Measuring Syntactic Variation in Dutch Dialects

Marco René Spruit

Meertens Instituut
Joan Muyskenweg 25
Postbus 94264
1090 GG Amsterdam
The Netherlands

marco.rene.spruit@meertens.knaw.nl
http://www.meertens.knaw.nl/medewerkers/marco.rene.spruit
+31 20 462 85 00

submitted to J. Nerbonne and W. Kretzschmar, Jr. (eds.)
Progress in Dialectometry, special issue of
Literary and Linguistic Computing,
selected proceedings of a workshop at
Methods in Dialectology XII, Moncton, Aug. 5, 2005.

March 16, 2006

Abstract

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

1. Introduction

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

The Daan and Blok (1969) map of the Dutch dialects shown in Figure 1 can be seen as an early attempt to represent dialectal differences in the aggregate. The classification of the Dutch dialects on this map is derived using subjective judgements of local speakers, local experts and Daan and Blok themselves. However, Spruit (2005) notes a number of practical and methodological problems which may have significantly influenced the outcome of this classification of Dutch dialect areas based on perceptual differences. Therefore, objective methods are required to assign numerical values to linguistic phenomena to aggregate individual dialect differences. These dialectometric methods were first described in Seguy (1971) and further investigated in Goebl (1984) and Heeringa and Nerbonne (2001), among others. However, these dialectometric studies were mainly limited to lexical and phonological data. Most notable in this context is the application of the Levenshtein method to aggregate differences in dialect pronunciation in Heeringa (2004).

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

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

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

This paper is structured as follows. The data with respect to syntactic variation in Dutch dialects are introduced in Section 2. The syntactic measurement procedure and the analysis technique are described in Sections 3 and 4 respectively. The resulting geographical color map of the Dutch dialect area based on a syntactic measure is presented in Section 5 and is related to distributions based on perception and pronunciation in Section 6. The latter section also includes an analysis of the correlation between syntactic variation and geographical distance. An alternative measure of syntactic distance based on feature variables is presented in Section 7. The measurement results based on feature variables are compared with the results based on atomic variables in Section 8. The paper concludes with a recapitulation of the most significant results and directions for future research in Sections 9 and 10.

2. Syntactic Atlas of the Dutch Dialects

Until recently dialectometric research was mainly limited to lexical and phonological data because no extensive collection of purely syntactic data was available. This situation has changed with the arrival of the first part of the Syntactic Atlas of the Dutch Dialects (SAND1, Barbiers et al. 2005). It contains 145 maps showing the geographic distribution of syntactic variables in 267 Dutch dialects. Geographic distributions of individual syntactic variables are shown in 134 maps.[1] The other 11 maps display correlations between syntactic variables. The second volume of the SAND will appear in 2007 and will contain data with respect to syntactic variation in verbal clusters and negation.

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

At each measuring point in the Netherlands the interview was not carried out by the field workers themselves but by local dialect speaking assistants, since most field workers did not speak the local dialect. The field worker would first instruct the assistant. Then, the assistant conducted the interview with the informant in the local dialect to avoid acommodation effects. The field worker's main role was to ensure adherance to the interview protocol. In Belgium no separate interview assistants were employed because the Belgian field workers were regional dialect speakers themselves. All in all, it may be safely assumed that the extensive SAND methodology provides a solid foundation for the results presented in this paper.

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

Table 1 Map 68a in SAND1 shows the five syntactic variables in the context of weak reflexive pronoun as object of inherently reflexive verb

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

Table 2 Map 82b in SAND1 shows the six syntactic variables in the context of short object relative

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

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

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

3. Hamming Distance Measure

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

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

	Lunteren	Veldhoven	distance
[sand1,68a]: zich	+	+	0
[sand1,68a]: hem	-	-	0
[sand1,68a]: zijn eigen	+	-	1
[sand1,68a]: zichzelf	-	-	0
[sand1,68a]: hemzelf	-	-	0
			1

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

This measurement based on binary comparisons of syntactic variables is carried out for all 507 variables, and the procedure is repeated for all (267 * 266) / 2 = 35511 unique dialect pairs.[3] The final result is a Hamming distance matrix a part of which is shown in Table 4. In this matrix each distance value represents the total number of different syntactic variable realisations between one pair of dialects. For example, the matrix shows that 47 different variable realisations were recorded between the dialects of Lunteren and Veldhoven after comparing all 507 syntactic variables.

