( $r$ ), i.e. the mean of all the familiar Pearson product-moment correlation coefficients, given in Figure 4-10.<sup>12</sup> Its values range between zero and one. Higher values indicate more reliability. As a rule of thumb, values higher than 0.7 are considered sufficient to obtain consistent results in social sciences (Nunnally, 1978). The Cronbach's alpha formula is shown in Figure 4-9. The formula to obtain the average inter-correlation value among the variables ( $r$ ) is listed in Figure 4-10.

Table 4-5 presents the Cronbach's alpha values which indicate the reliability of our measurement results at the pronunciational, lexical and syntactic levels. Based on the Cronbach's alpha value of 0.97 we can conclude that the Levenshtein analysis of the pronunciational data is very reliable. The GIW analysis of the syntactic data results in a value of 0.94 which also indicates very consistent results. The GIW analysis of the lexical data brings about a coefficient of consistency of 0.75, which is acceptable. It does indicate, however, that the analysis of the lexical data may be less reliable than the analyses at the pronunciational and syntactic levels.

$$\alpha = \frac{n_{var} \times r}{1 + (n_{var} - 1) \times r}$$

Figure 4-9: Cronbach's alpha ( $\alpha$ ) is a function of the number of linguistic variables ( $n_{var}$ ) and the average inter-correlation value among the variables ( $r$ ).

$$r = \frac{\sum_{i=2}^{n_{var}} \sum_{j=1}^{i-1} r(\text{var}_i, \text{var}_j)}{n_{var} \times (n_{var} - 1) / 2}$$

Figure 4-10: The average inter-correlation value ( $r$ ) is based on all Pearson's correlation coefficients between each pair of variables  $r(\text{var}_i, \text{var}_j)$ .

<sup>12</sup> The Pearson product-moment correlation coefficient (PMCC) is the most commonly used method of computing a correlation coefficient between variables that are linearly related.

Table 4-5: Reliability coefficients ( $\alpha$ ) of our measurement results at the pronunciation, lexical and syntactic levels.

Linguistic level	Number of variables ( $n_{var}$ )	Cronbach's alpha ( $\alpha$ )
Pronunciation	125	0.97
Lexis	107	0.75
Syntax	106	0.94

Finally, it may be helpful to explicitly point out the interpretational difference between Cronbach's alpha coefficients and MDS correlation coefficients which were described in the previous section. Cronbach's alpha coefficients estimate how well the measurement results of an analysed dataset, representing the variation within the linguistic level, can be expected to capture the variation within the entire linguistic domain. Simplified, it measures the level of reliability of the results. In our research context the MDS correlation coefficients indicate the amount of linguistic variance which is represented in the first three dimensions of the MDS solution. Simplified, it measures the level of accuracy of the scaling procedure.

#### 4.7. Correlations between linguistic levels

Table 4-6 answers the first of the two research questions central to this paper, as stated in Section 2: to what degree are aggregate pronunciation, lexical and syntactic distances associated with one another, when measured among varieties of a single language? We have calculated the Pearson product-moment correlation coefficients among the distance measurements for the three linguistic levels as a measure of the degree to which the three linguistic levels are associated. The results show that pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and that syntax is much more strongly associated with pronunciation (42%) than with lexis (25%).

Table 4-6: Associations between aggregate pronunciation, lexical and syntactic distances.

Linguistic level 1	Linguistic level 2	Correlation ( $r$ )	Explained variance ( $r^2 * 100$ )
Pronunciation	Lexis	0.617	38%
Lexis	Syntax	0.496	25%
Syntax	Pronunciation	0.648	42%

The results below are based on the 70 common varieties as described in Section 3. The pronunciation differences were measured using the Levenshtein distance and the GIW method was applied to measure the variation at the lexical and syntactic levels. The percentages in Table 4-6 indicate the amount of variation at the first linguistic level which can be explained with the amount of varia-

tion at the second linguistic level. All correlation coefficients are significant at the 0.001 level. In order to not confound significance calculations between distance tables, the significance levels of the correlation coefficients were calculated using the Mantel test (Mantel, 1967).

The Mantel test calculates the significance levels of correlation coefficients between distance tables while taking into account the structured, interdependent nature of distance matrices. The null hypothesis in this asymptotic test states that there is no correlation between the symmetrical dialect distances in the two matrices. In other words, the test assumes that changes in dialect distances at the first linguistic level do not influence the dialect relationships at the second linguistic level. The hypothesis is evaluated by randomly reallocating the order of elements in the first matrix many times, and recalculating the correlation between the permuted first matrix and the original second matrix after each permutation. The significance of the observed correlation results from the proportion of the permutations that lead to a higher correlation than the actual coefficient. With a significance level of  $\alpha = 0.05$  the number of repetitions should be equal to about 1,000 (Manly, 1997). This means that less than five percent of thousand permuted matrix correlations may yield higher coefficients than the correlation coefficient between the original matrices. The reasoning is that if the null hypothesis—there is no correlation between the two matrices—is correct, then the permuted matrix should be equally likely to produce a larger or a smaller correlation coefficient.

With this information we can also answer the subquestion of our first research question: are syntax and pronunciation more strongly associated with one another than either is associated with lexical distance? To our surprise, the expectations we laid out in Section 2 were not decisively met. Although syntax is clearly more strongly associated with pronunciation ( $r = 0.684$ ) than with lexis ( $r = 0.496$ ), the syntax-pronunciation association ( $r = 0.648$ ) is not much stronger than the lexis-pronunciation connection ( $r = 0.617$ ). At this point we can only speculate about these outcomes. We already pointed out that the Cronbach's alpha value for the lexical analysis is relatively low. This leaves room for less reliable results. Also, we already acknowledged that the pronunciational data includes subphonological variation. It might be the case that variation at the phonetic and morphological sublevels is distributed in different patterns than purely phonological variation. This could reduce the expected correlation with the syntactic data. However, we should not draw any conclusions before having checked the correlations for the influence of a third, underlying factor: geography.



#### 4.8. Linguistic levels correlated with geography

Geography has independently been shown to correlate strongly with each of the three linguistic levels under investigation. Heeringa and Nerbonne (2001) examined the degrees of association between geographical and pronunciational distances in Dutch dialects. Cavalli-Sforza and Wang (1986) related geographical distances with lexical similarities in a chain of Micronesian islands. The correlation between geographical and syntactic distances in Dutch dialects was analysed in Spruit (2006). In this study we present the scatterplots and correlation values of pronunciational Levenshtein distances versus geographical distances in Figure 4-11, lexical GIW distances versus geographical distances in Figure 4-12 and syntactic GIW distances versus geographical distances in Figure 4-13. All results are based on the 70 common varieties as described in Section 3. The scatterplots show the associations between each of the three linguistic levels as dependent variables on the Y-axes and geography as the independent variable on the X-axis.

The geographical distances have been calculated using the *ll2dst* programme, which is part of the freely available dialectometry software package RuG/L04. The programme takes longitude-latitude coordinates to calculate the corresponding geographical distances in kilometres ‘as the crow flies’. The algorithm assumes that the earth is a perfect sphere and that it has a circumference of 40,000 kilometres. Although neither of these two assumptions is entirely correct, it should not noticeably affect the accuracy of our distance calculations. The Dutch language area only covers a very small surface of the earth’s sphere. Therefore, the Dutch area surface remains relatively flat and the distance calculations remain accurate.<sup>13</sup>

The current operationalisation of the factor geography as Euclidean distances between longitude-latitude coordinates is an acceptable approximation of geographical distance in the case of the Dutch language area under investigation. However, a more refined measure of geographical distance may be required in situations where geographical barriers may influence the chance of social contact considerably. Gooskens (2004) notably illustrates the effect of geography on dialect variation in Norway, where the central mountain range prevented direct travel until recently. In Norway travel time turns out to be a much better predictor of linguistic distance than distance ‘as the crow flies’. Of course, there are no mountain ranges, dry deserts, tropical forests or other types of inhospitable geographical barriers within the Dutch language area. Van Gemert (2002) examines the influence of water barriers such as lakes and rivers in the Netherlands on pronunciational distances between dialects. Contrary to its expectations, however, it concludes that travelling costs between dialects never corre-

---

<sup>13</sup> The *ll2dst* manual at <http://www.let.rug.nl/~kleiweg/L04/Manuals/ll2dst.html> contains more information on this software programme.

late to a higher degree with pronunciational variation than geographical distances ‘as the crow flies’. The remainder of this work, therefore, feels confident in the application of distances ‘as the crow flies’ as an adequate operationalisation of geography.

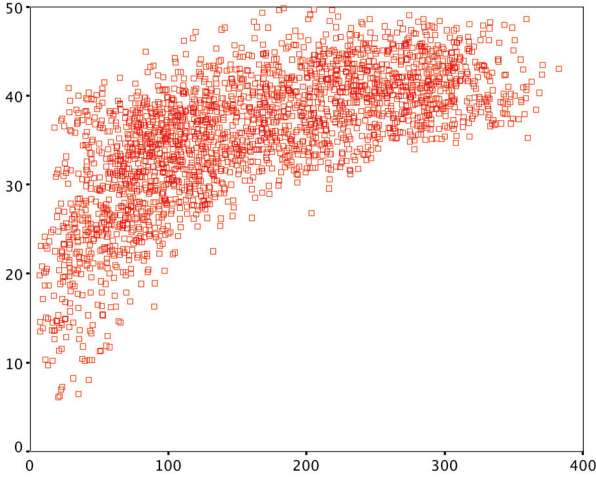


Figure 4-11: This scatterplot shows the relation between pronunciational Levenshtein distances on the Y-axis and geographical distances on the X-axis.

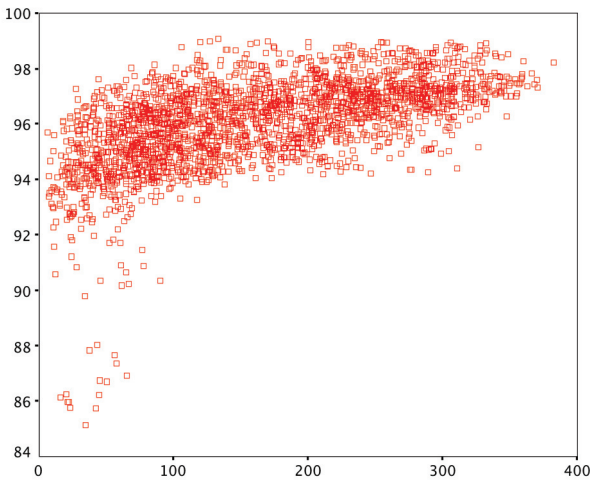


Figure 4-12: This scatterplot shows the relation between lexical GIW distances on the Y-axis and geographical distances on the X-axis.

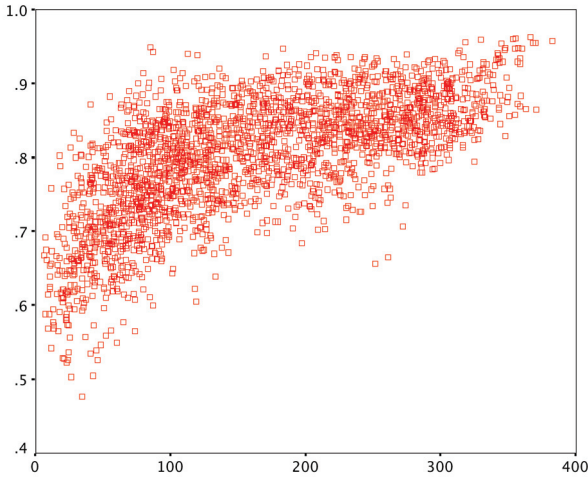


Figure 4-13: This scatterplot shows the relation between syntactic GIW distances on the Y-axis and geographical distances on the X-axis.

Table 4-7 shows the degrees of association between each linguistic level versus geography. The results clearly demonstrate that linguistic differences at the pronunciational and syntactic levels are more strongly associated with geographical distances (47% and 45%, respectively) than with variation at the lexical level (33%). All Pearson correlation coefficients are significant at the 0.001 level. The percentages in the right column are based on  $r^2$  values, which indicate the amount of variation at the specified linguistic level which can be explained with geographical distance. The results confirm the fundamental postulate in dialectology that language varieties are structured geographically (Nerbonne and Kleiweg, 2007).

Table 4-7: Correlations between geographical distances and pronunciational, lexical and syntactic distances.

Linguistic level	Correlation ( $r$ )	Explained variance ( $r^2 * 100$ )
Pronunciation	0.685	47%
Lexis	0.575	33%
Syntax	0.669	45%

#### 4.9. Linguistic correlations without the influence of geography

Section 7 presented the degrees of association among aggregate pronunciational, lexical and syntactic distances. However, in Section 8 we found that ge-

ography influences each of the three linguistic levels separately. Therefore, we need to refine the results in Section 7 by accounting for the structural influence of geography as an underlying, third factor. Based on the strong correlations between geography and each linguistic level separately, as shown in Section 8, we cannot assume that there is influence among the various linguistic levels. However, we can test for this.

The following three steps describe the procedure to calculate the correlation between two linguistic levels without geography as an influencing factor. This example takes pronunciation variation as the first linguistic level and lexical variation as the second linguistic level. First, we perform a regression analysis between the pronunciational distances and the geographical distances. This results in the pronunciational residuals. Residuals are those parts of the data which the regression model does not explain. Second, we likewise perform a regression analysis between the lexical distances and the geographical distances, which results in the lexical residuals. Third, we run a regression analysis between the pronunciational residuals and the lexical residuals which we obtained in steps one and two. This provides the correlation coefficient between pronunciational distances and lexical distances without the influence of geographical distances.

We repeat this procedure to calculate the correlation between the lexical and syntactic levels and the correlation between the syntactic and pronunciational levels. The results are presented in Table 4-8. Again, the results are based on the 70 common varieties as described in Section 3. The pronunciational differences were measured using the Levenshtein distance and the GIW method was applied to measure the variation at the lexical and syntactic levels. All correlation coefficients are significant at the 0.001 level using the Mantel test. Section 7 already explained why the significance levels of the calculated correlations between the linguistic levels are reliable, even when applied to the structured, interdependent data of distance matrices. The percentages in Table 4-8 indicate the amount of variation at the first linguistic level which can be explained with the amount of variation at the second linguistic level.

*Table 4-8: Associations between aggregate pronunciational, lexical and syntactic distances controlling for the influence of geography as an underlying factor.*

<i>Linguistic level 1</i>	<i>Linguistic level 2</i>	<i>Correlation (r)</i>	<i>Explained variance (r<sup>2</sup> * 100)</i>
Pronunciation	Lexis	0.374	14%
Lexis	Syntax	0.183	3%
Syntax	Pronunciation	0.350	12%

Table 4-9: The percentage of the correlation attributable to geography.

Linguistic level 1	Linguistic level 2	Geographical influence
Pronunciation	Lexis	39 %
Lexis	Syntax	63 %
Syntax	Pronunciation	46 %

With the degrees of association in Table 4-8 we can answer the second of our two research questions: is there evidence for influence among the linguistic levels, even once we control for the effect of geography? The answer is that some influence between pronunciation and syntax (12%) remains, although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (merely 3%).

Table 4-9 presents the influence of geography as a factor of influence underlying the associations between aggregate pronunciation, lexical and syntactic distances. Figure 4-14 shows the formula to calculate the influence of geography underlying the associations between the linguistic levels.

The formula in Figure 4-14 takes the correlation ( $r$ ) values from Table 4-6 and Table 4-8, respectively. Table 4-9 evidently shows the substantial influence of geography as a factor of influence underlying the associations between the linguistic levels. The degree of association between pronunciation and lexical distances turns out to be based on geography as an underlying factor for no less than 39%.<sup>14</sup> The association between syntactic and pronunciational distances is even more heavily based on geography as a third factor (46%). The apparent association between syntactic and lexical distances turns out to be principally due to geography as a third factor (63%).

$$\text{Geographical influence} = \left( 1 - \frac{\text{correlation controlling for influence of geography}}{\text{correlation not controlling for geography}} \right) * 100$$

Figure 4-14: Influence of geography underlying the associations between the linguistic levels as a percentage.

## 4.10. Conclusions

Without controlling for the effect of geography, pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and syntax is much more strongly associated with pronunciation (42%) than with lexis (25%). Pronunciation and syntax are more strongly associated with geography (47% and 45%, respectively) than lexis is (33%).

<sup>14</sup> The geographical influence underlying the association between pronunciational and lexical distances is calculated as follows:  $(1 - (0.374 / 0.617)) * 100 = 39\%$ .

However, once the influence of geography is filtered away as a factor of influence underlying the associations among the linguistic levels under investigation, the association between pronunciation and syntax turns out to be largely based on geography as an underlying factor (46%). Some influence between pronunciation and syntax remains (12%), although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (3%).

#### **4.11. Discussion and future research**

We wish to point to two consequences beyond the raw correlations of the distances among the linguistic levels, as interesting as these are on their own. First, the modest correlation ( $r = 0.35$ ) between syntactic and pronunciational variables in Table 4-8 indicates that 12% of the proportion of variance in common between the two variables cannot be explained by geography. It might be explained by typological constraints—i.e. by constraints obtaining between syntactic and phonological structure—which would be very interesting. If we had found no interesting level of correlation between these levels on the one hand and the lexical level on the other, one might postulate immediately that typological constraints are responsible for this modest correlation. But we, in fact, did find a comparable level of correlation between pronunciation and lexical choice, for which structural, typological constraints seem unlikely. We therefore must allow that extralinguistic, but clearly non-geographical explanations are equally plausible as candidates to explain the correlation.

Second, we turn to the modest correlation ( $r = 0.37$ ) between pronunciational and lexical variation on the one hand and the low, but significant correlation ( $r = 0.18$ ) between lexical and syntactic variation on the other. These coefficients in Table 4-8 indicate that 14% of the proportion of variance in common between lexical and pronunciational distances on the one hand, and 3% of the proportion of variance in common between lexical and syntactic distances on the other hand, cannot be explained by geography. As we have argued above, it is unlikely that these correlations may be explained by linguistic constraints, and since the correlations were obtained from the residues of a regression analysis in which geography was the independent variable, they are not explained by geography. This suggests that there must be further extralinguistic conditioning of variation that we as dialectologists should set in our sites. The literature on language variation suggests many candidates for such conditioning variables, but there have been too few data collection efforts aimed at cataloguing linguistic variation and candidate explanatory variables, including e.g. sex, education, class, social network, etc. This would indeed be a daunting task, but the present paper has sketched the sorts of analysis one could perform on the data, once it is available.

To summarise, the degrees of association among the linguistic levels presented in Section 9 are substantial but not overwhelming. There is influence between the various linguistic levels, even once we control for the dominant effect of geography. We assume that a more evenly geographically distributed set of dialect varieties may result in stronger degrees of association, since the current set of common varieties overrepresents the average variation spectrum in the Dutch language area. Regardless, the results further strengthen the fundamental postulate in dialectology that language varieties are structured geographically.

We note, however, that the results at the lexical level are consistently less strong in comparison to the results at the pronunciational and syntactic levels. We speculate that the unfavourable lexical results reflect the lower quality of the lexical data set. The consistency analysis of the lexical data in Section 6 hints at this direction. Future work will further examine the lexical data using a bootstrapping technique to analyse the influence of the selection of words on the results.

Once geography is controlled for as an underlying factor of influence, the lack of association between lexis and syntax accords with our expectations as stated in Section 2. However, we are surprised that the association between lexis and pronunciation is somewhat stronger than the correlation between syntax and pronunciation. We would have expected the highly structured syntactic and pronunciational systems to share more distributional patterns, in contrast to the volatile lexicon. We suspect that this outcome is another reflection of the somewhat lower quality of the lexical data. Also, the unbalanced nature of the syntactic data may be a factor of influence. SAND1 only describes variation in the left periphery of the clause and pronominal reference. However, the second volume of the SAND (SAND2, Barbiers et al., t.a. 2008) will concentrate on syntactic variation with respect to verbal clusters, negation and quantification. We will integrate the variation in these right peripheral domains in our syntactic measurements to further enhance the accuracy of our results.

Finally, pronunciational differences can arise from variation at the phonetic, phonological and morphological levels. Future research will attempt to dissect the complex interplay of these linguistic levels underlying pronunciational differences as follows. First, we are currently processing the purely morphological data in the first volume of the Morphological Atlas of the Dutch Dialects (MAND, De Schutter et al., 2005).<sup>15</sup> Second, we are also investigating the purely phonological data in the Phonological Atlas of the Dutch Dialects (FAND, Goossens et al., 1998-2005). We expect these extensive sources of purely morphological data and purely phonological data to provide new insights in the roles of the various linguistic levels underlying pronunciational differ-

---

<sup>15</sup> More information regarding De Schutter et al. (2005) is available on the official MAND website at <http://www.meertens.knaw.nl/projecten/mand>.

ences, and to enrich our understanding of the associations among linguistic levels.

#### 4.12. References

- Barbiers, S., Bennis, H., Devos, M., Vogelaar, G. de, Ham, M. van der (eds), 2005. *Syntactic Atlas of the Dutch Dialects*, Volume 1. Amsterdam University Press, Amsterdam.
- Barbiers, S., Bennis, H., Vogelaar, G. de, Auwera, J. van der, Ham, M. van der (eds), t.a. 2008. *Syntactic Atlas of the Dutch Dialects*, Volume 2. Amsterdam University Press, Amsterdam.
- Blancquaert, E., Peé, W. (eds), 1925-1982. *Reeks Nederlands(ch)e dialectatlassen*. De Sikkel, Antwerpen.
- Cavalli-Sforza, L., Wang, W., 1986. *Spatial distance and lexical replacement*. In: *Language*, Volume 62(1), 38–55.
- Cronbach, L., 1951. *Coefficient alpha and the internal structure of tests*. In: *Psychometrika*, Volume 16, 297–334.
- Daan, J., Blok, D., 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Volume XXXVII, Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.
- Donegan, P., Stamp D., 1983. *Rhythm and the holistic organization of language structure*. In: Richardson, J.F. et al. (eds), *Papers from the Parasession on the interplay of phonology, morphology and syntax*. Chicago Linguistic Society, Chicago, 337–353.
- Gemert, I. van, 2002. *Het geografisch verklaren van dialectafstanden met een geografisch informatie-systeem (GIS)*. Master's thesis, Rijksuniversiteit Groningen, Groningen.
- Goebel, H., 1982. *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Austrian Academy of Science, Wien.
- Goebel, H., 1984. *Dialektometrische Studien: Anhand italo-romanischer, rätoro-romanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Volume 3, Max Niemeyer, Tübingen.
- Gooskens, C., Heeringa, W., 2004. *Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data*. *Language variation and Change*, Volume 16(3), 189–207.
- Gooskens, C., 2004. *Norwegian Dialect Distances Geographically Explained*. In: Gunnarson, B., Bergström, L., Eklund, G., Fridella, S., Hansen, L., Karstadt, A., Nordberg, B., Sundgren, E., Thelander, M. (eds), *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2*, Uppsala, Sweden: Uppsala University, 195–206.
- Goossens, J. 1977. *Inleiding tot de Nederlandse dialectologie*. Wolters-Noordhoff, Groningen.
- Goossens, J., Taeldeman J., Verleyen, G., 1998. *Fonologische atlas van de Nederlandse dialecten*, Volume I, Het korte vocalisme. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.
- Heeringa, W., 2001. *De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen*. In: Hoeksema, J. et al. (eds), *TABU: Bulletin voor Taalwetenschap*, Volume 31(1/2). Rijksuniversiteit Groningen, Groningen, 61–103.



