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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

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.
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).1

(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.2,3 However, it is a well-known fact that

1 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.
2 The Dutch reflexive pronoun zich cannot be literally translated to English. It is normally annotated with REFL or Sg in word-by-word translations.
3 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). 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.”

---

4 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 pronunciational 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 unchartered 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 /i/ 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).\(^5\) 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.\(^6\) Isogloss one overlaps to a large extent with isogloss seven (the Uerdinger 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).\(^7\) 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-

\(^5\) 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).

\(^6\) Van Oostendorp (2006) notes that these distinctive tonal contours in Limburg dialects of Dutch are traditionally called stoottoon (‘bumping tone’) and sleeptoon (‘dragging tone’).

\(^7\) 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). 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

---

8 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).

Goebl (1982) marks the beginning of a series of major contributions to dialectometrical research. Goebl 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. Goebl’s methodologies resemble Séguy’s techniques considerably, although their measurement strategies differ with respect to the research focus. Whereas Séguy calculates dialect distances, Goebl 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. Goebl’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, Goebl’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; Goebl and Böhmer, 1985-2011). Goebl (2006) extensively describes the current state of Goebl’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 pronunciational 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 pronunciational 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. 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, Goebl, the Hoppenbrouwers brothers, and others. Unfortunately, the Levenshtein distance algorithm also has a fundamental shortcoming as a tool to accurately measure linguistic distances. Heeringa (2004:25) notes that:

---

9 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.

10 Although Goebl’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, Koeneman 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 moet zwemmen (*’I think that everyone can must swim’), and Barbiers and Bennis (2003) investigate why certain logically conceivable strong reflexives in Dutch such as *ben-eigen ‘him-own’ in *Jan herinnert ben-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).

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 northwestern 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). 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

---

11 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.

12 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 je 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 weak you strong healthily live, live- you weak you strong 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
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 phonological 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 21st 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 pronunciational 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 pronunciational, 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 pronunciational, 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.