Quantitative perspectives on syntactic variation in Dutch dialects
Spruit, M.R.

Citation for published version (APA):

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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.
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 pr onunciational, 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 pr onunciational, lexical and syntactic distances influenced by geography as an underlying factor?

Table 4-9 shows that the degree of association between pr onunciational and lexical distances turns out to be based on geography as an underlying factor for no less than 39%. The association between syntactic and pr onunciational 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 4-5 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.
v. 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.

vi. 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.

vii. 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 Identitätszert (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 combinatorial 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. 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.

---

1 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.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
</table>
| 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): | \[
\frac{\text{diff}(A,B) + \sum_{i} \text{freq}(i)/m}{\text{diff}(A,B) + \text{ident}(A,B)}
\] | 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. \(\sum_{i} \text{freq}(i)\)) 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 noted as follows: \( 1 - \frac{|A \cap B|}{|A \cup B|} \). 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.

<table>
<thead>
<tr>
<th></th>
<th>Lunteren (A)</th>
<th>Veldhoven (B)</th>
<th>Ident. ((i))</th>
<th>Diff. ((d))</th>
<th>Freq. ((f))</th>
</tr>
</thead>
<tbody>
<tr>
<td>[sand1,68a]: zich</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: hem</td>
<td></td>
<td></td>
<td></td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: zijn eigen</td>
<td>√</td>
<td></td>
<td>√</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: zichzelf</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: hemzelf</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>Lunteren (A)</th>
<th>Veldhoven (B)</th>
<th>Ident. ((i))</th>
<th>Diff. ((d))</th>
<th>Freq. ((f))</th>
</tr>
</thead>
<tbody>
<tr>
<td>{zich, zijn eigen}</td>
<td></td>
<td>{zich}</td>
<td></td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: reflexive</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: possessive</td>
<td>√</td>
<td></td>
<td>√</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: ownness</td>
<td>√</td>
<td></td>
<td></td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>[sand1,68a]: focus</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

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. 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.

---

2 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.

<table>
<thead>
<tr>
<th></th>
<th>SpecCP</th>
<th>Wh-form</th>
<th>Reduced</th>
<th>Neuter₁</th>
<th>Neuter₂</th>
<th>Casus</th>
<th>Nominative</th>
</tr>
</thead>
<tbody>
<tr>
<td>wat</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wie</td>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dat₁</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>die</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>der</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>den/dem</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>'t/Ø</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dat₂/as</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 at, 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(+) » Wh-form(+) » Neuter₂(+)], whereas the path of the atomic variable wat is [SpecCP(+) » Wh-form(+) » Neuter₁(+)]. 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*, *Neuter2* and *Cains*. 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.

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.

<table>
<thead>
<tr>
<th>1:die 2:as/at/da(t)</th>
<th>1:die 2:-t</th>
<th>1:wie 2:-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:die 2:as/at/da(t)</td>
<td>(0/4 + 1/2) / 2 = 0.25</td>
<td>(1/2 + 1/2) / 2 = 0.50</td>
</tr>
<tr>
<td>1:die 2:-t</td>
<td>(0/4 + 1/2) / 2 = 0.25</td>
<td>(1/2 + 0/2) / 2 = 0.25</td>
</tr>
<tr>
<td>1:wie 2:-t</td>
<td>(1/2 + 1/2) / 2 = 0.50</td>
<td>(1/2 + 0/2) / 2 = 0.25</td>
</tr>
</tbody>
</table>

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

\[ \text{The Fractional distance is divided by two to obtain the normalised distance between dialects A and B because the comparison between the syntactic variables consists of two linguistic elements.} \]
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 *wilt infinitive*.
  - ‘nobody has that ever want infinitive’

- **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’
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. 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.

---

4 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.

5 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.
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 pronunciational level (the pronunciational 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.
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
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.\(^6\) 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.

\(^6\) Applying the formula in Figure 4.14, which calculates the influence of geography underlying the associations between two linguistic levels, results in: \( (1 - (0.401 / 0.459)) \times 100 = 12.6 \) 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. 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.

---

7 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 subdomains (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 visualizing 1182 syntactic variables in the aggregate based on a Hamming distance measure ($r = 0.954$).