- Heeringa, W., Nerbonne, J., 2001. *Dialect Areas and Dialect Continua*. In: Sankoff, D. (eds), *Language Variation and Change*, Volume 13, Cambridge University Press, New York, 375–400.
- Heeringa, W., 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis, Rijksuniversiteit Groningen, Groningen.
- Heeringa, W., Nerbonne, J., Bezooijen, R. van, Spruit, M., 2007. *Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied*. In: *Nederlandse Taalen Letterenkunde*, Volume 123(1), Uitgeverij Verloren, Hilversum, 70–82.
- Kessler, B., 1995. *Computational dialectology in Irish Gaelic*. In: *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, Dublin, 60–67.
- Levenshtein, V., 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. In: *Cybernetics and Control Theory*, Volume 10(8), 707–710.
- Manly, B., 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, Second edition.
- Mantel, N., 1967. *The detection of disease clustering and a generalized regression approach*. In: *Cancer Research*, Volume 27, 209–220.
- Nerbonne, J., Heeringa, W., Kleiweg, P., 1999. *Edit Distance and Dialect Proximity*. In: Sankoff, D., Kruskal, J. (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Press, Stanford, v–xv.
- Nerbonne, J., Siedle, C., 2005. *Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede*. In: *Zeitschrift für Dialektologie und Linguistik*, Volume 72(2), 2005, 129–147.
- Nerbonne, J., Kleiweg, P., 2007. *Toward a Dialectological Yardstick*. In: *Journal of Quantitative Linguistics*, Volume 14(2), Routledge, New York, 148–167.
- Nerbonne, J., Heeringa, W., t.a. 2007. *Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation*. In: Featherston, S., Sternefeld, W. (eds), *Roots: Linguistics in Search of its Evidential Base*, Mouton De Gruyter, Berlin.
- Norušis, M., 1997. *SPSS Professional Statistics 7.5*. SPSS Inc, Chicago.
- Nunnally, J., 1978. *Psychometric Theory*. McGraw-Hill, New York.
- Sankoff, D., Kruskal, J. (eds), 1999. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Press, Stanford.
- Schutter, G. de, 1994. *Dutch*. In: König, E., Auwera, J. van der (eds), *The Germanic languages*. Routledge Language Family Descriptions, London, Routledge, New York.
- Schutter, G. de, Berg, B. van den, Goeman, T., Jong, T. de (eds), 2005. *Morphological Atlas of the Dutch Dialects*. Volume 1, Amsterdam University Press, Amsterdam.
- Spruit, M., 2005. *Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited*. In: Doetjes, J., Weijer, J. van de (eds), *Linguistics in the Netherlands*, 2005, John Benjamins, Amsterdam, 179–190.
- Spruit, M., 2006. *Measuring syntactic variation in Dutch dialects*. In: Nerbonne, J., Kretzschmar, W. Jr. (eds), *Literary and Linguistic Computing*, special issue on Progress in Dialectometry: Toward Explanation, Volume 21(4), 493–506.

Weijnen, A., 1958. *Nederlandse dialectkunde*. Van Gorcum & Comp. N.V. - G.A. Hak & Dr. J. Prakke, Assen.

## 5. Discovery of association rules between syntactic variables

“Data mining the Syntactic Atlas of the Dutch dialects”<sup>\*</sup>

*Spruit, M.R., 2007. Discovery of association rules between syntactic variables. Data mining the Syntactic Atlas of the Dutch dialects. In: Dirix, P., Schuurman, I., Vandeghinste, V., Eynde, F. van (eds), Computational Linguistics in the Netherlands 2006. Selected papers from the seventeenth CLIN meeting, 83–98.*

This research applies an association rule mining technique to purely syntactic dialect data. The paper answers the research question of how relevant associations between syntactic variables can be discovered. The method calculates the proportional overlap between geographical distributions of syntactic microvariables and incorporates rule quality factors such as accuracy, coverage and completeness to measure the interestingness of the variable associations. The exploratory review of the results discusses several highly ranked association rules and also examines an implicational chain of syntactic variables.

### 5.1. Introduction

This work investigates a data mining technique to discover associations between syntactic variables in Dutch dialects using a rule induction system based on proportional overlap. The research aims to contribute to the understanding of the associations between syntactic variables by examining geographical distributions of syntactic microvariation. The current paper addresses the following two research questions:

- I. How can relevant associations between syntactic variables be discovered?
- II. What are interesting associations between syntactic variables?

This research integrates expertise from the research fields of data mining and ecology to answer these questions quantitatively. In essence this investigation exhaustively evaluates levels of association between combinations of syntactic variables based on the proportional overlap between their geographical distributions.

---

<sup>\*</sup> This paper was presented in the Dialects session at the seventeenth Computational Linguistics in the Netherlands meeting in Leuven, Belgium, on 12 January 2007. The research is being carried out in the context of the NWO project The Determinants of Dialectal Variation, number 360-70-120, P.I. J. Nerbonne. Please visit <http://dialectometry.net> for more information and relevant software.

This work proceeds from the observation that linguistic research frameworks such as generative syntax and functional typology share a primary interest in understanding the structural similarities and differences between language varieties. The frameworks aim to identify which universal syntactic properties can vary across language varieties and which remain constant. The ultimate goal is to characterise the superficial structural diversity of all language varieties as particular settings of relatively few parametric patterns. Unfortunately, the search for syntactic universals is still very much a topic of ongoing research. Gianollo, Guardiano and Longobardi (t.a. 2007) most notably define an extensive parametric framework to model language variation in the internal structure of Determiner Phrases based on a relatively wide sample of languages and language families.

Haspelmath (2007) compiles a list of seven universal syntactic parameters for which there is a wide consensus in the field. One well-known example of a syntactic universal is the pro-drop/null-subject parameter, which states that the subject position in a clause may be empty or must be filled by a subject pronoun. It was originally thought to universally correlate with syntactic phenomena such as null thematic subjects and null expletives (Rizzi, 1986). However, the generalisation quickly became untenable once more language varieties were analysed (Newmeyer, 2005). This example adequately illustrates that a large data set of comparable language varieties is required to investigate syntactic variable relationships more reliably. Such an examination needs to be automated using verifiable methods because of the exhaustive and repetitive nature of the comparison procedure.

The current research aims to contribute to the global research effort of parameterisation of the structural diversity of language varieties by proposing a computational method to discover syntactic variable associations automatically. The technique facilitates exploration of previously unknown variable relationships and validation of existing parametric generalisations. The second research question is addressed through an exploratory review of the method's application to a large syntactic microvariation database.

The paper is structured as follows. Section 2 describes the unique syntactic variation database under investigation. Section 3 introduces the sample data subset used in Section 4 to illustrate the association rule mining procedure based on proportional overlap. Section 5 reviews the evaluation factors to accurately measure the quality of the association rules. Section 6 explores the most interesting rules discovered in the sample data. Section 7 highlights results of the association rule mining application to the entire syntactic variation database under investigation. Section 8 recapitulates the main findings. The paper concludes with a discussion and directions for future research in Section 9.

## 5.2. Syntactic variation database



Figure 5-1: Distribution of the 267 Dutch dialects in the *Syntactic atlas*.



Figure 5-2: The provinces in the Dutch language area under investigation.

This research examines the first volume of the *Syntactische Atlas van de Nederlandse Dialecten* (SAND1; ‘Syntactic Atlas of the Dutch Dialects’; Barbiers et al., 2005) from a quantitative perspective. SAND1 contains 145 geographical distribution maps of individual syntactic variables in 267 Dutch dialects in the Netherlands, the Northern part of Belgium and a small north-western part of France. Figure 5-1 and Figure 5-2 show the geographical distribution of the SAND dialect locations and the relevant province names, respectively. SAND1 covers syntactic variation related to the left periphery of the clause and pronominal reference. This includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. The second and final volume of the SAND is due to appear in 2008 and will describe syntactic variation in Dutch dialects with respect to verbal clusters, negation and quantification. Cornips and Jongenburger (2001) review the methodological aspects of the written and oral syntactic elicitation techniques which were employed to reliably collect the SAND data.

From a quantitative research perspective SAND1 also represents a syntactic microvariation database containing 106 syntactic contexts and 485 syntactic variables among varieties of a single language. This work defines a syntactic variable as a form or word order in a syntactic context in which two dialects can differ (Spruit, 2006). The number of available syntactic contexts is somewhat lower than the number of geographical maps because SAND1 also contains numerous correlation maps which show syntactic variables from different perspectives. Also, some syntactic contexts are presented using multiple maps.

Table 5-1: Map 14b in SAND1 shows seven syntactic variables in the complementisers domain.

Context: Complementiser of comparative if-clause  
 Variables: { of, \*of dat, dat, as/of + V2, at, as, et }  
 Example: 't lijkt wel of dat er iemand in de tuin staat.  
 'it looks [affirmative] if that there someone in the garden stands'  
 "It looks as if there is someone in the garden."

Table 5-2: Map 54a in SAND1 shows four syntactic variables in the subject doubling domain.

Context: Subject doubling 2 singular  
 Variables: { V<sub>FINITE</sub> \_\_, \* \_\_ V<sub>FINITE</sub> \_\_, C \_\_, \*C<sub>COMPARATIVE</sub> \_\_ }  
 Example: Ge gelooft gij zeker niet dat hij sterker is as -ge gij.  
 'you<sub>weak</sub> believe you<sub>strong</sub> certainly not that he stronger is than you<sub>weak</sub> you<sub>strong</sub>'  
 "You do not seem to believe that he is stronger than you."

Table 5-3: Map 68a in SAND1 shows five syntactic variables in the reflexives domain.

Context: Weak reflexive pronoun as object of inherent reflexive verb  
 Variables: { zich, hem, \*zijn eigen, zichzelf, hemzelf }  
 Example: Jan herinnert zijn eigen dat verhaal wel.  
 'John remembers his own that story [affirmative]'  
 "John certainly remembers that story."

Table 5-4: Map 84a in SAND1 shows four syntactic variables in the fronting domain.

Context: Short subject relative, complementiser following relative pronoun  
 Variables: { \*1:die 2:as/at/da(t), 1:die 2:-t, 1:dien 2:at/da(t), 1:die/dat 2:wat }  
 Example: Dat is de man die dat het verhaal verteld heeft.  
 'that is the man who that the story told has'  
 "That is the man who told the story."

Table 5-1 to Table 5-4 provide examples of syntactic variation in the complementisers, subject doubling, reflexives and fronting domains, respectively. For example, Table 5-1 shows the attested variation throughout the Dutch language area in the realisation of the complementiser position in comparative if-clauses as presented in SAND1 map B on page 14. In standard Dutch people say 't *lijkt wel of er iemand in de tuin staat* 'it looks [affirmative] if there someone in the garden stands', but in colloquial Dutch the following form also frequently occurs in the southern provinces: 't *lijkt wel of dat er iemand in de tuin staat*. There are even a few northern and southern regions within the Dutch language area

where the verb occurs in the second position of the if-clause: *'t lijkt wel of er staat iemand in de tuin*. The last example also illustrates that both word form and word order may vary within a syntactic context.

### 5.3. Sample data illustration and diagram



Figure 5-3: This SAND1 sample marks the occurrences in seven dialects (1-7) of the four syntactic variables (A-D) in Table 5-1 to Table 5-4.

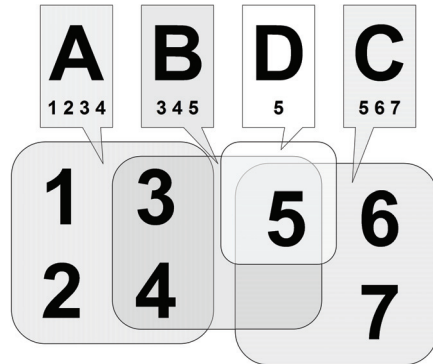


Figure 5-4: Symbolic representation of the SAND1 sample shown in Figure 5-3.

Figure 5-3 and Figure 5-4 illustrate the data mining procedure presented in the next section by defining a small subset of the actual SAND1 data. Figure 5-3 marks the geographical occurrences in seven Dutch dialects (1-7) of the four example variables (A-D) shown in Table 5-1 to Table 5-4. For example, Figure 5-3 shows that in the dialects of Ouddorp (1), Merckeghem (2), Brussel (3) and Gemert (4), people can say *'t Lijkt wel of dat er iemand in de tuin staat* (A). This variable does not occur in the dialects of Nieuwmoer (5), Boskoop (6) and Nijkerk (7). Likewise, only in the village of Nieuwmoer have all of the following three variables been attested: *Als gij gezond leeft, leef-de gij langer* (B), *Jan herinnert z'n eigen dat verhaal wel* (C), and *Dat is de man die dat het verhaal verteld heeft* (D). Figure 5-4 shows a symbolic representation of the sample data in Figure 5-3. The remainder of the current article uses the symbolic variable characters (A-D) and dialect numbers (1-7) to refer to the sample data components to enhance readability.

### 5.4. Association rule mining based on proportional overlap

The SAND1 sample data described above are used to illustrate how relationships between variables in a database can be discovered using a technique best known as data mining but arguably more accurately described with its synonym Knowledge Discovery in Databases (KDD). Data mining is an umbrella term

for various knowledge representation techniques such as association *rules*, decision *trees* and neural *networks*. Frawley, Piatetsky-Shapiro and Matheus (1992) define data mining as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Hand, Mannila and Smyth (2001) formulate data mining more generally as the science of extracting useful information from large data sets or databases.

This work explores associations between syntactic variables in Dutch dialects using a rule induction system based on proportional overlap. Generally speaking, association rules show attribute-value conditions that occur frequently together in a given dataset. The left side of an association rule is called the antecedent and may consist of multiple predicting attributes. The right side of a rule is called the consequent and defines the predicted class(es). Association rules are typically written as ‘ $A \rightarrow C$ ’ and should be read as ‘*if* variable  $A$  *then* variable  $C$ ’. A widely-used example of association rule mining is Market Basket Analysis, a method which examines a long list of supermarket transactions to determine which items are most frequently purchased together. It applies the Apriori algorithm to generate candidate association rules which relate the items within each transaction or basket (Agrawal, Imielinski and Swami, 1993).

The application of association rule mining between syntactic variables in the current paper examines all  $k$ -combinations (or  $k$ -subsets) of syntactic variables to determine which variable subsets most frequently co-occur geographically.<sup>1</sup> A  $k$ -combination is an unordered collection with  $k$  unique elements.<sup>2</sup> Figure 5-5 illustrates how to calculate the binomial coefficient of the number of combinations with three elements in the sample data set of the four variables  $\{A,B,C,D\}$ . In this example the binomial coefficient is four and represents the combinations  $\{A,B,C\}$ ,  $\{A,B,D\}$ ,  $\{A,C,D\}$  and  $\{B,C,D\}$ .

$$C_n^k = C_4^3 = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times (1)} = \frac{24}{6} = 4$$

Figure 5-5: Calculation of the number of combinations with  $k=3$  elements from the sample data set with  $n=4$  variables.

Table 5-5 lists the association rule mining algorithm in pseudocode. The procedure is scalable to even larger data sets because it is non-recursive. Therefore, memory usage remains constant. Line 1 specifies that the procedure iterates through all combinations with  $k=2$  to  $k=n$  variables. Line 2 selects the first

<sup>1</sup> The Rule INduction Console (*vinc*) programme implements the association rule mining procedure. It has been developed with the wxWidgets C++ toolkit and the next\_combination STL template. The console programme is available for all software platforms and can be downloaded from <http://dialectometry.net/syntax>.

<sup>2</sup> This is in contrast with a  $k$ -permutation, which is an *ordered* collection with  $k$  unique elements.



combination subset  $s$  with  $k$  variables. Then, lines 3 to 11 repeatedly process subset  $s$  and select the next subset. Line 4 iterates through all combinations of subset  $s$  with  $m=1$  to  $m=k-1$  variables. Line 5 generates the first combination subset  $a$  as the antecedent variables subset from  $s$  with  $m$  variables. Then, lines 6 to 9 repeatedly process subset  $a$  and select the next subset. Line 7 determines the corresponding consequent variables by selecting the complementary set of  $a$  from  $s$ . Finally, line 8 evaluates the quality of the generated association rule using the unique antecedent-consequent tuple based on the proportional overlap between the geographical distributions of the rule variables. The candidate association rule is accepted when it satisfies previously specified criteria of interestingness.

Table 5-5: Algorithm to non-recursively evaluate all association rules.

```

1.  FOR EACH  $k$ -combination of variable set  $v$  with  $n$  elements
2.      INITIALISE combination subset  $s$  from  $v$ 
3.      REPEAT
4.          FOR EACH  $m$ -combination of  $s$ 
5.              INITIALISE antecedent  $a$  from  $s$  with  $m$  elements
6.              REPEAT
7.                  INITIALISE consequent  $c$  as the complement of  $a$  with  $k-m$  elements
8.                  CALL evaluateAssociationRule with  $a$  and  $c$ 
9.                  UNTIL all antecedent combinations  $a$  have been processed
10.             ENDFOR
11.         UNTIL all combination subsets  $s$  have been processed
12.     ENDFOR

```

The procedure remains modest in automatically discarding uninteresting candidate rules. The current version of the algorithm only prunes the combination space in two cases. In the first, self-explanatory situation the interestingness value is either equal to or below zero. The second condition applies when the coverage value has the maximum value. This indicates that the antecedent encompasses the entire data set, which implies that the rule does not have any explanatory power. Of course, manual factor threshold values may be applied as well in addition to these conditions to further minimise the amount of uninteresting rules.

The proportional overlap procedure in this work consists of the following three steps. First, the lists of geographical occurrences of all syntactic variables in the rule antecedent are disjunctively merged into the rule antecedent vector of geographical occurrences. Variable occurrences are not merged conjunctively because the procedure attempts to combine microvariables to discover more general patterns. Then, the procedure constructs the rule consequent vector of geographical occurrences. Finally, the intersection and union sets of the two

vectors of geographical co-occurrences are calculated as factor components to help determine the quality of the candidate rule using a combination of indicators as listed in Table 5-6. The intersection set  $|A\&C|$  in Table 5-6 represents the geographical conjunction of antecedent and consequent variable occurrences. The concept of proportional overlap is predominantly applied in research areas such as ecology and biogeography and is notably explored in (Horn, 1966).

### 5.5. Evaluating the quality of a rule

Table 5-6 lists several widely used factors to help determine the quality of an association rule: accuracy, coverage, completeness and interestingness. Many more factors have been proposed over the years to further enhance rule evaluation quality. McGarry (2005) reviews a range of objective and subjective measures such as actionability, surprisingness, unexpectedness, misclassification cost, class distribution and attribute ranking, among others. These factors are not taken into account in this work. However, the current paper does incorporate complexity as the total number of variable disjuncts in both the antecedent and consequent sets. Higher complexity results are interpreted as being less interesting.

Table 5-6: Evaluation factors to help determine the quality of association rule ' $A \rightarrow C$ '.

Accuracy:	$ A\&C  /  A $	The number of dialects which have both variables A and C divided by the number of dialects which have variable A.
Coverage:	$ A  / N$	The number of dialects which have variable A divided by the total number of dialects in the data set.
Completeness:	$ A\&C  /  C $	The number of dialects which have both variables A and C divided by the number of dialects which have variable C.
Interestingness:	$ A\&C  -  A  C /N$	The number of dialects which have both variables A and C minus the product of the number of dialects which have variable A with the number of dialects which have variable C divided by the total number of dialects in the data set.

It is important to note that although a pattern is expressed as a rule, it does not mean that it is true all the time. An association rule does not imply causality. The antecedent of a rule does not necessarily cause the consequent of a rule to happen. Therefore, the uncertainty in a rule should be made explicit. This is what the accuracy of a rule indicates. It signifies how often a rule is correct and

is also called the confidence of a rule. The coverage of a rule expresses how often a rule applies and is also called support. The factor completeness may be used to explore how much of the target class a rule covers. This work multiplies all accuracy, coverage and completeness values by one hundred to express the rule quality factors as percentages.

The three rudimentary interestingness factors described above are always integrated in proposed measures of rule interestingness. Intuitively, rules are interesting when they have high accuracy, high coverage and deviate from the norm. The effort, then, is to formulate the optimal trade off between coverage, accuracy and potentially other factors for a specific problem domain. The domain specificity of interestingness is one of the many reasons why the ability to interactively explore the generated association rules is always desirable and maybe even inevitable. Although data mining algorithms may use objective factors to decide whether a rule is genuinely interesting or not, domain-specific, subjective notions of interestingness may be required as well to decide whether a potentially or technically interesting rule is also genuinely interesting in a specific domain. For example, a discovered association rule may be too well-known or too trivial.

Table 5-7: *Piatetsky-Shapiro's principles for rule interestingness (RI) measures.*

1.  $RI = 0$  if  $|A\&C| = |A| |C| / N$ .
2. RI monotonically increases with  $|A\&C|$  when other parameters are fixed.
3. RI monotonically decreases with  $|A|$  or  $|C|$  when other parameters are fixed.

This work applies the three principles for rule interestingness measures proposed in (Piatetsky-Shapiro, 1991). They are reprinted in Table 5-7. The principles formulate the relations between the factors accuracy, coverage and completeness as objective evaluation criteria of interestingness measures. The first principle states that the rule interestingness is zero if the antecedent and consequent of the rule are statistically independent. The second principle defines that more co-occurring elements in the antecedent and consequent of the rule will result in higher accuracy and completeness values when all other parameters remain fixed, which increases the interestingness of the rule. The third principle's interpretation is two-fold. It formulates that rule interestingness monotonically decreases with completeness when all other parameters remain fixed. Similarly, rule interestingness also monotonically decreases with coverage when all other parameters remain fixed (Freitas, 1999). Note that, in contrast with accuracy, coverage and completeness values, interestingness values do not necessarily range between zero and one.

Several enhancements and alternative measures of interestingness have been proposed since (Piatetsky-Shapiro, 1991). Lenca (2008) most notably describes

numerous measures of interestingness in detail. The current work restricts itself to Piatetsky-Shapiro's measure of interestingness because of its historical position and formulaic simplicity. Note, however, that its symmetric nature is a property where this measure seems lacking. This is not the case for the factors accuracy, coverage and completeness. To a certain extent the influence of symmetry can be compensated by ranking the entire result set of association rules firstly on descending interestingness, secondly on ascending complexity, thirdly on descending accuracy and finally on descending coverage.

### 5.6. Discovery of association rules between syntactic variables

Table 5-8 lists the eight most interesting association rules based on occurrences in seven dialects of the four syntactic variables in the sample data as shown in Figure 5-3 and Figure 5-4. The algorithm in Table 5-5 generates fifty variable combinations for the sample data. Fourteen candidate rules are potentially interesting based on the Piatetsky-Shapiro measure of interestingness and have at least some explanatory power. From a technical perspective this means that fourteen association rules have an interestingness value greater than 0 and a coverage value smaller than 100 percent. The list in Table 5-8 is sorted on descending interestingness, ascending complexity and descending accuracy, respectively.<sup>3</sup>

Table 5-8: *The eight most interesting association rules in the sample data set as shown in Figure 5-3 and Figure 5-4 sorted on descending interestingness, ascending complexity and descending accuracy.*

#	Antecedent $\rightarrow$ Consequent	Interestingness	Complexity	Accuracy	Coverage	Completeness
1.	B $\rightarrow$ A $\vee$ D	0.86	1	100	42	60
2.	A $\vee$ D $\rightarrow$ B	0.86	1	60	71	100
3.	D $\rightarrow$ B	0.57	0	100	14	33
4.	D $\rightarrow$ C	0.57	0	100	14	33
5.	B $\rightarrow$ D	0.57	0	33	42	100
6.	C $\rightarrow$ D	0.57	0	33	42	100
7.	B $\rightarrow$ A	0.29	0	66	42	50
8.	A $\rightarrow$ B	0.29	0	50	57	66

The list of association rules is primarily sorted on descending interestingness since the main goal of this work is to discover the most interesting association rules between the variables. The list's secondary sort factor uses ascending values of complexity which can be interpreted as an extension of the measure of

<sup>3</sup> The list of potentially interesting association rules can be sorted interactively using an external software programme such as Excel or SPSS.

interestingness. An increasing number of variable components in a rule decrease its comprehensibility and, therefore, its interestingness. Coincidentally, the application of the complexity factor in the sample data does not actually change the rule order. The list of association rules in Table 5-8 is ternarily sorted on descending accuracy. However, it would be equally valid to apply descending completeness as an alternative ternary sort factor. Favouring accuracy over completeness simply signifies that it is considered more important that a rule is correct than it is to discover the degree to which the consequent variables are predicted by the antecedent variables. The definitions of accuracy and completeness in Table 5-6 also illustrate these alternate perspectives on rule importance quite evidently. The first two rules in Table 5-8 demonstrate the effect of choosing completeness over accuracy to optimally sort the association rules. The rules have identical levels of interestingness and complexity but differ in the degree of accuracy and completeness. The first rule states that *if* variable B occurs in a dialect *then* variable A or D always occur as well; the rule is 100 percent accurate. However, it does not imply that the inverse is true as well. Indeed, in dialects one and two either variable A or D occurs but not variable B. This is specified in the second rule which states that if either variable A or D occurs in a dialect, then there is a 60 percent certainty that variable B occurs as well. This example adequately illustrates the asymmetric nature of the relationship between the antecedent variables and the consequent variables of an association rule. Furthermore, an asymmetric variable association may be interpreted as a variable dependency with potentially hierarchical implications.

### **5.7. Data mining the Syntactic Atlas of the Dutch Dialects**

The following pages highlight a small selection of potentially interesting association rules between the 485 syntactic variables in the SAND1 database based on their geographical co-occurrences in 267 Dutch dialects. The algorithm evaluated 234,740 rules without any variable disjunctions, i.e. all antecedents and consequents consist of only one variable, and found 10,730 interesting associations with an accuracy value of 90 percent or higher. This observation manifests the considerable proportional overlap between the syntactic variables in SAND1. Additionally, it could arguably be interpreted as an indication that highly interesting association rules with high coverage and high accuracy values effectively reduce the importance of the geographical occurrences in the data set. The information value of geography---by definition---becomes limited to generic density and distributional information when variable distributions overlap nearly perfectly. Ascending from the observational level of geographical distributions to more abstract variable associations would facilitate syntactic analyses to identify implicational chains and other association patterns.

The number of variable combinations rises to 113,614,160 candidate rules as soon as either the antecedent or consequent of a rule may include one variable

disjunction. No less than 56,267,729 generated association rules are at least 90 percent accurate.<sup>4</sup> This is to be expected since the algorithm disjunctively combines variables. Once a strong association between two variables has been found, any disjunctively added variable will further strengthen the association.

*Table 5-9: Example of a highly ranked association rule in SAND1 with one variable disjunct: “if either antecedent variable A1 or A2 occurs, then it is certain that the consequent variable also occurs”.*

Antecedent A1: p46b:julle(n)/jullie (Subject pronouns 2 plural, strong forms, complex)

We geloven dat julle(n)/jullie niet zo slim zijn als wij.

‘we believe that you<sub>plural,strong</sub> not so smart are as we’

“We believe that you are not as smart as we are.”

Antecedent A2: p46b:julder/jielder (Subject pronouns 2 plural, strong forms, complex)

We geloven dat julder/jielder niet zo slim zijn als wij.

‘we believe that you<sub>plural,strong</sub> not so smart are as we’

“We believe that you are not as smart as we are.”

Consequent: p46a:j-[lieden-compositum] (Subject pronouns 2 plural, strong forms)

We geloven dat j-lieden niet zo slim zijn als wij.

‘we believe that you<sub>plural,strong</sub> not so smart are as we’

“We believe that you are not as smart as we are.”

Statistics: Rank=9, Combination=5,327,848, Interestingness=61.31,  
Accuracy=100%, Coverage=40%, Completeness=93%,  
Complexity=1, A-Locations=107, C-Locations=114,  
AC-Overlap=107, AC-Disjunction=114.

Interpretation: The infrequent pronoun ‘julder/jielder’ perfects the implicational association of the frequent ‘julle(n)/jullie’ variant with the (abstract) ‘j-lieden’ group of complex pronouns.

Table 5-9 presents an association rule with one variable disjunction as an example of a potentially interesting rule with a higher complexity. However, higher complexity association rules become exceedingly more difficult to interpret linguistically.<sup>5</sup> As a matter of fact, it can already be quite challenging to

<sup>4</sup> The corresponding output file is 33 GB. The programme execution time was around 18 hours on a MacMini PowerPC G4 (1.5 GHz) computer.

<sup>5</sup> Illustratively, Spruit (2007) interprets the antecedents and consequent in the association rule shown in Table 5-9 as atomic variables (in this work’s terminology). However, Sjeff Barbiers recently pointed out that the consequent variable *j-lieden* actually represents an abstract group of variables which includes the atomic variables *jullie* and *julder*, among others. The *j-lieden* variable does not occur literally. Even though it is defined as an atomic variable in SAND1, the *j-lieden* variable should actually be interpreted as a composite variable. Therefore, this example also illustrates the applicability of the association rule mining technique as a data validation tool.

linguistically interpret rules without variable disjunctions. Interactive explorations can only partly facilitate the evaluation process. Therefore, the remainder of the current paper concentrates on association rules without variable disjunctions.

Table 5-10 shows the potentially most interesting association rule in SAND1 without variable disjunctions. The rule associates one of the variables in map A on page 46 in SAND1 with a variable in map B on page 38. It states that, in the context of a strong *plural* subject pronoun in second person, if the complex pronoun ‘g-lieden’ occurs, then the strong *singular* subject pronoun in second person ‘gij’ (or ‘gie’) nearly always occurs as well. This is indicated by the accuracy value of 99 percent. This value is calculated using the definition in Table 5-6 as follows:  $|A\&C| / |A| * 100 = AC\text{-Overlap} / A\text{-Locations} * 100 = 104 / 105 * 100 = 0.99 * 100 = 99$  percent. Similarly, the interestingness value results as follows:  $|A\&C| - |A| |C| / N = AC\text{-Overlap} - (A\text{-Locations} * C\text{-Locations} / 267) = 104 - (105 * 116 / 267) = 104 - 45.62 = 58.38$ .

Table 5-10: The most interesting rule in SAND1 without variable disjuncts.

Antecedent:	p46a:g-lieden (Subject pronouns 2 plural, strong forms) We geloven dat <u>g-lieden</u> niet zo slim zijn als wij. ‘we believe that you <sub>plural,strong</sub> not so smart are as we’ “We believe that you are not as smart as we are.”
Consequent:	p38b:gij/gie (Subject pronouns 2 singular, strong forms) Ze gelooft dat <u>gij/gie</u> eerder thuis bent dan ik. ‘she believes that you <sub>singular,strong</sub> earlier home are than I’ “She thinks that you’ll be home sooner than me.”
Statistics:	Rank=1, Combination=10,321, Interestingness=58.38, Accuracy=99%, Coverage=39%, Completeness=89%, Complexity=0, A-Locations=105, C-Locations=116, AC-Overlap=104, AC-Disjunction=117.
Interpretation:	The plural pronoun ‘g-lieden’ belongs to the same paradigm as the singular pronoun ‘gij’.

The geographical distributions of the rule variables in Table 5-10 are patterned quite coherently (not shown). All occurrences are found in the southern half of the Dutch language area. Although it may not be particularly surprising to discover a strong association between two typically southern word forms, it does not automatically follow that it may not be considered interesting or even significant to discover that the geographical overlap between, specifically, these two southern word forms is nearly all-inclusive. It is sufficient to interactively sort all association rules on antecedent name, descending interestingness and

descending accuracy, respectively, to verify this hypothesis. This action reveals that only nine potentially interesting association rules exist with the complex pronoun ‘g-lieden’ as their antecedent and which also have an accuracy of 90 percent or higher.

The top six ‘g-lieden’ rules state that if in a dialect people can say *We geloven dat g-lieden niet zo slim zijn als wij* ‘we believe that you<sub>strong</sub> not so smart are as we’, then people in that dialect can also say, in descending degree of certainty, (a) *Ze gelooft dat jij/gie eerder thuis bent dan ik* ‘she believes that you earlier home are than I’, (b) *Ik denk da Marie hem zal moeten roepen* ‘I think that Mary him will must call’, (c) *U [niet-beleefdheidsvorm] gelooft dat Lisa even mooi is als Anna* ‘you [non-honorific] believe that Lisa as beautiful is as Anna’, (d) *Fons zag een slang naast hem* ‘Fons saw a snake next to him’, (e) *Erik liet mij voor hem werken* ‘Erik let me for him work’ and (f) *De jongen wie/die z’n moeder gisteren hertrouwd is* ‘the boy who/that his mother yesterday remarried is’. Table 5-11 lists more details for rules (c) and (d).

Table 5-11: More potentially interesting consequents in association rules which have the complex pronoun ‘g- + lieden’ as their antecedent, in addition to the rule consequent in Table 5-10.

Combination: 3,962 (c)

Consequent: p41b:[no\_honorifics] (Honorifics)

U [niet-beleefdheidsvorm] gelooft dat Lisa even mooi is als Anna.  
 ‘you [no\_honorifics] believe that Lisa as beautiful is as Anna’  
 “You believe that Lisa is as beautiful as Anna.”

Statistics: Interestingness=28.97, Accuracy=92%, Coverage=39%,  
 Completeness=56%, A-Locations=105, C-Locations=173, AC-Overlap=97, AC-Disjunction=181.

Interpretation: The complex pronoun ‘g-lieden’ may either block honorific pronouns or ‘g-lieden’ is a honorific pronoun itself.

Combination: 10,182 (d)

Consequent: r70b:hem (Reflexive pronoun as object in a locative prepositional phrase)

Fons zag een slang naast hem.  
 Fons saw a snake to him  
 ‘Fons saw a snake next to him.’

Statistics: Interestingness=22.85, Accuracy=91%, Coverage=39%,  
 Completeness=51%, A-Locations=105, C-Locations=186,  
 AC-Overlap=96, AC-Disjunction=195.

Interpretation: The second person, plural pronoun ‘g-lieden’ nearly always co-occurs with the third person, singular, reflexive pronoun ‘him’. (sic)



Rules (d) and (e) also strongly indicate a relationship between the second person, plural complex pronoun ‘g-lieden’ and the third person, singular, reflexive pronoun ‘hem’. It is unclear how this association should be interpreted linguistically. Although the rules might describe a previously unknown linguistic relationship, it could also merely reflect that the variables are geographically clustered. The latter case would signify the methodological reminder that a strong variable association does not necessarily imply a linguistic causation. All in all, the analysis above adequately illustrates how exploration of one association rule may easily trigger interactive investigations of several more potentially interesting rules and may raise new questions to answer.

Table 5-12: The most interesting implicational chain of association rules between four syntactic variables:  $d54a:after\_v \rightarrow d55a:after\_v \rightarrow p46a:g-lieden \rightarrow p38b:gij/gie$ .

Variable 1/4:  $d54a:after\_v$  (Subject doubling 2 singular)

As  $gij$  gezond leeft, leef- de  $gij$  langer.  
 ‘if you<sub>singular</sub> healthily live, live- you<sub>singular,weak</sub> you<sub>singular,strong</sub> longer’  
 “If you live healthily you will live longer.”  
 # Rank=6, Combination=6,509, Interestingness=52,78, Accuracy=92%.

Variable 2/4:  $d55a:after\_v$  (Subject doubling 2 plural)

As  $gulder$  gezond leeft, leef- de gulder langer.  
 ‘if you<sub>plural</sub> healthily live, live- you<sub>plural,weak</sub> you<sub>plural,strong</sub> longer’  
 “If you live healthily you will live longer.”  
 # Rank=3, Combination=7.503, Interestingness=54,07, Accuracy=93%.

Variable 3/4:  $p46a:g-lieden$  (Subject pronouns 2 plural, strong forms)

We geloven dat g-lieden niet zo slim zijn als wij.  
 ‘we believe that you<sub>plural,strong</sub> not so smart are as we’  
 “We believe that you are not as smart as we are.”  
 # Rank=1, Combination=10,321, Interestingness=58,38, Accuracy=99%.

Variable 4/4:  $p38b:gij/gie$  (Subject pronouns 2 singular, strong forms)

Ze gelooft dat gij/gie eerder thuis bent dan ik.  
 ‘she believes that you<sub>singular,strong</sub> earlier home are than I’  
 “She thinks that you’ll be home sooner than me.”  
 # Rank=8, Combination=6,552, Interestingness=52,73, Accuracy=98%.

Another approach of interactively exploring the result set of rules focuses on the examination of implicational chains between syntactic variables. Table 5-12 lists the highest ranked implicational chain of four syntactic variables in the set of association rules without variable disjunctions to illustrate this phenomenon. First, rule six states that if subject doubling occurs after V in second person singular, then it also appears after V in second person plural. Second, the third highest rule asserts that if subject doubling occurs after V in second person plural, then the second person plural pronoun ‘g-lieden’ nearly always arises as well. As an aside, this rule effectively demonstrates the implicit capacity to discover variable associations across syntactic domains. Third, the highest ranked rule convincingly associates the second person plural pronoun ‘g-lieden’ with the second person singular pronoun ‘gij/gie’. Finally, rule eight confirms the transitive nature of the rules with the association between subject doubling after V in second person singular and the second person singular pronoun ‘gij/gie’.

From a statistical perspective many more linguistically interesting variable associations can be expected to surface upon closer investigation. The explorations described above merely attempt to indicate the great potential of association rule mining as a meaningful contribution to linguistic theory in general and syntactic theory in particular. Another promising approach could employ association rule mining to quantitatively validate existing and new typological hypotheses. This is in contrast with the current approach which focuses on exploration and identification of variable patterns. However, every approach will require extensive consultation with syntactic theorists to meaningfully interpret the data. SAND1 provides geographical maps of many individual variable distributions to facilitate interpretation and validation of potentially interesting association rules. The generated sets of induced association rules and the rule induction programme are publicly available for interactive exploration at <http://dialectometry.net/syntax>.

## **5.8. Conclusions**

This research has successfully demonstrated how associations between syntactic variables in Dutch dialects can be discovered computationally using an association rule mining technique based on proportional overlap. The rule induction system facilitates identification and exploration of previously unknown variable relationships and validation of existing parametric generalisations. The ability to define variable associations asymmetrically is considered to be an important property of the technique in the syntactic domain. The analysis of the sample data has indicated that the Piatetsky-Shapiro measure of interestingness adequately formulates the relationships between the evaluation factors of accuracy, coverage and completeness.

The application of the association rule mining technique to the Syntactic atlas of the Dutch dialects has revealed the existence of many potentially interesting associations with high accuracy and coverage values and showed considerable overlaps between the geographical distributions of syntactic variable pairs. The exploratory review has examined the highest ranked association rules and also discussed an implicational chain of variable associations. The results strongly indicate that many more potentially interesting associations between syntactic variables are likely to be uncovered upon further investigation.

## **5.9. Discussion**

The approach presented in this paper to discover associations between syntactic variables can be extended and refined in several ways. For example, the candidate generation algorithm listed in Table 5-5 could be extended to incorporate exception rules as well. These are rules which cannot be predicted from existing knowledge. Hussain (2000) defines a relative entropy measure to identify exception rules. Exception rules typically combine high accuracy with poor coverage values. Further refinements of the data mining procedure may include experimentation with alternative measures of interestingness and incorporation of additional rule quality evaluation factors such as surprisingness, among others.

An interesting property of data mining applications such as association rule mining arises as more variables become available to the procedure. The formula in Figure 5-5 shows that the number of generated candidate association rules increases factorially with the number of variables. Also, increasing complexity is another source of combinatorial explosion. These observations are relevant in the current context because the second volume of the SAND (SAND2) is due to appear in 2008. Incorporation of the SAND2 data into the association rule discovery process will result in a linguistic database containing around 750 syntactic variables and covering all major syntactic microvariation domains. Although the linguistically trained mind may be extremely effective in heuristically associating variables, the astronomical SAND combination space will undoubtedly exceed human limits of association precision and capacity. Additionally, the compartmented and repetitive nature of data mining algorithms makes them good candidates for computational scaling and parallelisation using grid computing techniques. Therefore, a combination of the unsurpassed human heuristic capabilities with the verifiable precision and processing power available to data mining tools may well contribute to the understanding of the structural diversity of language varieties. There is, of course, no reason to stop incorporating more data into the procedure. For example, it could be really interesting to combine available phonological data with these syntactic data to discover potential associations between variables among linguistic levels (cf. Spruit, Heeringa and Nerbonne, t.a. 2008).

An entirely different application of association rule mining analyses the set of variable associations to define clusters of geographically overlapping variables known as composite variables (Spruit, 2006). This application assumes that if a group of variables nearly always occur together, then a single variable of such a group does not add to the variation between two language varieties by itself. Therefore, from a quantitative perspective the cluster of variables can be interpreted as one entity which should more accurately quantify syntactic variation. Preliminary visualisations of the distance relationships between Dutch dialects based on the Jaccard distance between composite syntactic variables appear to classify the Dutch dialect areas quite accurately.<sup>6</sup> The dialect maps appear to be in line with expert opinion (cf. Schutter, 1994) and correspond with dialect distance visualisations in (Spruit, 2006) and (Spruit, Heeringa and Nerbonne, t.a. 2008) but require further research.

Finally, it would be interesting to compare the discovered variable associations with results based on more classic statistical methods such as Cramér's V or correspondence analysis. Cramér's V is a statistic which measures the strength of association between two categorical variables based on the  $\chi^2$ -statistic. Time permitting, this approach could be well worth investigating. One of the method's attractive benefits is that it calculates the statistical significance of each variable pair association. Another statistical technique which may hold promise is correspondence analysis (cf. Cichocki, 2006). This method resembles the factor analysis technique but has specifically been designed to help explore associations between categorical variables. However, the interpretability of the resulting correspondence visualisations may become an issue given the considerable geographical overlaps between the syntactic variable distributions. Furthermore, a more fundamental shortcoming of the two alternative approaches described above is the inherent symmetric nature of the discovered variable associations.

## 5.10. References

- Agrawal, R., Imielinski, T., Swami, A., 1993. *Mining association rules between sets of items in large databases*. In: Buneman, P., Jajodia, S. (eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, Washington, D.C., 207–216.
- Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M. van der (eds), 2005. *Syntactic Atlas of the Dutch Dialects*, Volume 1. Amsterdam University Press, Amsterdam.
- Cichocki, W., 2006. *Geographic variation in Acadian French /r/: What can correspondence analysis contribute toward explanation?*. In: Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation*, Volume 21, Oxford University Press, Oxford, 529–541.

---

<sup>6</sup> See (Jaccard 1901) for information on the nominal Jaccard measure.

- Cornips, L., Jongenburger, W., 2001. *Elicitation techniques in a Dutch syntactic dialect atlas project*. In: Broekhuizen, H., Wouden, T. van der (eds), *Linguistics in the Netherlands*, 2001, John Benjamins, Philadelphia/Amsterdam, 53–63.
- Frawley, W., Piatetsky-Shapiro, G., Matheus, C., 1992. *Knowledge discovery in databases: An overview*. *AI Magazine*, Volume 13, 213–228.
- Freitas, A., 1999. *On rule interestingness measures*. *Knowledge-based Systems*, Volume 12, 309–315.
- Gianollo, C., Guardiano, C., Longobardi, G., t.a. 2007. *Three fundamental issues in parametric linguistics*. In: Biberauer, T. (ed), *The Limits of Syntactic Variation*, John Benjamins, Philadelphia/Amsterdam.
- Hand, D., Mannila, H., Smyth, P., 2001. *Principles of Data Mining*. The MIT Press, Cambridge, MA.
- Haspelmath, M., t.a. 2007. *Parametric versus functional explanations of syntactic universals*. In: Biberauer, T., Holmberg, A. (eds), *The Limits of syntactic variation*, Benjamins, Amsterdam.
- Horn, H., 1966. *Measurement of overlap in comparative ecological studies*. *The American Naturalist*, Volume 100, 419–424.
- Hussain, F., Liu, H., Suzuki, E., Lu, H., 2000. *Exception rule mining with a relative interestingness measure*. In: Terano, T., Liu, H., Chen, A. (eds), *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer-Verlag, London, UK, 86–97.
- Jaccard, P., 1901. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. In: *Bulletin de la Societe Vaudoise des Sciences Naturelles*, Volume 37, 547–579.
- Jellinghaus, H., 1892. *Die niederländischen Volksmundarten: Nach den Aufzeichnungen der Niederländer*. In: Norden, H. (ed), *Forschungen/Verein für Niederdeutsche Sprachforschung*, Soltau.
- Lenca, P., Meyer, P., Vaillant, B., Lallich, S., 2008. *On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid*. In: *European Journal of Operational Research*, Elsevier, Volume 184(2), 610–626.
- McGarry, K., 2005. *A survey of interestingness measures for knowledge discovery*. *The Knowledge Engineering Review*, Volume 20, 39–61.
- Newmeyer, F., 2005. *Possible and probable languages: a generative perspective on linguistic typology*. Oxford University Press, Oxford.
- Piatetsky-Shapiro, G., 1991. *Discovery, analysis and presentation of strong rules*. In: Piatetsky-Shapiro, G., Frawley, W. (eds), *Knowledge Discovery in Databases*, AAAI/MIT Press, 229–248.
- Rizzi, L., 1986. *Null objects in Italian and the theory of pro*. In: *Linguistic Inquiry*, Volume 17, 501–557.
- Schutter, G. de, 1994. *Dutch*. In: König, E., Auwera, J. van der (eds), *The Germanic languages*, Routledge, London/New York.
- Spruit, M., 2006. *Measuring syntactic variation in Dutch dialects*. In: Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation*, Volume 21, Oxford University Press, Oxford, 493–506.

Spruit, M., Heeringa, W., Nerbonne, J., t.a. 2008. *Associations among linguistic levels*. In: *Lingua*, Special issue on Syntactic databases. Selected papers presented in the special session Comparing Aggregate Syntaxes, Digital Humanities conference, Paris, 2006.

## 6. Summary and conclusions

This dialectometrical research has investigated three quantitative perspectives on syntactic variation in Dutch dialects. The first perspective shows how to quantify syntactic differences between language varieties and classifies the Dutch dialect varieties based on a measure of syntactic distance. This objective classification is compared with—and highly resembles—the traditional, perceptual classification based on subjective judgements. This approach also affirmatively answers the question whether syntactic variation patterns are geographically coherent. The second perspective describes how to quantify the degrees of association between pronunciational, lexical and syntactic differences. This approach reveals that the degrees of association among the linguistic levels of pronunciation, lexis and syntax are genuine but modest. Also, syntactic and pronunciational differences are not more strongly associated with one another than either one is associated with lexical differences. The third perspective demonstrates how to discover relevant associations between syntactic variables using a data mining technique based on geographical co-occurrences. This approach contributes to the validation of existing typological hypotheses and facilitates the identification and exploration of variable relationships in general.

### 6.1. Chapter summary

Chapter 1 motivates the importance of quantitative linguistic research at a syntactic level. The chapter starts with two examples of syntactic variation in Dutch dialects, which indicate that each syntactic variation phenomenon may have a unique geographical distribution. Therefore, a quantitative methodology with a robust, empirical foundation is required to compensate for the idiosyncrasies of individual variables. This allows the data to be examined from more general perspectives. The chapter continues with an introduction of the research fields of *dialect cartography*, *dialectometry* and *syntactic microvariation* to sketch the scientific context and relevance of this first investigation of dialectometrical applications to purely syntactic dialect data. Then, it presents the first volume of the *Syntactische Atlas van de Nederlandse Dialecten* (SAND1; ‘Syntactic Atlas of the Dutch Dialects’; Barbiers et al., 2005) as the first compendium of Dutch syntactic variation and the main data source for this work. The current study is also highlighted from four different research dimensions to indicate what this research is *not* about. An introductory overview of the chapters in this dissertation follows after the four following research questions have been formulated and clarified:

- I. How can syntactic variation be measured adequately? (*Model*)
- II. What are the syntactic distances among the Dutch dialects? (*Application*)

- III. To what extent are the linguistic levels of syntax, lexis and pronunciation associated with each other? (*Context*)
- IV. What are relevant dependencies between syntactic variables? (*Associations*)

Research questions I and II jointly address the relation between syntactic and geographical distance. The first question focuses on how to *model* syntactic differences between language varieties such that syntactic variation can be examined reliably in the aggregate to provide more general perspectives on syntactic variation. The second research question concentrates on the *application* of the measurement model to the first compendium of purely syntactic Dutch dialect data and analyses the results. These two research questions are answered in Chapters 1 and 3. Research question III addresses the degree to which geographical distributions of syntactic distances correlate with distributions of pronunciational and lexical distances. The question helps to put the syntactic measurement results into a broader linguistic *context* by calculating the extent to which syntactic variation correlates with pronunciational and lexical variation. This research question is the topic of Chapter 4. Research question IV addresses the discovery of relevant *associations* between syntactic variables. It contributes to the global linguistic research effort of parameterisation of the structural diversity of language varieties by identifying which syntactic variables nearly always co-occur geographically. This research question is investigated in Chapter 5.

Chapter 2 investigates the relationship between syntactic variation and geographical distance by addressing the Dutch dialect classification problem from a dialectometrical perspective. It compares the syntactic measurement results projected on a geographical map with the traditional *Daan and Blok (1969) map* of the Dutch dialects based on *subjective judgements*. The chapter presents a quantitative measure of syntactic distance to objectively and verifiably differentiate dialect borders and dialect continua. It discusses the *arrow method* and the methodological challenges underlying the perceptual classification of the Dutch dialects based on subjective judgements. These problems bring about the introduction of the research field of dialectometry and SAND1 as a purely syntactic database containing 510 syntactic variables suitable for quantitative analysis. The dialectometrical method described in this chapter *aggregates syntactic differences* between dialect varieties using a *Hamming distance* algorithm until the highly repetitive measurement procedure results in the SAND1 *Hamming distance matrix*. The dialect relationships in the distance matrix are analysed by applying the Classical *Multidimensional scaling* (MDS) procedure to optimally represent the most differentiating syntactic variables for each dialect in relation to all other dialects. The variation in the Dutch language area is visualised geographically using *full-colour dialect maps*, in which the MDS map colours correspond with the first three dimensions of the MDS solution. The review of the results first dis-



cusses the application of the MDS procedure to each of the seven SAND1 domains separately. Then, the aggregate SAND1 MDS dialect map based on a syntactic Hamming distance measure is calculated, which results in a homogeneous colour continuum with discernable dialect regions. The SAND1 MDS map evidently shows that syntactic variation is structured in a *geographically coherent* way when viewed in the aggregate. Furthermore, the objective classification of Dutch dialect varieties based on a syntactic measure highly resembles the classification based on subjective judgements on the Daan and Blok dialect map, which confirms and validates the syntactic measurement method.

Chapter 3 extends the work described in Chapter 2 in several ways. First, the SAND1 MDS dialect map based on a syntactic measure is also compared with the *Heeringa (2004)* map of the Dutch dialects based on *pronunciational differences*. A visual comparison between the syntactic map and the pronunciational map shows that the maps correspond to a reasonable extent, even though the syntactic map shows a less smooth colour continuum overall. Second, *geographical distances* are correlated with syntactic Hamming distances using *regression analyses* to investigate how much of the recorded syntactic variation can be accounted for by geographical distance. Regression analyses based on an optimal cross-section of 21 dialect varieties and on all 267 dialect varieties show that 56 percent and 30 percent of syntactic distance can be explained with geographical distance in a linear relationship, respectively. To put these percentages into a broader perspective, 30 percent may be considered a relatively large amount of explainable variation when compared to the 6 percent of syntactic variation which can be explained by *population sizes* (Heeringa et al., 2007). The remaining 70 percent of syntactic variation unexplained by geography should be explainable with other linguistic, social, cultural and political factors. Figure 6-1 visualises the hypothetical case of a 100 percent correlation between syntactic and geographical distances on an MDS dialect map to illustrate the results. The result is a perfect colour *continuum map* without any dialect borders. Contrastingly, an example of an MDS dialect *mosaic map* is shown in Figure 6-2. This map illustrates the visual result of a low correlation between syntactic and geographical distances. Third, the chapter presents measurement results based on binary comparisons between *feature variables*, which are formulated by manually annotating syntactic variables with *linguistic feature information*. The measurement results based on defined feature variables are compared with the results based on the observed atomic variables for the *reflexives* subdomain in SAND1. The geographical distributions seem nearly identical after application of the MDS procedure. The visual resemblance is confirmed by the results of a regression analysis. Application of the *local incoherence* validation method suggests that atomic variable distances may somewhat better reflect local conditioning of dialect differences than distances based on the current set of feature variables. Section 6.3.1 proposes a refined measurement method based on linguistic feature information to more accurately model syntactic differences.



Figure 6-1: A perfect MDS dialect continuum map resulting from a 100 percent correlation between syntactic and geographical distances.



Figure 6-2: An example of an MDS dialect mosaic map resulting from a low correlation between syntactic and geographical distances.

Chapter 4 measures the degrees of association among *aggregate pronunciational, lexical and syntactic differences*. This joint research—in collaboration with Wilbert Heeringa and John Nerbonne—quantifies lexical and syntactic differences at a nominal level using the *gewichteter Identitätswert* (GIW) method—a frequency-weighted similarity measure—and measures pronunciational differences numerically using the *Levenshtein distance* metric. It examines the subset of 70 common Dutch dialect varieties in two data sources: the *Reeks Nederlandse Dialectatlassen* (RND; ‘Series of Dutch Dialect atlases’; Blancquaert and Peé, 1925–1982) and SAND1. The RND data are used to measure both pronunciational and lexical distances; the SAND1 data are used to measure syntactic distances. The chapter presents colour maps of the Dutch dialect areas based on pronunciational, lexical and syntactic differences in pairwise comparisons to provide a general impression of the associations between the pronunciational, lexical and syntactic levels. The colour maps employ the MDS procedure to visualise the variation in the Dutch language area geographically. *Cronbach’s alpha* consistency coefficients are calculated to determine the minimum reliability of the distance measurements when applied to the data sources. The correlation coefficients among the distance measurements for the three linguistic levels are calculated as a measure of the degree to which the three linguistic levels are associated. Since regression analyses clearly show that geography influences each of the three linguistic levels separately, the correlations between all linguistic levels are recalculated in *multiple regression analyses* to filter out geography as an underlying factor of influence. These analyses result in modest degrees of association among the linguistic levels. However, the measured association levels are substantial and may reflect typological constraints between syntactic and phonological structure, which would be very interesting.

Chapter 5 investigates a *data mining* technique to discover relevant associations between 485 syntactic variables in SAND1 using a rule induction system based on *proportional overlap*. The method of *association rule mining* calculates the proportional overlap between geographical distributions of syntactic variables and incorporates rule quality factors such as *accuracy*, *coverage*, *completeness* and *complexity* to measure the *interestingness* of variable associations. This work restricts itself to the *Piatetsky-Shapiro (1991)* measure of interestingness because of its historical priority and its formulaic simplicity. First, the chapter presents the non-recursive association rule mining algorithm in pseudocode and illustrates the procedure using a minimal subset of the actual SAND1 data. The example procedure exposes the *asymmetric nature* of syntactic variable associations, which may be interpreted as *variable dependencies* with potentially hierarchical implications. Then, the association rule mining method is applied to 485 syntactic variables in 267 Dutch dialects in SAND1. Finally, the exploratory review of the results discusses the *highest ranked association rules* with and without *variable disjunctions* and also examines an *implicational chain* of variable associations. The results manifest the high degrees of proportional overlap between the geographical distributions of the syntactic variables in SAND1, which effectively reduce the importance of the geographical occurrences in the data set. This observation may facilitate syntactic analyses to ascend from the observational level of geographical distributions to more abstract variable association patterns.

## 6.2. Conclusions in questions and answers

- i. How can dialect borders and dialect continua be differentiated objectively?

*Chapters 2, 3 and 4 describe computational methods to objectively classify syntactic variation in Dutch dialects using several measures of syntactic distance. Multidimensional scaling is applied to analyse and interpret the resulting distance matrices to classify the Dutch dialect areas.*

- ii. How can syntactic variation be measured?

*This research shows that a measure of syntactic distance based on the Hamming distance suffices. This nominal measure is based on binary comparisons of syntactic variables between dialect pairs. The Jaccard distance measure emphasises Hamming distances because this method effectively filters out irrelevant variable comparisons. The gewichteter Identitätswert method incorporates dialect frequency of variable occurrences into the measurement procedure. MDS dialect maps based on these measures of syntactic distance indicate that application of these methods to measure syntactic differences produces comparable results.*

- iii. Is syntactic variation in Dutch dialects structured in a geographically coherent way when viewed in the aggregate?

*The front cover of this dissertation, as well as Figure 6-12, satisfactorily answers this question affirmatively. Although syntactic variation appears in many dimensions, this research demonstrates that aggregate geographical distributions can be represented accurately in merely three dimensions after reduction via multidimensional scaling. This is a computational confirmation of the intuition that syntactic variation is organised in groups of related patterns.*

- iv. To what extent do the geographical distributions of the syntactic variation domains in SAND1 correspond with each other?

*Although Figure 2-2 to Figure 2-5 show that the individual syntactic subdomains in SAND1 have rather different distribution patterns, the main geographical dialect areas are easily discernable. The dialect areas emerge continuously more pronounced and robust as more syntactic differences are aggregated.*

- v. To what extent does the objective map of the Dutch dialects based on a syntactic measure visually correspond with the traditional dialect map based on perceptual judgements?

*The syntactic and perceptual dialect maps are remarkably similar. The classification of the Dutch dialects in the southern half of both maps is nearly identical, although significant differences are visible as well in the central eastern and central western regions. The syntactic map only reveals a few relatively subtle dialect area borders in the northern half of the map, whereas the perceptual map shows many dialect area borders within this region.*

- vi. What is the nature of the relation between syntactic variation and geographical distance?

*Chapters 2 and 3 demonstrate that there is, in fact, geographical cohesion in syntactic variation when viewed in the aggregate. The regression analysis shown in Figure 3-6 reveals that around 30 percent of syntactic distance can be explained with geographical distance.*

- vii. How can linguistic knowledge be incorporated into a measure of syntactic distance?

*Chapter 3 formulates feature variables to abstract away from the atomic variables as they occur. The idea is to measure differences between dialects at a more structural level which may only be obtained after syntactic analysis. Feature variables can help capture the notion that some variables are less different from each other than other variables.*

- viii. Does incorporating linguistic knowledge into a measure of syntactic distance contribute to more accurate quantifications of syntactic variation?

*The distance measure using feature variables, as described in Section 3.7, yields highly similar results compared to the same measure using atomic variables with respect to syntactic variation in the reflexives domain. Even though these results using feature variables do not directly increase accuracy of the syntactic measure, they do provide new and promising pathways to more accurately quantify syntactic variation. This includes differentiation between dissimilar variable pairs and the inclusion of the number of similarities as well as differences in the syntactic measure. Section 6.3.1 proposes a more refined measure of syntactic distance based on a feature variable hierarchy.*

- ix. To what degree are aggregate pronunciation, lexical and syntactic distances associated with one another when measured among varieties of a single language? Particularly, are syntax and pronunciation more strongly associated with one another than either (taken separately) is associated with lexical distance?

*Chapter four calculates correlation coefficients among the distance measurements for the three linguistic levels as a measure of the degree to which the three linguistic levels are associated. The results in Table 4-6 show that—without controlling for the effect of geography—pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and that syntax is much more strongly associated with pronunciation (42%) than with lexis (25%).*

- x. To what degree are the associations between aggregate pronunciation, lexical and syntactic distances influenced by geography as an underlying factor?

*Table 4-9 shows that the degree of association between pronunciation and lexical distances turns out to be based on geography as an underlying factor for no less than 39%. The association between syntactic and pronunciation distances is even more heavily based on geography as a third factor (46%). The apparent association between syntactic and lexical distances turns out to be principally due to geography as a third factor (63%).*

- xi. Is there evidence for influence among the linguistic levels, even once we control for the effect of geography? Particularly, do syntax and pronunciation more strongly influence one another than either—taken separately—influences or is influenced by lexical distance?

*The effects of linguistic levels on one another—once geography is included as an independent variable—have been measured in multiple regression designs. The results in Table 4-8 show that some influence between pronunciation and syntax (12%) remains after geography as an underlying factor of influence is filtered away, although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (merely 3%). The correlation between pronunciation and syntax might either be explained by typological constraints or other extralinguistic factors.*

- xii. To what extent does the map of the Dutch dialects based on syntactic distances visually correspond with the dialect map based on lexical distances?

*The syntactic and lexical dialect maps are rather dissimilar. The two maps differ in the degree of separation with respect to the Frisian area in the central North. Also, the south-eastern Limburg area on the syntactic map is quite prominently present, whereas this area can hardly be made out on the lexical map.*

- xiii. To what extent does the map of the Dutch dialects based on syntactic distances visually correspond with the dialect map based on pronunciational distances?

*The syntactic and pronunciational dialect maps are partially similar. Although the two maps differ in the degree of separation with respect to the Frisian area in the central North, they do correspond to a certain degree in the southern areas.*

- xiv. How can relevant associations between syntactic variables be discovered?

*Chapter five exhaustively evaluates levels of association between combinations of syntactic variables based on the proportional overlap between their geographical distributions. The application of association rule mining between syntactic variables in this work examines all combinations of syntactic variables to determine which variable subsets most frequently co-occur geographically.*

- xv. Why is it considered important to discover associations between syntactic variables?

*Linguistic research frameworks such as generative syntax and functional typology share a primary interest in understanding the structural similarities and differences between language varieties. The frameworks aim to identify which universal syntactic properties can vary across language varieties and which remain constant. The ultimate goal is to characterise the superficial structural diversity of all language varieties as particular settings of relatively few parametric patterns. Unfortunately, the search for syntactic universals is still very much a topic of ongoing research. This investigation aims to contribute to this global research effort of parameterisation by proposing a computational method to discover syntactic variable associations automatically.*

- xvi. What factors can help determine the quality of an association rule?

*Table 5-6 lists several widely used rule quality evaluation factors: accuracy, coverage, completeness and interestingness. The accuracy of a rule indicates how often a rule is correct. The coverage of a rule expresses how often a rule applies. The factor completeness may be used to explore how much of the target class a rule covers. These three rudimentary interestingness factors are integrated in a measure of rule interestingness. Complexity is another rule quality factor. It is defined as the total number of variable disjuncts in a rule.*

- xvii. What is an example of an interesting association between syntactic variables?

*Table 5-10 shows the most interesting association rule in SAND1. It associates one of the variables in map A on page 46 in SAND1 with a variable in map B on page 38. The rule states that, in the context of a strong plural subject pronoun in second person—i.e. ‘We geloven dat g-lieden niet zo slim zijn als wij’—if the complex pronoun ‘g-lieden’ occurs, then the strong singular subject pronoun in second person ‘gij’ (or ‘gie’) nearly always occurs as well—i.e. ‘Ze gelooft dat gij/gie eerder thuis bent dan ik’. This rule suggests that the plural pronoun ‘g-lieden’ belongs to the same paradigm as the singular pronoun ‘gij’.*

### 6.3. Directions for future research

Due to inevitable time restrictions and the explorative nature of this research, several relevant research strands necessarily remain uncompleted at this time. Sections 4.11 and 5.9 discuss the results, implications and directions for future research with respect to the associations among linguistic levels and the discovery of association rules between syntactic variables, respectively. The current section reviews two other intriguing directions for future research. Section 6.3.1 discusses the development of alternative measures of syntactic distance. Section 6.3.2 explores the results of the incorporation of a preliminary version of the SAND2 data into the measurement procedure and concludes with a preview of the SAND MDS map in Figure 6-12.

#### 6.3.1. Alternative measures of syntactic distance

The previous chapters focused on the Hamming distance and the *gewichteter Identiteitswert* (GIW) method to measure syntactic differences. However, Sections 5.9 and 6.2.ii mention experiments with the Jaccard distance. Section 5.9 also refers to preliminary results based on composite variables. It may be helpful to list the types of syntactic variables and distance measures in order to clarify the combinatory space under investigation. Table 6-1 recapitulates the classification of syntactic variable types as discussed in Section 3.1. For example, the alternative distance measure based on feature variables presented in Section 3.7 applies the same Hamming distance measure as discussed in Section 3.3 to analyse the syntactic variation data from a different perspective. Likewise, Section 5.9 briefly refers to encouraging, experimental results of the application of the Jaccard distance measure to the SAND1 data based on composite variables. Individual variables are combined into a composite variable when the geographical distributions of the individual variables overlap beyond a certain level of accuracy. The composition procedure is covered in Chapter 5. However, it should be noted that the application of the Hamming distance measure to the same set of composite SAND1 variables results in less encouraging results. It would be an interesting direction for future research to investigate the applicability, accuracy and implications of analyses based on composite variables.

Table 6-2 lists a selection of nominal measures of syntactic distance with informal definitions which aim to optimise comparability.<sup>1</sup> All distance measures return normalised values between zero and one. Table 6-2 illustrates the increasing levels of refinement in the measurement formulas. The first method states that the Hamming distance measure straightforwardly divides the number of different variable realisations by the total number of variable comparisons.

---

<sup>1</sup> Section 4.4 notes that a distance measure is considered nominal when the variables under comparison are either equal or unequal.



Table 6-1: A classification of syntactic variable types.

Atomic:	Syntactic variables as they have been recorded, without interpretations. Atomic variables are compared binarily: they are either found or not found in a language variety.
Feature:	Syntactic variables with manually annotated, linguistic feature information obtained after a syntactic analysis. A translation matrix is required to map atomic variables to collections of corresponding feature variables. Feature variables are compared binarily or ternarily: they occur, do not occur, but may also be undefined.
Composite:	Collections of syntactic variables with (nearly) identical geographical distributions. A variable distance matrix based on geographical co-occurrences and a specified threshold value are required to determine whether a collection of variables should be treated as a composite variable. Composite variables are compared according to the type of the individual variables in the variable collection.

Table 6-2: Definitions of a selection of nominal measures of syntactic distance.

Hamming:	$\frac{\text{diff}(A,B)}{n}$	The number of variables which occur in only one of the two dialects (i.e. $\text{diff}(A,B)$ ) divided by the total number of variable comparisons (i.e. $n$ ).
Fractional (Jaccard):	$\frac{\text{diff}(A,B)}{\text{diff}(A,B) + \text{ident}(A,B)}$	The number of variables which occur in only one of the two dialects (i.e. $\text{diff}(A,B)$ ) divided by the total number of variables which occur in at least one of the two dialects (i.e. $\text{diff}(A,B) + \text{ident}(A,B)$ ).
Frequency-weighted (GIW):	$\frac{\text{diff}(A,B) + \sum(\text{freq}(i)/m)}{\text{diff}(A,B) + \text{ident}(A,B)}$	The number of variables which occur in only one of the two dialects (i.e. $\text{diff}(A,B)$ ), plus the summation of the number of geographical occurrences of each variable occurring in both dialects (i.e. $\sum(\text{freq}(i))$ ) divided by the total number of dialects (i.e. $m$ ). The resulting (fractional) number is divided by the total number of variables which occur in at least one of the two dialects (i.e. $\text{diff}(A,B) + \text{ident}(A,B)$ ).

The second equation defines the Fractional distance measure as a refinement of the Hamming measurement by only taking into account variables for which there is empirical data in at least one of the dialects under comparison. This method effectively removes redundant variables and results in more pronounced dialect differences. The third method describes a frequency-weighted distance measure—such as the GIW method—which also incorporates the relative frequency of occurrence of each individual variable to further refine the syntactic distance relationships between the dialects.

Table 6-2 states that in the current research context the Jaccard distance measure can be described with the same equation as the Fractional distance measure. However, the measurement types are generally applied to different variable types and in different application domains. The Jaccard distance measure is predominantly applied in research areas such as ecology and biogeography to quantify the proportional overlap between geographical locations based on binary comparisons of variable occurrences. It is more commonly notated as follows:  $1 - |A \cap B| / |A \cup B|$ . If A and B are two sets of dialects then the Jaccard similarity index is calculated by dividing the number of variables which occur in both dialects (i.e. the intersection set  $A \cap B$ ) by the number of variables which occur in either dialect (i.e. the union set  $A \cup B$ ). The complementary Jaccard distance measure subtracts the similarity index from one. This is functionally equivalent to the Fractional distance measure as defined in Table 6-2 if undefined variable values are not incorporated. However, the Jaccard measure is predominantly applied to attested occurrences of binary variables, whereas the Fractional method is designed to be applied to abstract variables. Table 6-1 notes that up to three different values may be differentiated with respect to abstract feature variables (yes/no/undefined), whereas only two values are generally required in the context of attested atomic variables (found/not found).

The following example illustrates the differences between the measures of syntactic distance described in Table 6-2. Table 6-3 and Table 6-4 present calculations to demonstrate the syntactic measures based on atomic and feature variables using the sample data from Table 3-3 and Table 3-6, respectively. The feature frequencies in Table 6-4 are derived from the mapping matrix from feature to atomic variables with respect to reflexive pronouns shown in Table 3-5. The frequency-weighted measure results in the highest syntactic distances because every variable comparison—by definition—adds to the accumulative distance. Frequently occurring variables are considered to be less important than infrequently occurring variables. Therefore, the technique *emphasizes* rather than ignores infrequently occurring variables. Section 4.4 provides more information regarding this measurement concept. A visualisation of the typical relation between frequency-weighted syntactic distances and geographical distances is shown in Figure 4-12. The two example calculations in Table 6-3 and Table 6-4 demonstrate two points. First, different measurement techniques may result in rather different syntactic distances between dialects. Note that this observation does not imply that different dialect distances also result in different dialect relationships. Second, it is important to find out what type of variables should be measured in order to optimally quantify the differences between language varieties at the syntactic level. In conclusion, additional research is recommended to further explore and refine measures of syntactic distance.

Table 6-3: Example distance measurements using atomic variables based on Table 3-3.

	Lunteren (A)	Veldhoven (B)	Ident. (i)	Diff. (d)	Freq. (f)
[sand1,68a]: zich	√	√	√		121
[sand1,68a]: hem					112
[sand1,68a]: zijn eigen	√			√	43
[sand1,68a]: zichzelf					2
[sand1,68a]: hemzelf					1
			$i = 1$ $d = 1$		
Variables (n) = 5					
Dialects (m) = 267					

Hamming:  $d / n = 1 / 5 = 0.2$

Fractional:  $d / (d + i) = 1 / (1 + 1) = 1 / 2 = 0.5$

Frequency-weighted:  $d + \sum(f(i)/m) / (d + i) = (1 + (121/267)) / 2 = 1.45 / 2 = 0.73$

Table 6-4: Example distance measurements using feature variables based on Table 3-6.

	Lunteren (A)	Veldhoven (B)	Ident. (i)	Diff. (d)	Freq. (f)
	{zich, zijn eigen}	{zich}			
[sand1,68a]: personal					112
[sand1,68a]: reflexive	√	√	√		122
[sand1,68a]: possessive	√			√	43
[sand1,68a]: ownness	√			√	43
[sand1,68a]: focus					3
			$i = 1$ $d = 2$		
Variables (n) = 5					
Dialects (m) = 267					

Hamming:  $d / n = 2 / 5 = 0.4$

Fractional:  $d / (d + i) = 2 / (2 + 1) = 2 / 3 = 0.66$

Frequency-weighted:  $d + \sum(f(i)/m) / (d + i) = (2 + (122/267)) / 3 = 2.46 / 3 = 0.82$

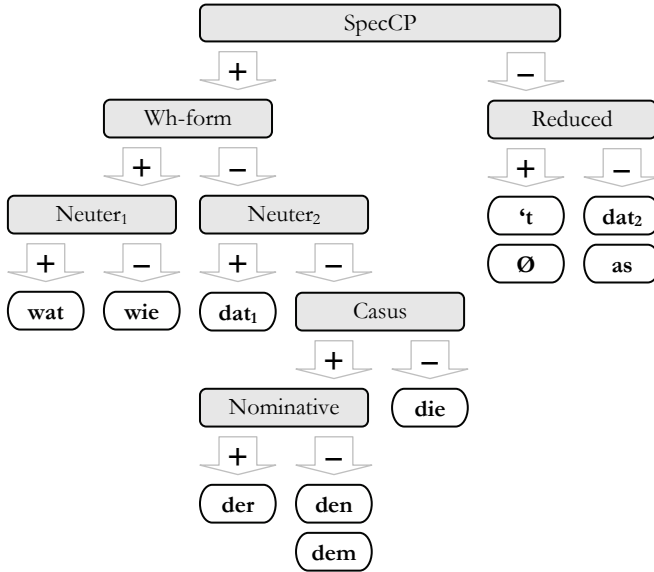


Figure 6-3: Fragment of a feature variable hierarchy with respect to fronting phenomena.

The final topic of this section discusses a proposal for a more refined measure of syntactic distance based on a feature variable hierarchy. Sections 3.10, 6.1 and 6.2.viii already referred to this approach. The original set of binary feature variables with respect to the Reflexives domain is listed in Table 3-5, Table 3-7 and Table 3-8. The current section outlines a more refined method in line with the work of Longobardi et al. (2005) who study the parametric variation of the structure of nominal phrases in 20 languages based on a list of 35 binary linguistic parameters.<sup>2</sup> In their work the syntactic distance between any two languages—described by a unique configuration of syntactic parameters which are formulated in the generative syntax research framework—is expressed by a coefficient derived from the number of identities and differences between the language varieties. Phylogenetic methods are applied to the resulting distance tables to provide historically correct taxonomies of language families. The current proposal implements a syntactic measure based on Fractional distances between hierarchically ordered feature variables with respect to the SAND1 Fronting domain. A fragment of an experimental version of the Fronting feature variable hierarchy is shown in Figure 6-3.

<sup>2</sup> The number of supported languages and syntactic parameters has been expanded to at least 24 languages and 46 syntactic parameters since the cited work.

Table 6-5: The corresponding matrix for the feature variable hierarchy in Figure 6-3.

	SpecCP	Wh-form	Reduced	Neuter <sub>1</sub>	Neuter <sub>2</sub>	Casus	Nominative
wat	+	+		+			
wie	+	+		-			
dat <sub>1</sub>	+	-			+		
die	+	-			-	-	
der	+	-			-	+	+
den/dem	+	-			-	+	-
‘t/∅	-		+				
dat <sub>2</sub> /as	-		-				

The hierarchy is formulated in such a way that each syntactically differentiating atomic variable in the SAND1 Fronting domain translates to a unique feature variable path. Each recorded atomic variable is shown in a white rounded rectangle and each feature variable is shown in a grey rounded rectangle. Elements with subscripts (*Neuter*, *dat*) may occur in several positions within the hierarchy, depending on the linguistic context. The plus or minus sign within each arrow indicates whether the atomic variable (directly below the arrow) contains the feature variable (directly above the arrow). Multiple atomic variables below one arrow—such as *dat* and *as*, or *den* and *dem*—indicate that the different realisations are not explainable at the syntactic level in this model. The variation may be explainable at the lexical or morphological level instead. Although feature variables are binary in nature—indicated by plus and minus arrows—a feature variable is undefined for an atomic variable if it is not in the feature variable path. For example, the feature *Casus* is undefined for the atomic variable *wat*. The feature variable path of *Casus* is [SpecCP<sup>(+)</sup> » Wh-form<sup>(-)</sup> » Neuter<sub>2</sub><sup>(-)</sup>], whereas the path of the atomic variable *wat* is [SpecCP<sup>(+)</sup> » Wh-form<sup>(+)</sup> » Neuter<sub>1</sub><sup>(+)</sup>]. Table 6-5 shows the corresponding matrix for the fragment of the SAND1 Fronting feature variable hierarchy in Figure 6-3. The matrix elements contain one of the following three values. A plus sign indicates that the abstract feature (in the first row) is represented in the atomic variable (in the first column). A minus sign means that the abstract feature is not represented in the atomic variable. An empty slot shows that the feature variable is not applicable to the atomic variable.

Table 6-7 illustrates the Fractional distance measurement based on the feature variable hierarchy with the data sample shown in Table 6-6. If the syntactic variable 1:*die* 2:*as/at/da(t)* occurs in dialect A and not in dialect B in the syntactic context of *short object relative*, and the variable 1:*die* 2:*-t* occurs in dialect B and not in dialect A, then the Fractional distance between dialects A and B is calcu-

lated as follows. The two dialects use the same relative pronoun (*die* ‘who’) but vary with respect to the complementiser position (*dat* ‘that’ versus *-t*). The relative pronoun *die* ‘who’ consists of the feature variables *SpecCP*, *Wh-form*, *Neuter<sub>2</sub>* and *Casus*. A feature variable comparison of *die* ‘who’ with itself results in zero differences and four similarities (i.e. 0/4). However, in the second variable position the complementiser *dat* ‘that’ occurs in dialect A, whereas dialect B chooses *-t* in this syntactic context. The two atomic variables *dat* ‘that’ and *-t* have identical values with respect to the feature variable *SpecCP*, but they differ with respect to the feature variable *Reduced*. A feature variable comparison between these two atomic variables results in one difference and one similarity (i.e.  $1/(1+1) = 1/2$ ). Therefore, the normalised Fractional distance between dialects A and B based on the sample data is  $((0/4 + 1/2) / 2) = 0.25$ .<sup>3</sup>

Table 6-6: Map 84a in SAND1 shows three syntactic variables in the fronting domain.

Context: Short object relative, complementiser following relative pronoun  
 Variables: { \*1:die 2:as/at/da(t), 1:die 2:-t, 1:wie 2:-t }  
 Example: Dat is de man die dat ze geroepen hebben.  
           ‘that is the man who that they called have’  
           “That is the man who they have called.”

Table 6-7: Fractional distance matrix in the short object relative context in Table 6-6, based on the feature variable mapping in Table 6-5.

	1:die 2:as/at/da(t)	1:die 2:-t	1:wie 2:-t
1:die 2:as/at/da(t)		$(0/4 + 1/2) / 2 = 0.25$	$(1/2 + 1/2) / 2 = 0.50$
1:die 2:-t	$(0/4 + 1/2) / 2 = 0.25$		$(1/2 + 0/2) / 2 = 0.25$
1:wie 2:-t	$(1/2 + 1/2) / 2 = 0.50$	$(1/2 + 0/2) / 2 = 0.25$	

A feature variable hierarchy has the potential to measure syntactic differences between language varieties at a more structural level which may only be obtained after syntactic analysis. However, the main difficulty with annotation-based techniques like the current proposal remains in the design of a reasonable feature variable hierarchy which differentiates all atomic variables. The current proposal satisfies the latter condition of variable differentiation for the SAND1 data with respect to the Fronting domain, but in its current experimental form, the proposal is still theoretically weak. It would be an interesting direction for future research to design a ‘theoretically sound’ feature hierarchy. For example, syntactic features could be more closely defined according to the Generative

<sup>3</sup> The Fractional distance is divided by two to obtain the normalised distance between dialects A and B because the comparison between the syntactic variables consists of two linguistic elements.

syntax research framework (cf. Chomsky 1995). In theory, all atomic variables should be differentiated implicitly if the framework becomes able to explain all syntactic variation patterns. In this respect the current proposal also represents a framework to validate syntactic theory.

### 6.3.2. Incorporation of SAND2 data

Sections 1.2.3, 4.11 and 5.9 already mentioned the forthcoming second volume of the Syntactic atlas of the Dutch dialects (SAND2; Barbiers et al., t.a. 2008) as an additional syntactic microvariation data source. SAND2 contains syntactic variation related to verbal clusters, cluster interruption, morphosyntactic variation, the negative particle, and negative concord and quantification. Table 6-8 provides examples of syntactic variables in each of these syntactic domains to indicate the wealth of syntactic variation in SAND2.

Table 6-8: Examples of syntactic variables in context for each syntactic domain/ chapter in SAND2.

*Chapter 1:* Verbal clusters:

Ik weet dat hij weeste zwemmen is.

‘I know that he been swimming is’

*Chapter 2:* Cluster interruption:

Ik denk dat je veel zou weg moeten gooien.

‘I think that you much should away must throw’

*Chapter 3:* Morphosyntactic variation:

Niemand heeft dat ooit willen.

‘nobody has that ever want<sub>infinitive</sub>’

*Chapter 4:* Negative particle:

Els en wil niet zingen.

‘Els [negative particle] wants not sing’

*Chapter 5:* Negative concord and quantification:

Ik heb niemand niet gezien.

‘I have nobody not seen’

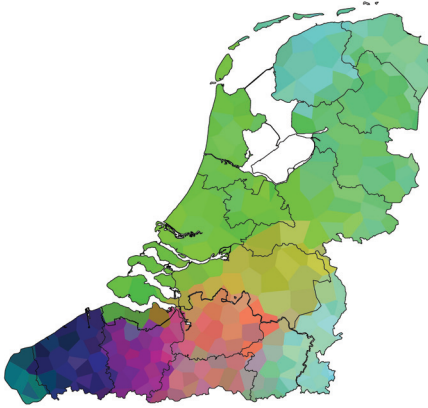


Figure 6-4: SAND1 MDS map visualising 485 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.955$ ).

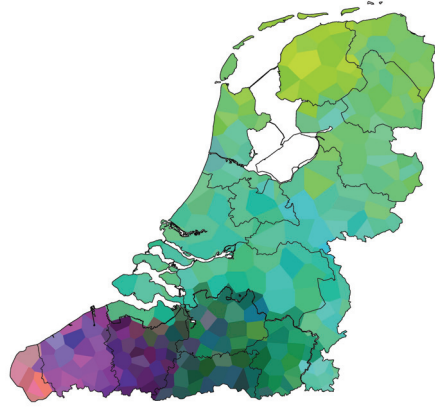


Figure 6-5: Preliminary SAND2 MDS map visualising 697 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.932$ ).

SAND2 consists of 697 syntactic variables in 83 syntactic contexts, whereas SAND1 contains 485 variables in 106 contexts. Roughly speaking, SAND2 focuses on syntactic variation related to the right periphery of the clause, whereas SAND1 mainly documents syntactic variation related to the left periphery of the clause. Therefore, it seems reasonable to assume that aggregating the SAND2 variables into the overall SAND data set will result in more balanced measurements of syntactic variation. This, in turn, should provide more accurate quantitative perspectives on syntactic variation in Dutch dialects.

Figure 6-5 shows a preliminary SAND2 MDS dialect map. It visualises 697 syntactic variables in SAND2 based on the Hamming distance measure as described in Section 2.4. It is important to note that the results are based on a pre-final version of the SAND2 data.<sup>4</sup> Therefore, additions, removals and modifications to the syntactic data are to be expected. However, since most of the SAND2 data have already been analysed and verified at the time of this writing, the version of the SAND2 data presented below may be assumed to be adequately robust for an exploratory visual analysis at high levels of aggregation. The SAND1 MDS map in Figure 6-4 is shown next to the SAND2 MDS map in Figure 6-5 to facilitate a visual comparison.<sup>5</sup>

<sup>4</sup> The SAND2 data snapshot was taken at September 17, 2007. Syntactic variation data related to correlation and summary maps are not included in the snapshot.

<sup>5</sup> The three colour components red, green and blue are arbitrarily assigned to the first three dimensions of the MDS solutions which are visualised in the dialect maps. Therefore, the colours differ in meaning in the two maps. For example, the yellow shading in the map in Figure 6-4 found in Noord-Brabant does not illustrate the same patterns of syntactic variation as the yellow



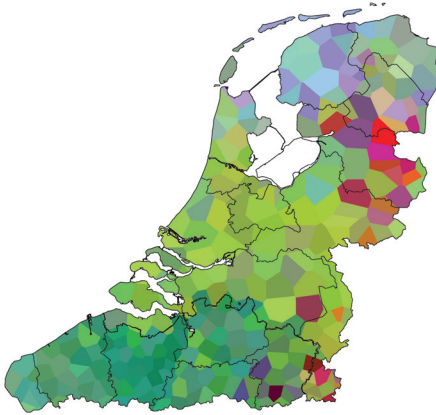


Figure 6-6: Preliminary MDS map visualising 149 syntactic variables related to verbal clusters based on a Hamming distance measure ( $r = 0.894$ ).



Figure 6-7: Preliminary MDS map visualising 231 variables related to morphosyntactic variation based on a Hamming distance measure ( $r = 0.879$ ).

At first sight the Dutch dialect area classifications on the SAND1 and SAND2 MDS maps in Figure 6-4 and Figure 6-5 seem similar to a certain extent. The south-western dialect areas are nearly identical on both maps and largely correspond with the political borders of French Flanders and the Belgian provinces. Also, the central-northern Frisian area (in blue on the SAND1 map) can be identified on both maps. A number of differences between the maps are noticeable as well. For example, a subtle difference visible on the SAND2 map is the existence of a small transitional area in the north-eastern (West-Frisian) region of the Noord-Holland province. The MDS maps in Figure 6-6, Figure 6-7 and Figure 6-8 visualise syntactic variation in SAND2 related to verbal clusters, morphosyntax, and negative concord and quantification, respectively. The maps provide less general perspectives on syntactic variation than the overall SAND2 map shown in Figure 6-5. These more detailed maps can be used to determine which syntactic variation subdomains are responsible for the correspondence between the transitional area and the Frisian varieties. Figure 6-6 shows that the transitional area most prominently corresponds with the Frisian varieties in the context of verbal clusters. Figure 6-7 indicates that the transitional area also exists at a morphosyntactic level, although less prominently. The area is nearly invisible in the context of negative concord and quantification phenomena, as shown in Figure 6-8. It doesn't exist at all in the SAND2

---

shading in the map in Figure 6-5 found in Friesland. This explains why the French Flanders area in the far South-West should be classified as nearly identical on both maps. The colour difference between the shades of green-blue colours and the dark-blue neighbouring region on the SAND1 map seems comparable to the colour difference between the shades of pink and the purple neighbouring area on the SAND2 map.

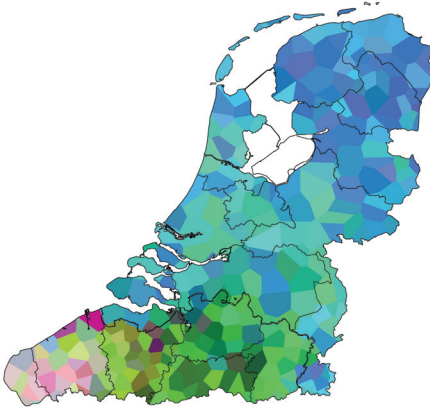


Figure 6-8: Preliminary MDS map visualising 186 syntactic variables related to negative concord and quantification phenomena based on a Hamming distance measure ( $r = 0.919$ ).

subdomains of verbal cluster interruption and negative particle variation (not shown).

Returning to the Dutch dialect area classifications on the SAND1 and SAND2 MDS maps shown in Figure 6-4 and Figure 6-5, the most fascinating difference between the two maps is arguably the complete disappearance of the Noord-Brabant and Nederlands Limburg dialect areas in yellow-brown and light-blue, respectively, on the SAND2 map. The two dialect areas are clearly visible on the SAND1 map. Figure 6-6, Figure 6-7 and Figure 6-8 show that especially the Noord-Brabant province in yellow-brown on the SAND1 map does not exist at all in any of the five SAND2 subdomains. Interestingly, the northern border of the yellow-brown Noord-Brabant area closely corresponds with the traditional Catholic-Protestant boundary as documented by Van Heek (1954). Manni, Heeringa and Nerbonne (2006) note a strong correlation with Dutch surname diversity and suggest that this religious distinction may have acted as a social boundary, thus increasing surname differences between populations on the border's sides. However, they could not find linguistic evidence of such a separation at the pronunciation level (the pronunciation MDS dialect map is shown in Figure 4-6). The SAND1 MDS map in Figure 6-4 shows that linguistic correspondences with this social boundary exist at the syntactic level. It would be an interesting direction for future research to investigate correspondences between linguistic and demographic boundaries in more detail. This type of research may lead to a deeper understanding of the role of local migrations and cultural diffusion in language variation.

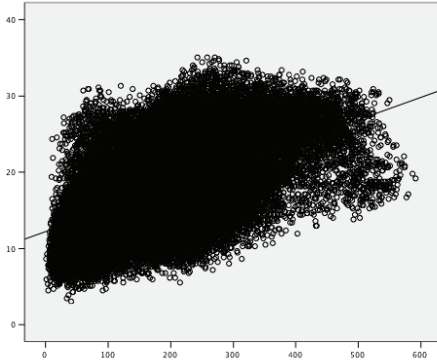


Figure 6-9: SAND1 Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.553$ ).

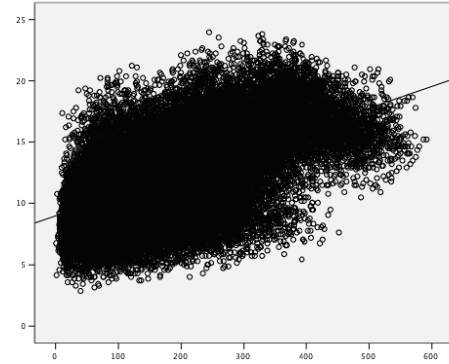


Figure 6-10: SAND2 Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.552$ ).

The final visual distinction between the SAND1 and SAND2 maps discussed in this section are the somewhat smoother colour continua within the dialect areas on the SAND1 map. Furthermore, the SAND2 MDS solution in three dimensions results in a correlation coefficient of  $r = 0.932$ , whereas the comparable SAND1 solution produces a slightly better MDS correlation coefficient of  $r = 0.955$ . These two indicators might suggest a slightly more complex relation between geography and syntactic variation in verbal clusters, negation and quantification. However, a regression analysis between SAND2 Hamming distances and geographical distances results in a correlation coefficient of  $r = 0.552$ . This means that geographical distances can explain as much syntactic variation in SAND2 as in SAND1 ( $r = 0.553$ ). A regression analysis using a logarithmic transformation results in a lower correlation of  $r = 0.521$ . This confirms the conclusion in Section 3.6.3 that the relationship between syntactic and geographical distances can more accurately be described with a linear function than with a logarithmic transformation. The SAND2 regression analysis plot is shown in Figure 6-10 next to the SAND1 regression analysis plot in Figure 6-9 to facilitate a visual comparison. Also, in the spirit of Chapter 4, the association between the final SAND1 and preliminary SAND2 data domains was calculated as well. The regression analysis between the SAND1 and SAND2 distances results in a correlation coefficient of  $r = 0.459$  ( $r^2 = 0.21$ ), which means that 21 percent of the syntactic variation in SAND1 can be explained with variation in SAND2, and vice versa. The residual analysis procedure described in Section 4.9 filters out the influence of geography and produces a correlation coefficient of  $r = 0.401$ . It indicates that 16 percent of the syntactic variation in SAND1 can be explained with variation in SAND2 without geography as an underlying factor of influence, and vice versa. Although these results are based

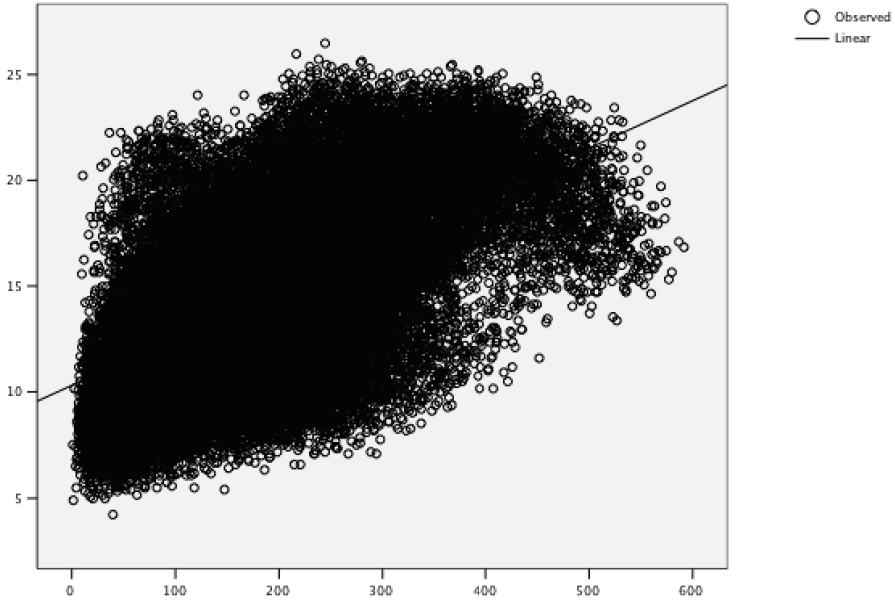


Figure 6-11: SAND Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.592$ ).

on preliminary data, the correlation coefficients might suggest that the SAND1 and SAND2 data domains describe different patterns of syntactic variation. It seems reasonable to assume that the removal of geographical influences from the raw correlation between SAND1 and SAND2 syntactic distances ( $r = 0.459$ ) would result in a much lower coefficient than  $r = 0.401$  if geography were a major underlying factor for the correspondences between the SAND1 and SAND2 geographical patterns. Geography only influences the association between syntactic variation in the SAND1 and SAND2 domains as an underlying structuring factor for less than 13 percent.<sup>6</sup> This is in rather sharp contrast with the results presented in Section 4.9 which showed the major role of geography as an underlying, structuring factor with respect to the associations between the linguistic levels. Of course, this result is to be expected, since the SAND1 and SAND2 domains describe language variation patterns within the same linguistic level, which is widely assumed to be structured by one uniform set of grammatical rules (cf. Chomsky, 1995). Based on the correlation analyses it may be expected that the joint analyses of the two syntactic data sources, described in the next paragraphs, will result in more accurate quantifications of syntactic variation in Dutch dialects.

<sup>6</sup> Applying the formula in Figure 4-14, which calculates the influence of geography underlying the associations between two linguistic levels, results in:  $(1 - (0.401 / 0.459)) * 100 = 12.6$  percent.

The final topic of this dissertation presents and discusses the preliminary SAND MDS dialect map. This classification of the Dutch dialect area aggregates all 1182 syntactic variables in SAND1 and the preliminary version of SAND2. The SAND MDS dialect map is shown in Figure 6-12 and is based on the Hamming distance measure. The high MDS correlation coefficient ( $r = 0.954$ ) indicates that the colours on the SAND dialect map accurately represent the syntactic variation in both volumes of the syntactic atlas. A regression analysis was performed to analyse the relation between the SAND Hamming distances and the geographical distances. The resulting correlation coefficient of  $r = 0.592$  indicates that geographical distances can explain 35 percent of the SAND Hamming distances. A regression analysis using a logarithmic transformation results in a lower correlation ( $r = 0.568$ ), once again confirming the conclusion in Section 3.6.3 that a linear function better describes the relationship between syntactic and geographical distances than a logarithmic transformation. The SAND regression analysis plot is shown in Figure 6-11.

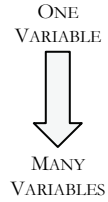
The visual representation of the Dutch dialect relationships in the SAND MDS colour map in Figure 6-12 shows that the differences between the SAND1 and SAND2 maps in Figure 6-4 and Figure 6-5 level out, as one would expect with an application of an additive measurement method. The dialect area differences between the two syntactic variation domains merge into a more harmonious dialect colour map at the highest level of aggregation. The most fascinating difference between the two maps nicely illustrates this quantitative property. As described earlier in this section, the Noord-Brabant and Nederlands Limburg dialect areas in yellow-brown and light-blue on the SAND1 map do not exist on the SAND2 map. The most general perspective on syntactic variation in the Dutch dialect area in Figure 6-12 shows that combining the syntactic variables in the SAND1 and SAND2 data domains for joint quantitative analysis based on the Hamming distance measure results in much less pronounced Noord-Brabant and Nederlands Limburg dialect areas in yellow-green and light-green, respectively. It should be noted that the SAND2 data carry more weight in the Hamming distance calculations because SAND2 contains 30 percent more syntactic variables than SAND1.<sup>7</sup> It would be an interesting direction for future research to investigate the results and implications of alternative measures of syntactic distance when applied to the entire SAND data set, such as the *gewichteter Identitätswert* method and the Jaccard distance in combination with either atomic, feature or composite variables. This type of research could also revisit the associations among the various linguistic levels and quantify the influence of the syntactic variation domains in SAND1 and SAND2.

---

<sup>7</sup> SAND2 contains  $(1 - (485 / 697)) * 100 = 30.4$  percent more variables than SAND1.

Table 6-9: *Visualisation perspectives on syntactic variation in Dutch dialects.*

- I. Symbolic representations of individual syntactic variables (*Figure 1-9*).
- II. Mosaic-like distributions of dialect varieties (*Figure 6-2*).
- III. Groups of geographically coherent patterns (*Figure 6-4*).
- IV. A continuum of geographical patterns (*Figure 6-12*).



Finally, this dissertation has visualised the relation between geography and syntactic variation in Dutch dialects at various levels of aggregation. Table 6-9 summarises the main visualisation perspectives in increasing degrees of data generalisation. At the first level, symbolic representations of individual syntactic variables visualise the relation between each individual syntactic variable and geography. In other words, this qualitative perspective does not present syntactic variation in the aggregate. For example, the map in Figure 1-9 uses a separate colour symbol for each of the seven syntactic variables. The second level of aggregation often arises when a relatively small number of variables are included in the measurement procedure. Mosaic-like distributions of dialect varieties, such as shown in Figure 2-5, often reveal a low correlation between syntactic and geographical distances. However, such distributions may also indicate that either the number of variables or the average inter-correlation between the variables is too small. Therefore, a mosaic-like distribution of dialect varieties may often be predictable by calculating Cronbach's alpha values. As a rule of thumb, relatively low values should correspond with mosaic-like MDS dialect maps. At the third visualisation level, groups of geographically coherent patterns emerge from the mosaic-like variation patterns. The SAND1 dialect map in Figure 2-6 convincingly shows that geographically coherent patterns arise at higher levels of aggregation, even though the dialect areas in the SAND1 sub-domains (shown in Figure 2-2 to Figure 2-5) classify several regions differently. The fourth and final level visualisation perspective on syntactic variation arises when the number of included variables becomes so large that differences between geographically coherent patterns tend to level out. This effect can be derived from the differences between the SAND1 and SAND2 maps in Figure 6-4 and Figure 6-5, respectively. Several differences on these maps fade away at the most general perspective on syntactic variation in Dutch dialects as shown on the SAND MDS map in Figure 6-12. The SAND map may be best described as a continuum of geographical patterns. Given the fact that the pronunciational MDS map of the Dutch dialects in Figure 4-6 shows an even smoother dialect continuum than the syntactic MDS map in Figure 6-12, it would be an interesting direction for future research to investigate the extent to which the generalisation perspectives in Table 6-9 remain applicable when language variation data from several linguistic levels are jointly analysed and visualised in MDS dialect maps, based on the final and complete version of the

SAND data set. This line of research could very well be explored in conjunction with a study into potential associations between variables among linguistic levels such as syntax and phonology, as mentioned in Section 5.9.

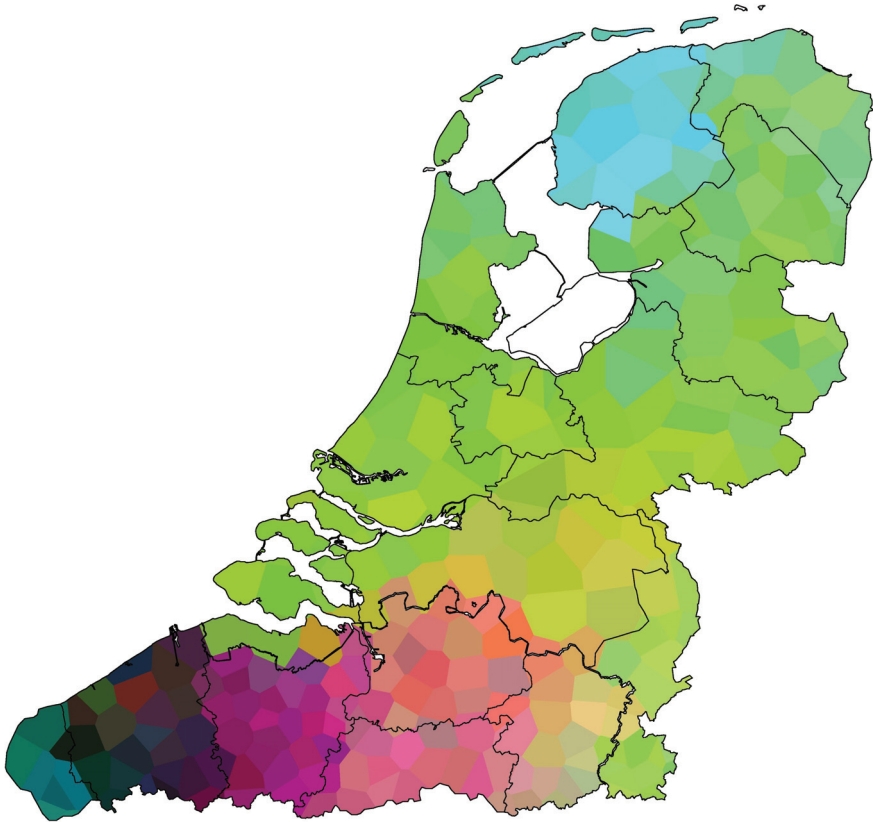


Figure 6-12: SAND MDS map visualising 1182 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.954$ ).





## References

- Agrawal, R., Imielinski, T., Swami, A., 1993. *Mining association rules between sets of items in large databases*. In: Buneman, P., Jajodia, S. (eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, Washington, D.C., 207–216.
- Barbiers, S., Cornips, L., 2001. *Introduction to Syntactic Microvariation*. In: Barbiers, S., Cornips, L., Kleij, S. van der (eds), *Syntactic microvariation*, Electronic Publication, Meertens Instituut, Amsterdam, 1–11.  
<http://www.meertens.knaw.nl/books/synmic/introduction.pdf>.
- Barbiers, S., Bennis, H., 2003. *Reflexives in Dutch Dialects*. In: Koster, J., Riemsdijk, H. van (eds), *Germania et alia: a Linguistic Webschrift for Hans den Besten*, Universiteit Groningen, Groningen, 25–44.
- Barbiers, S., Bennis, H., 2004. *Reflexieven in dialecten van het Nederlands. Chaos of structuur?*. In: Caluwe, J. de, Schutter, G. de, Devos, M. et al. (eds), *Schatbewaarder van de taal. Johan Taeldeman. Liber Amicorum*. Academia Press Gent en Vakgroep Nederlandse Taalkunde Universiteit Gent, Gent, 43–58.
- Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M. van der (eds), 2005. *Syntactic Atlas of the Dutch Dialects*, Volume 1. Amsterdam University Press, Amsterdam.
- Barbiers, S., 2005. *Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics*. In: Cornips, L., Corrigan, K. (eds), *Syntax and Variation. Reconciling the Biological and the Social. Current Issues in Linguistic Theory 265*, John Benjamins, Amsterdam/Philadelphia, 233–264.
- Barbiers, S., Koenenman, O., Lekakou, M., t.a. 2007. *Syntactic doubling and the structure of chains*. *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Cascadia Proceedings Project, Somerville MA.
- Barbiers, S., Cornips, L., Kunst, J., 2007. *The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas*. In: Beal, J., Corrigan, K., Moisl, H. (eds), *Creating and digitizing language corpora, Volume 1, Synchronic databases*, Palgrave Macmillan, Hampshire, 54–90.
- Barbiers, S., Bennis, H., Vogelaer, G. de, Auwera, J. van der, Ham, M. van der (eds), t.a. 2008. *Syntactic Atlas of the Dutch Dialects*, Volume 2. Amsterdam University Press, Amsterdam.
- Blancquaert, E., Peé, W. (eds), 1925-1982. *Reeks Nederlands(ch)e dialectatlassen*. De Sikkel, Antwerpen.
- Bolognesi, R., Heeringa, W., 2002. *De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten*. In: Bakker, D., Sanders, T., Schoonen, R., Wijst, P. van der (eds), *Gramma/TIT: tijdschrift voor taalwetenschap*, Nijmegen University Press, Nijmegen, Volume 9(1), 2002, 45–84.
- Bree, C. van, 1992. *The stability of language elements, in present-day eastern Standard-Dutch and eastern Dutch dialects*. In: Leuvensteijn, J. van en Berns, J., *Dialect and Standard language [...] in the English, Dutch, German and Norwegian language areas*, Amsterdam, 178–203.

- Bree, C. van, 1994. *The development of so-called Town Frisian*. In: Bakker, P. and Mous, M. (eds), *Mixed Languages. 15 Case Studies in Language Intertwining*, Studies in Language and Language Use, Volume 13, IFOTT Amsterdam, 69–82.
- Bucheli, C., Glaser, E., 2002. *The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems*. In: Barbiers, S., Cornips, L., Kleij, S. van der (eds), *Syntactic microvariation*, Electronic Publication, Meertens Instituut, Amsterdam, 41–74.  
<http://www.meertens.knaw.nl/books/synmic/pdf/buch-glas.pdf>.
- Cavalli-Sforza, L., Wang, W., 1986. *Spatial distance and lexical replacement*. In: *Language*, Volume 62(1), 38–55.
- Chambers, J. and Trudgill, P., 1998. *Dialectology*. Cambridge University Press, Cambridge, Second edition.
- Chomsky, N., Halle, M., 1968. *The Sound Pattern of English*. Harper & Row, New York.
- Chomsky, N., 1995. *The Minimalist program*. The MIT Press, Cambridge (MA).
- Cichocki, W., 2006. *Geographic variation in Acadian French /r/: What can correspondence analysis contribute toward explanation?*. In: Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing*, special issue on Progress in Dialectometry: Toward Explanation, Volume 21(4), Oxford University Press, Oxford, 529–541.
- Cornips, L., Jongenburger, W., 2001. *Elicitation techniques in a Dutch syntactic dialect atlas project*. In: Broekhuizen, H., Wouden, T. van der (eds), *Linguistics in the Netherlands*, 2001, John Benjamins, Philadelphia/Amsterdam, 53–63.
- Cronbach, L., 1951. *Coefficient alpha and the internal structure of tests*. In: *Psychometrika*, Volume 16, 297–334.
- Daan, J., Blok, D., 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Volume XXXVII, Amsterdam.
- Donegan, P., Stamp D., 1983. *Rhythm and the holistic organization of language structure*. In: Richardson, J.F. et al. (eds), *Papers from the Parasession on the interplay of phonology, morphology and syntax*. Chicago Linguistic Society, Chicago, 337–353.
- Frawley, W., Piatetsky-Shapiro, G., Matheus, C., 1992. *Knowledge discovery in databases: An overview*. *AI Magazine*, Volume 13, 213–228.
- Freitas, A., 1999. *On rule interestingness measures*. *Knowledge-based Systems*, Volume 12, 309–315.
- Gemert, I. van, 2002. *Het geografisch verklaren van dialectafstanden met een geografisch informatie-systeem (GIS)*. Master's thesis, Rijksuniversiteit Groningen, Groningen.
- Gianollo, C., Guardiano, C., Longobardi, G., t.a. 2007. *Three fundamental issues in parametric linguistics*. In: Biberauer, T. (ed), *The Limits of Syntactic Variation*, John Benjamins, Philadelphia/Amsterdam.
- Ginneken, J. van, 1913. *De sociologische structuur der Nederlandsche taal*. Handboek der Nederlandsche taal, Volume I, Malmberg, Nijmegen.
- Goebel, H., 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Volume 157, Philosophisch-Historische Klasse

- Denkschriften. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.  
With assistance of Rase, W., Pudlatz, H.
- Goebel, H., 1984. *Dialektometrische Studien: Anband italommanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Volume 3, Max Niemeyer, Tübingen.
- Goebel, H., 2006. *Recent advances in Salzburg dialectometry*. In: Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation*, Volume 21(4), Oxford University Press, Oxford, 411–435.
- Goeman, A., 1989. *Dialectes et jugements subjectifs des locuteurs. Quelques remarques de méthode à propos d'une controverse*. In: *Espaces romanes. Etudes de dialectologie et de sociolinguistique offertes à Gaston Tuillon*, Volume II, Grenoble, 532–544.
- Goeman, A., 2000. *Perception of Dialect Distance: Standard and Dialect in Relation to New Data on Dutch Varieties*. In: Long, D., Preston, D. (eds), *Handbook of perceptual dialectology*, Volume II, 2000, John Benjamins, 137–151.
- Gooskens, C., 2004. *Norwegian Dialect Distances Geographically Explained*. In: Gunnarson, B., Bergström, L., Eklund, G., Fridella, S., Hansen, L., Karstadt, A., Nordberg, B., Sundgren, E., Thelander, M. (eds), *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2*, Uppsala, Sweden: Uppsala University, 195–206.
- Goossens, J. 1977. *Inleiding tot de Nederlandse dialectologie*. Wolters-Noordhoff, Groningen.
- Goossens, J., Taeldeman J., Verleyen, G., 1998. *Fonologische atlas van de Nederlandse dialecten*, Volume I, Het korte vocalisme. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.
- Hand, D., Mannila, H., Smyth, P., 2001. *Principles of Data Mining*. The MIT Press, Cambridge, MA.
- Haspelmath, M., t.a. 2007. *Parametric versus functional explanations of syntactic universals*. In: Biberauer, T., Holmberg, A. (eds), *The Limits of syntactic variation*, Benjamins, Amsterdam.
- Heek, F. van, 1954. *Het geboortenniveau der Nederlandse Rooms-Katholieken*. Leiden.
- Heeringa, W., 2001. *De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen*. In: Hoeksema, J. et al. (eds), *TABU: Bulletin voor Taalwetenschap*, Volume 31(1/2). Rijksuniversiteit Groningen, Groningen, 61–103.
- Heeringa, W., Nerbonne, J., 2001. *Dialect Areas and Dialect Continua*. In: Sankoff, D. (eds), *Language Variation and Change*, Volume 13, Cambridge University Press, New York, 375–400.
- Heeringa, W., Gooskens, C., 2003. *Norwegian Dialects Examined Perceptually and Acoustically*. In: Nerbonne, J., Kretzschmar, W. (eds), *Computers and the Humanities*, Kluwer Academic Publishers, Dordrecht, Volume 37, Number 3, 293–315.
- Heeringa, W., 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis Rijksuniversiteit Groningen, Groningen. <http://irs.ub.rug.nl/ppn/258438452>.
- Heeringa, W., Nerbonne, J., Bezooijen, R. van, Spruit, M., 2007. *Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied*. In: *Nederlandse Taal- en Letterenkunde*, Volume 123(1), Uitgeverij Verloren, Hilversum, 70–82.  
<http://marco.info/pro/pub/hnbs2007tntl.pdf>.

- Helke, M., 1970. *The Grammar of English Reflexives*. Doctoral Dissertation, MIT, Cambridge.
- Hoppenbrouwers, C., Hoppenbrouwers, G., 1988. *De featurefrequentiemethode en de classificatie van Nederlandse dialecten*. TABU: Bulletin voor taalwetenschap, Volume 18(2), 51–92.
- Hoppenbrouwers, C., Hoppenbrouwers, G., 2001. *De indeling van de Nederlandse streektaalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Koninklijke Van Gorcum B.V., Assen.
- Horn, H., 1966. *Measurement of overlap in comparative ecological studies*. The American Naturalist, Volume 100, 419–424.
- Hussain, F., Liu, H., Suzuki, E., Lu, H., 2000. *Exception rule mining with a relative interestingness measure*. In: Terano, T., Liu, H., Chen, A. (eds), Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag, London, UK, 86–97.
- Jaccard, P., 1901. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. In: Bulletin de la Societe Vaudoise des Sciences Naturelles, Volume 37, 547–579.
- Jellinghaus, H., 1892. *Die niederländischen Volksmundarten: Nach den Aufzeichnungen der Niederländer*. In: Norden, H. (ed), Forschungen/Verein für Niederdeutsche Sprachforschung, Soltau.
- Kessler, B., 1995. *Computational dialectology in Irish Gaelic*. In: Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL, Dublin, 60–67.
- Lenca, P., Meyer, P., Vaillant, B., Lallich, S., 2008. *On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid*. In: European Journal of Operational Research, Elsevier, Volume 184(2), 610–626.
- Levenshtein, V., 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. In: Cybernetics and Control Theory, Volume 10(8), 707–710.
- Longobardi, G., Guardiano, C., 2005. *Parametric Comparison and Language Taxonomy*. In: Battlori, M., Hernanz, M., Picallo, C., Roca, F. (eds), Grammaticalization and Parametric Variation, Oxford University Press, 149–174.
- Manly, B., 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, Second edition.
- Manni, F., Heeringa, W., Nerbonne, J., 2006. *To what Extent are Surnames Words? Comparing Geographic Patterns of Surnames and Dialect Variation in the Netherlands*. In: Nerbonne, J., Kretzschmar, W. (eds), Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation, Volume 21(4), Oxford University Press, Oxford, 507–528.
- Mantel, N., 1967. *The detection of disease clustering and a generalized regression approach*. In: Cancer Research, Volume 27, 209–220.
- McGarry, K., 2005. *A survey of interestingness measures for knowledge discovery*. The Knowledge Engineering Review, Volume 20, 39–61.
- Nerbonne, J., Heeringa, W., Hout, E. van den, Kooi, P. van de, Otten, S., Vis, W. van de, 1996. *Phonetic Distance between Dutch Dialects*. In: Durieux, G., Daelemans, W., Gillis, S. (eds), CLIN VI: Proceedings of the Sixth CLIN Meeting, Antwerp, Centre for Dutch Language and Speech (UIA), 185–202.

- Nerbonne, J., Heeringa, W., 1998. *Computationale vergelijking en classificatie van dialecten*. Taal en Tongval, Tijdschrift voor Dialectologie, Volume 50, Number 2, 164–193.
- Nerbonne, J., Heeringa, W., Kleiweg, P., 1999. *Edit Distance and Dialect Proximity*. In: Sankoff, D., Kruskal, J. (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Press, Stanford, v–xv.
- Nerbonne, J., Kretzschmar, W., 2003. *Introducing Computational Methods in Dialectometry*. In: Nerbonne, J., Kretzschmar, W. (eds), *Computational Methods in Dialectometry*, Special issue of *Computers and the Humanities*, Volume 37(3), 245–255.
- Nerbonne, J., Kleiweg, P., 2007. *Toward a Dialectological Yardstick*. In: *Journal of Quantitative Linguistics*, Volume 14(2), Routledge, New York, 148–167.
- Nerbonne, J., Heeringa, W., t.a. 2007. *Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation*. In: Featherston, S., Sternefeld, W. (eds), *Roots: Linguistics in Search of its Evidential Base*, Mouton De Gruyter, Berlin.
- Newmeyer, F., 2005. *Possible and probable languages: a generative perspective on linguistic typology*. Oxford University Press, Oxford.
- Norušis, M., 1997. *SPSS Professional Statistics 7.5*. SPSS Inc, Chicago.
- Nunnally, J., 1978. *Psychometric Theory*. McGraw-Hill, New York.
- Oostendorp, M. van, 2006. *Expressing inflection tonally*. *Catalan Journal of Linguistics*, Volume 4(1), 107–126.
- Piatetsky-Shapiro, G., 1991. *Discovery, analysis and presentation of strong rules*. In: Piatetsky-Shapiro, G., Frawley, W. (eds), *Knowledge Discovery in Databases*, AAAI/MIT Press, 229–248.
- Postma, G., 1997. *Logical entailment and the possessive nature of reflexive pronouns*. In: Bennis, H., Pica, P., Rooryck, J. (eds), *Perspectives on Binding and Atomism*, Foris, Dordrecht, 295–322.
- Rizzi, L., 1986. *Null objects in Italian and the theory of pro*. In: *Linguistic Inquiry*, Volume 17, 501–557.
- Sankoff, D., Kruskal, J. (eds), 1999. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Press, Stanford.
- Schutter, G. de, 1994. *Dutch*. In: König, E., Auwera, J. van der (eds), *The Germanic languages*. Routledge Language Family Descriptions, London, Routledge, New York.
- Schutter, G. de, Berg, B. van den, Goeman, T., Jong, T. de (eds), 2005. *Morphological Atlas of the Dutch Dialects*. Volume 1, Amsterdam University Press, Amsterdam.
- Séguy, J., 1971. *La relation entre la distance spatiale et la distance lexicale*, *Revue de linguistique Romane*, Volume 35, 335–357.
- Séguy, J., 1973. *La dialectométrie dans l'Atlas linguistique de la Gascogne*. *Revue de linguistique Romane*, Volume 37, 1–24.
- Spruit, M., 2005. *Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited*. In: Doetjes, J., Weijer, J. van de (eds), *Linguistics in the Netherlands, 2005*, John Benjamins, Amsterdam, 179–190.  
<http://marco.info/pro/pub/mrs2005lin.pdf>
- Spruit, M., 2006. *Measuring syntactic variation in Dutch dialects*. In: Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing*, special issue on Progress in Dialectome-

- try: Toward Explanation, Volume 21(4), Oxford University Press, Oxford, 493–506.  
<http://marco.info/pro/pub/mrs2006llc.pdf>.
- Spruit, M., 2006b. *Tellen met Taal. Het meten van variatie in zinsbouw in Nederlandse dialecten*. In: Gerritsen, D., Verburg, A. (eds), *Respons: Mededelingen van het Meertens Instituut*, Volume 8, Meertens Instituut, Amsterdam, 12–16.  
<http://marco.info/pro/pub/mrs2006respons.pdf>.
- Spruit, M., 2007. *Discovery of association rules between syntactic variables. Data mining the Syntactic Atlas of the Dutch dialects*. In: Dirix, P., Schuurman, I., Vandeghinste, V., Eynde, F. van (eds), *Computational Linguistics in the Netherlands 2006. Selected papers from the seventeenth CLIN meeting*, 83–98. <http://marco.info/pro/pub/mrs2007clin.pdf>.
- Spruit, M., Heeringa, W., Nerbonne, J., t.a. 2008. *Associations among linguistic levels*. *Lingua*, Special issue on Syntactic databases. <http://marco.info/pro/pub/shn2007dh.pdf>.
- Thráinsson, H., Angantýsson, Á., Svavarsdóttir, Á., Eythórrson, T., Jónsson, J., 2007. *The Icelandic (Pilot) Project in ScanDiaSyn*. In: Bentzen, K., Vangnes, Ø. (eds), *Scandinavian Dialect Syntax 2005*, Special issue of Nordlyd – Tromsø University Working Papers in Language & Linguistics. University Library of Tromsø, Tromsø, 87–124.
- Torgerson, W., 1952. *Multidimensional scaling: I. Theory and method*. *Psychometrika*, Volume 17, 401–419.
- Weijnen, A., 1946. *De grenzen tussen de oost-noord-Brabantse dialecten onderling*. In: Weijnen, A., Renders, J., Ginneken, J. van, *Oost-Noord-Brabantse Dialectproblemen. Bijdragen en Mededeelingen der Dialectcommissie van de Koninklijke Academie van Wetenschappen te Amsterdam*, Volume 8, 1–15.
- Weijnen, A., 1958. *Nederlandse dialectkunde*. Van Gorcum & Comp. N.V. - G.A. Hak & Dr. J. Prakke, Assen.
- Weijnen, A., 1966. *Nederlandse dialectkunde*. *Studia Theodisca*. Van Gorcum, Assen, Second edition. <http://www.dbnl.org/tekst/weij005nede01/>.
- Willems, P., 1886. *De enquête werd gehouden in 1886; de antwoorden zijn het eigendom van de Koninklijke Vlaamse Academie voor Taal- en Letterkunde, en worden daar bewaard* [The inquiry was done in 1886; the responses are the property of the Royal Flemish Academy of Languages and Literature in Gent where they are preserved]. Microcopies are at the Institute of Dialectology and Phonetics in Leuven, the Catholic university nijmegen, and the Meertens Institute Amsterdam.
- Winkel, J. te, 1901. *Geschiedenis der Nederlandsche taal*. Blom & Olivierse, Culemborg. Naar de tweede Hoogduitsche uitgave met toestemming van den schrijver vertaald door Dr. F. C. Wieder. Met eene Kaart.
- Winkler, J., 1874. *Algemeen Nederduitsch en Friesch dialecticon*. Two volumes, Martinus Nijhoff, 's-Gravenhage. <http://www.dbnl.nl/tekst/wink007alge00/>.

## Relevant software

This research has thankfully employed the excellent cartographic capabilities of the *RuG/L04* software package developed by Peter Kleiweg at the University of Groningen. The ANSI C programmes (with GPL license) generate Postscript dialect maps. All geographical maps in this dissertation have been created by this software. It is available at <http://www.let.rug.nl/~kleiweg/L04/>.

The SAND internally refers to dialect locations in Rijksdriehoeksmeting (RD) coordinates. This work incorporates the *rd2wgs* program developed by Ejo Schrama at the University of Delft. This ANSI C programme (with public domain license) converts Dutch RD coordinates to international WGS84 coordinates which the *RuG/L04* map generation software prefers. The conversion is accurate to about 50 centimetres.

The perfect dialect continuum map in Figure 6-1 was calculated using the Geodesy Foundation Classes (*GFC*) developed by Sam Blackburn. This is an ANSI C++ library (with freeware license) for distance calculations between earth locations. It is available at <http://www.samblackburn.com/gfc/>.

Most statistical analyses were performed with SPSS, although the *syndi* toolchain (see below) also generates an R script. The intersection of RND and SAND dialects underlying the associations among the three linguistic levels in Chapter 4, were determined with a set of nonpublic Pascal programmes and shell scripts developed by Wilbert Heeringa at the University of Groningen.

The SYNtactic DIAlectometry (*syndi*) software package is a toolchain of commandline programmes (with BSD license) which I have developed myself to perform this research. It depends on the wxWidgets cross-platform C++ library (with BSD-style license) which was originally developed by Julian Smart at the University of Edinburgh. The library provides a uniform programming interface for numerous operating systems and is available at <http://www.wxwidgets.org/>. The *syndi* toolchain has been extensively tested with current versions of the Windows and Mac operating systems.

The *syndi* toolkit consists of the following main components. First, the ‘Data Import Programme’ (*dimp*) imports the SAND data from a comma separated value (CSV) file, which can be exported directly from the SAND database. The *dimp* programme produces a self-describing data file in Extensible Markup Language (XML) format, a fragment of which is shown in Data fragment 1. The ‘SYNtactic measurement Console’ (*sync*) programme measures the syntactic distances between dialect pairs based on the syntactic variables in the XML data files. Feature variables can be defined in separate XML Attribute Language (XAL) files.

Data fragment 2 shows a small sample of a SAND1 data annotation file. The output of *sync* is further processed by the cartographic *RuG/L04* programmes. The dialect maps can be perfected with the ‘FINE-tune postscript map Console’ (*finc*) programme. The ‘Rule INduction Console’ (*rinc*) programme calculates the proportional overlap between geographical distributions of syntactic variables based on rule quality factors such as accuracy, coverage and completeness to measure the interestingness of the variable associations. The programme generates CSV files with association rules which can be imported for further analysis by standard data analysis programmes such as SPSS, R and Excel. Finally, the ‘BAtch Script Console’ (*bascc*) programme automates series of commandline invocations by executing platform-independent scripts in XML format. The *syndi* tools are available at <http://dialectometry.net/syntax/>.

*Data fragment 1: SAND1 data sample in Extensible Markup Language (XML) format.*

```
<?xml version="1.0" encoding="utf-8"?>
<data version="0.5" xml:lang="nl">
  <context name="ref1_01" map="68a" description="Zwak reflexief pronomen">
    <variable name="hem">
      <location name="A001p"/>
      <location name="B001a"/>
    </variable>
    <variable name="z'n eigen"/>
    <variable name="zich"/>
  </context>
</data>
```

*Data fragment 2: SAND1 annotation sample in XML Attribute Language (XAL) format.*

```
<?xml version="1.0" encoding="utf-8"?>
<data version="0.5" xml:lang="nl" annotated="yes">
  <context name="ref1_01" map="68a" reference="r4311" omschrijving="Zwak reflexief
pronomen als object van inherent reflexief werkwoord" description="Weak reflex-
ive pronoun as object of inherent reflexive verb" example="Jan herinnert __ dat
verhaal wel." gloss="John remembers __ that story [affirm]" translation="John
certainly remembers that story.">
    <variable name="zich" occurrences="121" features="reflexive"/>
    <variable name="hem" occurrences="112" features="personal"/>
    <variable name="z'n eigen" occurrences="43" features="possessive, ownness"/>
  </context>
</data>
```



# List of figures

Figure 1-1: “Toon wast ____”.....	11
Figure 1-2: Jellinghaus’ (1892) map based on an interpretation of various word and parable translations.....	16
Figure 1-3: Te Winkel’s (1901) map based on an interpretation of two linguistic questionnaires.....	16
Figure 1-4: Weijnen’s (1958) map based on 18 isophones and isomorphemes.....	16
Figure 1-5: Daan and Blok’s (1969) map based on subjective judgements.....	16
Figure 1-6: The 267 dialect locations in the Dutch language area under investigation.....	25
Figure 1-7: Distribution of the 267 Dutch dialects in the syntactic atlas.....	25
Figure 1-8: The provinces in the Dutch language area under investigation.....	25
Figure 1-9: SAND1 map 14b shows seven syntactic variables in the context of a complementiser of a comparative if-clause.....	27
Figure 2-1: The Daan and Blok dialect map (reprinted from Daan and Blok, 1969).....	34
Figure 2-2: MDS map visualising syntactic distances with respect to complementisers.....	40
Figure 2-3: MDS map visualising syntactic distances with respect to subject pronouns.....	40
Figure 2-4: MDS map visualising syntactic distances with respect to reflexive and reciprocal pronouns.....	41
Figure 2-5: MDS map visualising syntactic distances with respect to fronting.....	41
Figure 2-6: The SAND1 MDS dialect map based on a syntactic Hamming distance measure.....	42
Figure 2-7: The Daan and Blok dialect map based on subjective judgements (see also Figure 2-1).....	42
Figure 3-1: The Daan and Blok map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok, 1969).....	52
Figure 3-2: The SAND1 map of the Dutch dialects based on a syntactic measure after application of the Classical MDS procedure.....	52
Figure 3-3: Map of the Dutch dialects based on pronunciation differences after application of Kruskal’s Non-metric MDS procedure (reprinted from Heeringa, 2004).....	52
Figure 3-4: The selection of 21 dialect locations used in the regression analyses.....	52
Figure 3-5: Geographical distances versus syntactic distances using the subset of 21 dialect locations shown in Figure 3-4.....	55
Figure 3-6: Geographical distances versus syntactic distances using all 267 dialect locations.....	55
Figure 3-7: Geographical distances versus syntactic distances with respect to reflexives using atomic variables.....	60
Figure 3-8: Geographical distances versus syntactic distances with respect to reflexives using feature variables.....	60
Figure 4-1: Distribution of the 360 Dutch dialects in the RND atlas.....	76
Figure 4-2: Distribution of the 267 Dutch dialects in the SAND atlas.....	76
Figure 4-3: Distribution of the 70 common Dutch dialects in the RND and SAND atlases with the relevant province names.....	76
Figure 4-4: Expert consensus map of the Dutch dialects (translated from De Schutter, 1994).....	76
Figure 4-5: Perceptual map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok, 1969).....	77
Figure 4-6: Pronunciational MDS map of the Dutch dialects based on Levenshtein distances ( $r = 0.94$ ).....	77
Figure 4-7: Lexical MDS map of the Dutch dialects based on GIW distances ( $r = 0.74$ ).....	77

Figure 4-8: Syntactic MDS map of the Dutch dialects based on GIW distances ( $r = 0.89$ ). .....	77
Figure 4-9: Cronbach's alpha ( $\alpha$ ) is a function of the number of linguistic variables ( $n_{var}$ ) and the average inter-correlation value among the variables ( $\bar{r}$ ). .....	78
Figure 4-10: The average inter-correlation value ( $\bar{r}$ ) is based on all Pearson's correlation coefficients between each pair of variables ( $r(var_i, var_j)$ ). .....	78
Figure 4-11: This scatterplot shows the relation between pronunciational Levenshtein distances on the Y-axis and geographical distances on the X-axis. ....	82
Figure 4-12: This scatterplot shows the relation between lexical GIW distances on the Y-axis and geographical distances on the X-axis. ....	82
Figure 4-13: This scatterplot shows the relation between syntactic GIW distances on the Y-axis and geographical distances on the X-axis. ....	83
Figure 4-14: Influence of geography underlying the associations between the linguistic levels as a percentage. ....	85
Figure 5-1: Distribution of the 267 Dutch dialects in the Syntactic atlas. ....	93
Figure 5-2: The provinces in the Dutch language area under investigation. ....	93
Figure 5-3: This SAND1 sample marks the occurrences in seven dialects (1-7) of the four syntactic variables (A-D) in Table 5-1 to Table 5-4. ....	95
Figure 5-4: Symbolic representation of the SAND1 sample shown in Figure 5-3. ....	95
Figure 5-5: Calculation of the number of combinations with $k=3$ elements from the sample data set with $n=4$ variables. ....	96
Figure 6-1: A perfect MDS dialect continuum map resulting from a 100% correlation between syntactic and geographical distances. ....	114
Figure 6-2: An example of an MDS dialect mosaic map resulting from a low correlation between syntactic and geographical distances. ....	114
Figure 6-3: Fragment of a feature variable hierarchy with respect to fronting phenomena. ....	124
Figure 6-4: SAND1 MDS map visualising 485 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.955$ ). ....	128
Figure 6-5: Preliminary SAND2 MDS map visualising 697 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.932$ ). ....	128
Figure 6-6: Preliminary MDS map visualising 149 syntactic variables related to verbal clusters based on a Hamming distance measure ( $r = 0.894$ ). ....	129
Figure 6-7: Preliminary MDS map visualising 231 variables related to morphosyntactic variation based on a Hamming distance measure ( $r = 0.879$ ). ....	129
Figure 6-8: Preliminary MDS map visualising 186 syntactic variables related to negative concord and quantification phenomena based on a Hamming distance measure ( $r = 0.919$ ). ....	130
Figure 6-9: SAND1 Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.553$ ). ....	131
Figure 6-10: SAND2 Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.552$ ). ....	131
Figure 6-11: SAND Hamming distances on the Y-axis versus geographical distances on the X-axis ( $r = 0.592$ ). ....	132
Figure 6-12: SAND MDS map visualising 1182 syntactic variables in the aggregate based on a Hamming distance measure ( $r = 0.954$ ). ....	135

## List of tables

Table 1-1: Examples of syntactic variables in context for each syntactic domain/chapter in SAND1. Please refer to Table 5-1 to Table 5-4 for more detailed variable examples.....	26
Table 2-1: Example of a syntactic feature and its recorded variants. Map 68a in SAND1 shows the geographical distribution of the syntactic feature weak reflexive pronoun as object of inherent reflexive verb. Five feature variants have been recorded for this phenomenon throughout the Dutch language area: zich, hem, zijn eigen, zichzelf, hemzelf. ....	36
Table 2-2: Hamming distance algorithm applied to measure syntactic variation in dialects. ....	36
Table 2-3: Fragment of the SAND1 Hamming distance matrix. Each dialect pair distance is an integer between 0 and 510 which represents the total number of different feature variant realisations. ....	37
Table 2-4: Correlation between the original sets of SAND1 feature variants and the corresponding representation after reducing each set to three dimensions via MDS. ....	39
Table 3-1: Map 68a in SAND1 shows the five syntactic variables in the context of weak reflexive pronoun as object of inherent reflexive verb. ....	48
Table 3-2: Map 82b in SAND1 shows the six syntactic variables in the context of short object relative. ....	48
Table 3-3: Fragment of the distance measurement between two dialects using five syntactic variables. ....	49
Table 3-4: Fragment of the SAND1 Hamming distance matrix. ....	50
Table 3-5: Mapping from atomic variables (first column) to feature variables (first row) with respect to reflexive pronouns. ....	56
Table 3-6: Fragment of the distance measurement between two dialects using five feature variables (first column). ....	57
Table 3-7: Mapping from atomic variables (first column) to feature variables (first row) with respect to reciprocal pronouns. ....	63
Table 3-8: Mapping from atomic variables (first column) to feature variables (first row) with respect to one-pronominalisation. ....	63
Table 4-1: Map 14b in SAND1 shows seven syntactic variables in the context of complementiser of comparative if-clause. ....	69
Table 4-2: Map 54a in SAND1 shows four syntactic variables in the context of subject doubling 2 singular. ....	69
Table 4-3: String alignment and Levenshtein distance calculation between two pronunciations of the Dutch word hart 'heart'. ....	71
Table 4-4: Weighted similarity calculation between two dialects based on word choices for the three concepts of vriend 'friend', schip 'ship' and duwen 'to push' using the gewichteter Identitätswert (GIW) measure. ....	72
Table 4-5: Reliability coefficients ( $\alpha$ ) of our measurement results at the pronunciational, lexical and syntactic levels. ....	79
Table 4-6: Associations between aggregate pronunciational, lexical and syntactic distances. ....	79
Table 4-7: Correlations between geographical distances and pronunciational, lexical and syntactic distances. ....	83
Table 4-8: Associations between aggregate pronunciational, lexical and syntactic distances controlling for the influence of geography as an underlying factor. ....	84
Table 4-9: The percentage of the correlation attributable to geography. ....	85
Table 5-1: Map 14b in SAND1 shows seven syntactic variables in the complementisers domain. ....	94

Table 5-2: Map 54a in SAND1 shows four syntactic variables in the subject doubling domain.....	94
Table 5-3: Map 68a in SAND1 shows five syntactic variables in the reflexives domain. ....	94
Table 5-4: Map 84a in SAND1 shows four syntactic variables in the fronting domain. ....	94
Table 5-5: Algorithm to non-recursively evaluate all association rules.....	97
Table 5-6: Evaluation factors to help determine the quality of association rule ‘A → C’. ....	98
Table 5-7: Piatetsky-Shapiro’s principles for rule interestingness (RI) measures. ....	99
Table 5-8: The eight most interesting association rules in the sample data set as shown in Figure 5-3 and Figure 5-4 sorted on descending interestingness, ascending complexity and descending accuracy. ....	100
Table 5-9: Example of a highly ranked association rule in SAND1 with one variable disjunct: “if either antecedent variable A1 or A2 occurs, then it is certain that the consequent variable also occurs”. ....	102
Table 5-10: The most interesting rule in SAND1 without variable disjuncts.....	103
Table 5-11: More potentially interesting consequents in association rules which have the complex pronoun ‘g- + lieden’ as their antecedent, in addition to the rule consequent in Table 5-10. ....	104
Table 5-12: The most interesting implicational chain of association rules between four syntactic variables: d54a:after_v → d55a:after_v → p46a:g-lieden → p38b:gij/gie. ....	105
Table 6-1: A classification of syntactic variable types. ....	121
Table 6-2: Definitions of a selection of nominal measures of syntactic distance. ....	121
Table 6-3: Example distance measurements using atomic variables based on Table 3-3. ....	123
Table 6-4: Example distance measurements using feature variables based on Table 3-6. ....	123
Table 6-5: The corresponding matrix for the feature variable hierarchy in Figure 6-3. ....	125
Table 6-6: Map 84a in SAND1 shows three syntactic variables in the fronting domain. ....	126
Table 6-7: Fractional distance matrix in the short object relative context in Table 6-6, based on the feature variable mapping in Table 6-5.....	126
Table 6-8: Examples of syntactic variables in context for each syntactic domain/chapter in SAND2.....	127
Table 6-9: Visualisation perspectives on syntactic variation in Dutch dialects. ....	134

## List of terms

Arrow method.....	8	Gewichteter Identitätswert .....	59
Association rule mining		Hamming distance .....	24, 37
Apriori algorithm.....	84	Jaccard distance .....	96, 110
Association rules.....	84	Levenshtein distance.....	58
Implicational chain .....	93	Linguistic distance.....	9
Interestingness measures .....	87	Linguistic ruler.....	4
Market Basket Analysis .....	84	Measurement examples .....	111
Proportional overlap .....	85	Nominal measure .....	11
Quality factors.....	86	Nominal types.....	109
Atomic variables.....	37	Numerical measure .....	11
Comparative syntax .....	12	Feature frequency method .....	10
Composite variables.....	95	Feature variable hierarchy.....	112
Correlations.....	27, 62	Feature variables .....	44
Catholic-Protestant boundary ....	118	Influence of geography.....	73
Dutch surname diversity.....	118	Isoglosses.....	5
Frisian city dialect islands .....	41	Isogloss bundles .....	7
Correspondence analysis.....	96	Isogloss maps.....	7
Cramér's V .....	96	Isogloss method .....	7
Cronbach's alpha.....	66	Isomorpheme.....	7
Versus MDS coefficients .....	67	Isophone.....	6
Data mining .....	83	Knowledge discovery in databases	
Delaunay triangulation .....	62	(KDD) .....	83
Dialect.....	5	Language variation	
Dialect classification methods		Limits .....	13
Arrow method.....	8	Locus.....	13
Computational method.....	9	Patterns.....	13
Intuitive method.....	7	Language variation in context	
Isogloss method.....	7	Pronunciation versus lexis .....	63
Perceptual method.....	8	Syntax versus geography .....	42
Dialect maps		Syntax versus lexis.....	63
Barbiers et al. - Syntax (2005) .....	17	Syntax versus perception.....	40
Daan and Blok (1969) .....	8, 22	Syntax versus pronunciation..	41, 65
De Schutter (1994) .....	63	Linguistic atlases	
Heeringa - Lexis (2004).....	65	AIS .....	10
Heeringa - Pronunciation (2004) .	40	ALD .....	10
Jellinghaus (1892).....	5	ALF .....	10
Te Winkel (1901) .....	5	ALG .....	9
Van Ginneken (1913).....	7	ASIS .....	14
Weijnen (1958).....	7	Cordial-SIN.....	14
Dialect syntax .....	12	LASID .....	11
Dialectology .....	5	RND .....	10, 56
Dialectometry .....	9, 23	SADS .....	14
Distance measures		SAND1.....	14, 56, 81
Fractional distance.....	114	SAND2.....	115

ScanDiaSyn.....	14	Syntactic macrovariation.....	12
Linguistic variables.....	34	Syntactic microvariation .....	12
Local incoherence .....	48	Syntactic variables.....	14, 34, 81
Mantel test.....	68	A classification.....	109
Morphosyntactic variation.....	2	Atomic variable .....	34
Multidimensional scaling.....	26, 38, 61	Atomic versus feature variables....	46
Classical.....	39	Composite variable .....	34
Kruskal's Non-metric.....	39	Feature variable .....	34
Multiple regression analysis .....	72	SAND1 examples.....	82
Pearson product-moment correlation		SAND2 examples.....	115
coefficient .....	66	Syntactic variation.....	2, 18
Phonological features .....	10	Typological constraints .....	74
Qualitative research .....	9	Universal Grammar hypothesis .....	13
Quantitative research.....	9	Universal syntactic parameters.....	80
Residual analysis .....	72	Visualisation perspectives.....	122
SAND methodology.....	35	Voronoi polygons.....	62
Second language acquisition .....	42		

## Nederlandse samenvatting

Zoals de titel reeds vermeldt, onderzoekt dit proefschrift syntactische variatie in Nederlandse dialecten vanuit kwantitatieve perspectieven. Wat betekent dit?

In het vakgebied *syntactische variatie* onderzoekt men taalkundige verschillen op zinsniveau. Dit vakgebied bestudeert onder meer de verschillende volgordes van de woorden in een zin. Ook de regelmatigheden in de opbouw van woorden worden geanalyseerd indien deze samenhangen met de omgeving van de zin. Belangrijk is dat de zinsvarianten dezelfde betekenis uitdrukken. De volgende vier zinnen illustreren het onderzoeksgebied van syntactische variatie:

- (a) Het lijkt wel of er iemand in de tuin staat. (*Standaard Nederlands*)
- (b) Het lijkt wel dat er iemand in de tuin staat.
- (c) Het lijkt wel of dat er iemand in de tuin staat.
- (d) Het lijkt wel of er staat iemand in de tuin.

Hoofdstuk 1 van deze dissertatie opent met een beschrijving van enkele voorbeelden van syntactische variatie, waaronder de hierboven in (a) tot (d) opgesomde verschillen die men in het Nederlandse taalgebied aantreft in het gebruik van het voegwoord als aankondiging van een bijzin. In sommige dialecten zegt men (b), terwijl men in andere dialecten weer de vormen (c) of (d) gebruikt. De standaard Nederlandse zin in voorbeeld (a) gebruikt het voegwoord *of* om de bijzin te introduceren. De spreektaalvariant in voorbeeldzin (b) kiest het voegwoord *dat*, terwijl voorbeeld (c) in dialecten voorkomt waarin de voegwoordpositie wordt ingenomen door een samenstelling van de voegwoorden *of dat*. De zinnen (a), (b) en (c) zijn voorbeelden van microvariatie in de verbindingsfunctie van het voegwoord tussen de hoofd- en bijzin. In voorbeeldzin (d) bevindt het werkwoord *staat* zich op een andere positie in de bijzin dan in de voorgaande voorbeeldzinnen. Dit is een voorbeeld van dialectvariatie in woordvolgorde. Bovenstaande voorbeelden worden taalvariatie op syntactisch niveau genoemd aangezien de zinnen ondanks de verschillende functiewoorden en woordvolgordes dezelfde betekenis behouden. De zinnen drukken dezelfde betekenis uit.

Bovendien geldt voor de voorbeelden (b), (c) en (d) dat ze niet toegestaan zijn volgens de grammatica van het standaard Nederlands. Desondanks is er een schat aan dergelijke zinsbouwvariatie vastgelegd in het eerste deel van de *Syntactische atlas van de Nederlandse dialecten* (SAND1), waarin de zinsbouwvariatie in 267 Nederlandse dialecten in Nederland, België en Frankrijk is beschreven. Deze dissertatie maakt dankbaar gebruik van deze unieke verzameling gegevens als eerste bron van taalvariatie die geschikt is voor het uitvoeren van dialectometrisch onderzoek op puur syntactisch niveau. De voorbeelden in de zinnen (a)

tot (d) zijn in de context van dit onderzoek vier syntactische variabelen in één syntactische context. Spruit (2006) definieert een syntactische variabele als een functionele vorm of woordvolgorde in een syntactische context waarin twee dialecten kunnen verschillen.

In het vakgebied *dialectometrie* onderzoekt men taalkundige verschillen tussen dialecten vanuit *kwantitatieve perspectieven*. Kwantitatief taalkundig onderzoek richt zich op grote hoeveelheden taalkundige gegevens om inzicht te krijgen in de achterliggende principes. Het is een mooie aanvulling op het traditionele kwalitatieve taalkundig onderzoek, waarin slechts één of enkele taalkundige verschijnselen tot in detail onderzocht worden. Waar kwalitatief onderzoek in de diepte gaat, concentreert kwantitatief onderzoek zich op de breedte—een ander perspectief dus. De kern van kwantitatief onderzoek zit in de overgang van het meten van *afzonderlijke* taalkundige verschillen naar *samengevoegde* verschillen tussen taalvariëteiten. Hiervoor is het nodig om numerieke waarden toe te kennen aan taalkundige verschijnselen. Na de vertaalslag van tekst-verschillen naar taal-verschillen wordt het mogelijk om te tellen met taal.

Dit proefschrift beschrijft verschillende kwantitatieve methoden waarmee voor het eerst op objectieve en verifieerbare wijze inzicht gegeven kan worden in meer globale karakteristieken van syntactische variatie met behulp van geaggregeerde distributiepatronen van syntactische verschijnselen, de samenhang tussen syntactische variatie en andere taalkundige niveaus, en afhankelijkheden tussen syntactische verschijnselen. Spruit (2006b) introduceert dit onderzoek in meer detail voor geïnteresseerde niet-taalkundigen. Het vervolg van deze Nederlandstalige samenvatting beschrijft in het kort de hoofdstukken van deze dissertatie.

Hoofdstuk 1 motiveert het belang van kwantitatief taalkundig onderzoek op syntactisch niveau met behulp van enkele conflicterende syntactische variatiepatronen. Vervolgens worden de vakgebieden *dialectcartografie*, *dialectometrie* en *syntactische microvariatie* vanuit historisch perspectief geïntroduceerd om de wetenschappelijke context en relevantie te schetsen van dit eerste onderzoek naar dialectometrische toepassingen op puur syntactische data. Tevens wordt het huidige werk vanuit vier verschillende onderzoeksdimensies belicht om aan te geven waar dit onderzoek *niet* over gaat. Een inleidend overzicht van de hoofdstukken volgt na formulering en verduidelijking van de volgende onderzoeksvragen:

- I. Hoe kan syntactische variatie op adequate wijze worden gemeten? (*Model*)
- II. Wat zijn de syntactische afstanden tussen de Nederlandse dialecten? (*Toepassing*)



- III. In welke mate zijn de taalkundige niveaus van syntax, lexicon en uitspraak met elkaar geassocieerd? (*Context*)
- IV. Wat zijn relevante afhankelijkheden tussen syntactische variabelen? (*Associaties*)

De onderzoeksvragen I en II analyseren gezamenlijk de relatie tussen syntactische en geografische afstand door de verschillen tussen de Nederlandse dialecten op zinsbouwniveau te kwantificeren en deze geografisch in kaart te brengen. Deze twee vragen worden beantwoord in Hoofdstuk 2 en Hoofdstuk 3. Onderzoeksvraag III bestudeert de mate waarin geografische distributies van syntactische afstanden correleren met distributies van uitspraak- en woordkeusafstanden om de meetresultaten in een bredere taalkundige context te kunnen plaatsen. Deze vraag is het onderwerp van Hoofdstuk 4. Onderzoeksvraag IV behandelt computationele methoden om op objectieve en verifieerbare wijze relevante associaties en afhankelijkheden tussen syntactische variabelen te ontdekken op basis van geografische distributiepatronen. Deze vraag wordt onderzocht in Hoofdstuk 5.

Hoofdstuk 2 onderzoekt de relatie tussen syntactische variatie en geografische afstand door het Nederlandse dialectclassificatieprobleem vanuit een dialectometrisch perspectief te benaderen. Het vergelijkt de resultaten van het ontwikkelde syntactische meetinstrument—geprojecteerd op een geografische kaart—met de traditionele *Daan en Blok* kaart uit 1969 die de Nederlandse dialecten classificeert op basis van *subjectieve oordelen* van lokale dialectsprekers. Het hoofdstuk presenteert een kwantitatieve maat van syntactische afstand om op objectieve en verifieerbare wijze te kunnen differentiëren tussen dialectgrenzen en dialectcontinua. Het beschrijft de *pijljesmethode* en de methodologische uitdagingen die zich voordoen bij de perceptuele classificatie van de Nederlandse dialecten op basis van subjectieve oordelen. Deze problemen leiden tot de introductie van het onderzoeksgebied dialectometrie en SAND1 als een puur syntactische gegevensbank. De database bevat 510 syntactische variabelen en is daarmee bruikbaar voor kwantitatieve analysedoeleinden. De dialectometrische methode die dit hoofdstuk beschrijft, *aggregeert syntactische verschillen* tussen dialectvariëteiten met behulp van een *Hamming* afstands algoritme. De zeer repetitieve metingsprocedure resulteert uiteindelijk in de SAND1 *Hamming afstandstabel*. De dialectrelaties in de afstandstabel worden geanalyseerd door de Classical *Multidimensional scaling* (MDS) procedure toe te passen met als doel om voor ieder dialect de meest onderscheidende syntactische variabelen zo optimaal mogelijk te representeren in relatie tot alle andere dialecten. De variatie in het Nederlandse taalgebied wordt geografisch gevisualiseerd met behulp van *dialectkleurenkaarten*, waarin de MDS kaartkleuren corresponderen met de eerste drie dimensies van de oplossing van de MDS procedure. De bespreking van de resultaten onderzoekt eerst de toepassing van de MDS procedure op elk van de

zeven SAND1 domeinen afzonderlijk. Vervolgens wordt de geaggregeerde SAND1 MDS dialectkaart berekend op basis van de syntactische Hamming afstandsmaat, wat in een homogeen kleurcontinuüm met duidelijk waarneembare dialectgebieden resulteert. De SAND1 MDS kaart maakt duidelijk dat syntactische variatie *geografisch coherent* gestructureerd is wanneer het in het aggregaat bestudeerd wordt. Bovendien komt de objectieve classificatie van Nederlandse dialectvariëteiten op basis van een syntactische afstandsmaat in hoge mate overeen met de classificatie op basis van subjectieve oordelen op de Daan en Blok dialectkaart. Dit bevestigt en valideert de syntactische afstandsmetingmethode.

Hoofdstuk 3 verdiept het in Hoofdstuk 2 beschreven onderzoek vanuit meerdere perspectieven. Ten eerste wordt de SAND1 MDS dialectkaart op basis van een syntactische afstandsmaat nu ook vergeleken met de *Heeringa* (2004) dialectkaart op basis van *uitspraakverschillen*. Een visuele vergelijking tussen de syntaxkaart en de uitspraakkaart maakt duidelijk dat de kaarten tot op zekere hoogte corresponderen, alhoewel de syntactische kaart een minder geleidelijk kleurcontinuüm vertoont. Ten tweede worden de *geografische afstanden* gecorreleerd met syntactische Hamming afstanden met behulp van *regressieanalyses* om te onderzoeken hoeveel van de vastgelegde syntactische variatie toe te schrijven is aan geografische afstand. Enkele regressieanalyses, zowel op basis van een optimale doorsnede van 21 dialectvariëteiten als op basis van alle 267 dialecten, tonen aan dat respectievelijk 56 en 30 procent van de syntactische afstanden lineair verklaard kan worden met behulp van geografische afstanden. Ten derde presenteert dit hoofdstuk meetresultaten op basis van binaire vergelijkingen tussen *feature variables* die geformuleerd zijn door handmatig syntactische variabelen te annoteren met *taalkundige kenmerk-informatie*. De meetresultaten op basis van de geformuleerde feature variabelen worden vergeleken met de resultaten op basis van de waargenomen *atomic variables* voor het syntactische domein van de reflexieven in SAND1. De geografische distributies lijken vrijwel identiek na toepassing van de MDS procedure. De visuele overeenkomst wordt bevestigd door de resultaten van een regressieanalyse. Toepassing van de *local incoherence* validatiemethode suggereert dat de onderlinge afstanden tussen atomische variabelen enigszins beter de lokale conditionering van dialectverschillen lijken te reflecteren dan de onderlinge afstanden tussen kenmerkvariabelen.

Hoofdstuk 4 meet de mate van verwantschap tussen *geaggregeerde uitspraak-, woordkeus- en zinsbouwverschillen*. Dit onderzoek—het resultaat van een samenwerkingverband met Wilbert Heeringa en John Nerbonne—kwantificeert de lexicale en syntactische verschillen op nominaal niveau met behulp van de *gewichteter Identitätswert* (GIW) methode—een frequentie-gebaseerde gelijkheidsmaat—en meet uitspraakverschillen op numeriek niveau met behulp van de *Levenshtein* afstandsmaat. Het bestudeert de deelverzameling van 70 Nederlandse dialectvariëteiten die voorkomen in zowel de *Reeks Nederlandse Dialectatlassen*

(RND; Blancquaert en Peé, 1925-1982) als SAND1. De RND data worden gebruikt om zowel uitspraak- als woordkeusafstanden te meten, terwijl de SAND1 data gebruikt worden om syntactische afstanden te meten. Het hoofdstuk presenteert kleurenkaarten van de Nederlandse dialectgebieden op basis van uitspraak-, woordkeus- en zinsbouwverschillen in paarsgewijze vergelijkingen om een eerste visuele indruk te verkrijgen van de associaties tussen de taalkundige niveaus van uitspraak, lexicon en syntax. De kleurenkaarten visualiseren op geografische wijze de taalkundige variatie in het Nederlandse dialectgebied door toepassing van de MDS procedure. *Cronbach's alfa* consistentiecoëfficiënten worden berekend om de minimale betrouwbaarheid te bepalen van de afstandsmetingen op basis van de gebruikte databronnen. De correlatiecoëfficiënten tussen de afstandsmetingen voor de drie taalkundige niveaus worden berekend als de graden van verwantschap tussen de drie taalkundige niveaus. Aangezien regressieanalyses duidelijk aantonen dat geografie elk van de drie taalkundige niveaus afzonderlijk beïnvloedt, worden de correlaties tussen alle taalkundige niveaus opnieuw berekend in *meervoudige regressieanalyses* om geografie als een onderliggende factor van invloed uit te filteren. Deze analyses resulteren in substantiële maar bescheiden graden van verwantschap tussen de drie taalkundige niveaus.

Hoofdstuk 5 onderzoekt een *data mining* techniek om relevante associaties te ontdekken tussen 485 syntactische variabelen in SAND1 met behulp van een regelinductiesysteem dat gebaseerd is op het principe van *proportional overlap*. De methode van *association rule mining* berekent de proportionele overlap tussen geografische distributies van syntactische variabelen en maakt gebruik van kwaliteitsfactoren zoals *accuracy*, *coverage*, *completeness* en *complexity* om de graad van *interestingness* te meten van associaties tussen variabelen. Dit onderzoek beperkt zich tot de *Piatetsky-Shapiro (1991)* maat van interessantheid vanwege zijn historische positie en eenvoud van formulering. Ten eerste presenteert het hoofdstuk het niet-recursieve algoritme voor *association rule mining* in pseudocode en verduidelijkt het de procedure met behulp van een minimale deelverzameling van de daadwerkelijke SAND1 data. De voorbeeldprocedure onthult de *asymmetrische aard* van associaties tussen syntactische variabelen, die geïnterpreteerd zouden kunnen worden als afhankelijkheden tussen variabelen met potentiële hiërarchische implicaties. Daarna wordt de *association rule mining* methode toegepast op 485 syntactische variabelen in 267 Nederlandse dialecten in SAND1. Tenslotte beschrijft de verkennende bespreking van de resultaten de *hoogst gerangschikte* associatieregels met en zonder *variabelendisjuncties* en bestudeert de resultaatbespreking tevens een *implicatieve keten* van associaties tussen variabelen. De resultaten onthullen de hoge gradaties van proportionele overlap tussen de geografische distributies van de syntactische variabelen in SAND1, die effectief het belang van de geografische voorkomens in de gegevensverzameling reduceren. Deze observatie zou syntactische analyses kunnen faciliteren het

waarnemingsniveau van geografische distributies te ontstijgen naar meer abstracte associatiepatronen tussen variabelen.

## **Curriculum vitae**

Marco René Spruit werd geboren op 6 oktober 1969 in Ermelo.

In 1988 behaalde hij het VWO diploma aan het Christelijk College Groevenbeek in Ermelo, waarna hij ging studeren aan de Universiteit van Amsterdam. In 1989 ontving hij de propedeusebul Nederlandse Taal- en Letterkunde. In 1990 behaalde hij de propedeuse Muziekwetenschap en specialiseerde zich tot 1994 in de Systematische Muziekwetenschap. In 1995 voltooide hij de specialisatie Taal en Informatica in de bovenbouwstudie Alfa-informatica in het kader van het EU project “Neural Networks and Information Retrieval in a Libraries Context”.

Van 1993 tot 1997 werkte hij als software ontwikkelaar bij MSC Information Retrieval Technologies BV en het Ministerie van Defensie. Van 1997 tot 2003 leidde hij als zelfstandig software ontwikkelaar voor het midden- en kleinbedrijf de ondernemingen Wizzer en Insertable Objects. Van 1995 tot 2001 was hij tevens werkzaam als freelance redactiemedewerker van het Personal Computer Magazine bij VNU Business Publications BV waar hij software ontwikkelingsomgevingen, programmeertalen, databases en multimedia producten recenseerde.

Van 2003 tot 2007 heeft hij dit promotieonderzoek verricht als Onderzoeker in Opleiding aan het Meertens Instituut te Amsterdam. In 2005 mocht hij voor dit onderzoek de “Association for Linguistic and Literary Computing bursary” prijs in ontvangst nemen aan de Universiteit van Moncton in Canada. In 2006 bezocht hij als “visiting scholar” gedurende enkele maanden de Universiteit van Triëst in Italië. Hij is nu werkzaam aan de Universiteit van Utrecht als universitair docent/onderzoeker Informatiekunde bij de onderzoeksgroep Organisatie en Informatie.