Table 4 Fragment of the SAND1 Hamming distance matrix

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

4. Multidimensional Scaling Analysis

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

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

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

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

5. Map of the Dutch Dialects

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

A correlation coefficient of nearly 0.96 is achieved using the Classical MDS method. This value indicates how much of the syntactic variance is represented in the first three dimensions of the MDS solution, which, in this context, quantifies the amount of syntactic variance represented in the map colours. Correlation values between 0.9 and 1.0 are quite high, indicating that the MDS result faithfully represents the information in the original distance matrix. Thus, the claim can be made that the SAND1 MDS dialect map visualises the actual dialect relationships accurately.[6]

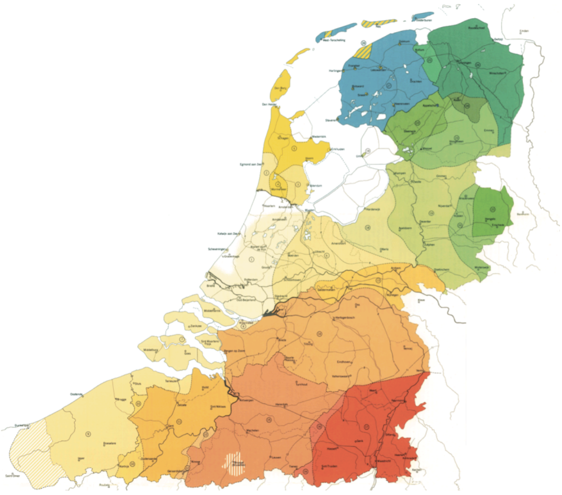

Fig. 1 The Daan and Blok map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok 1969)

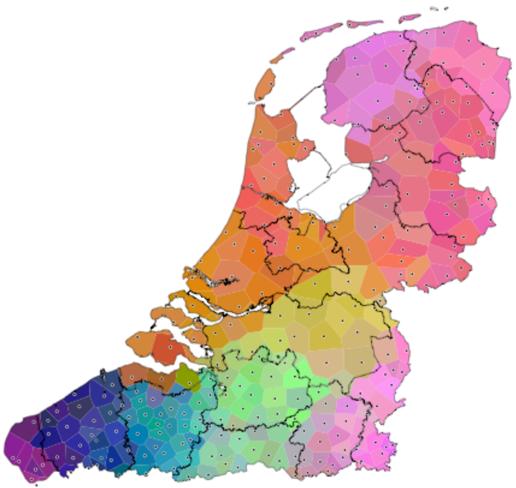

Fig. 2 The SAND1 map of the Dutch dialects based on a syntactic measure after application of the Classical MDS procedure

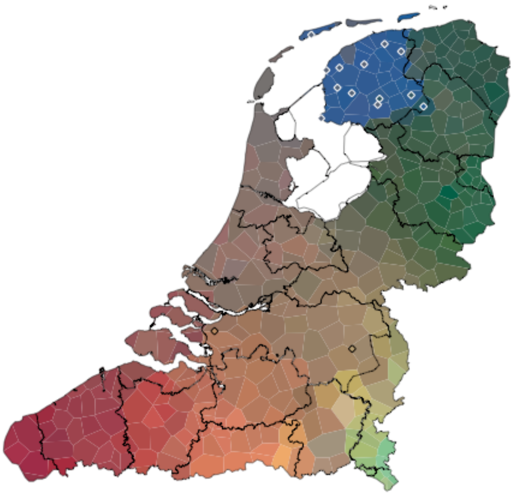

Fig. 3 Map of the Dutch dialects based on pronunciation differences after application of Kruskal's Non-metric MDS procedure (reprinted from Heeringa 2004)

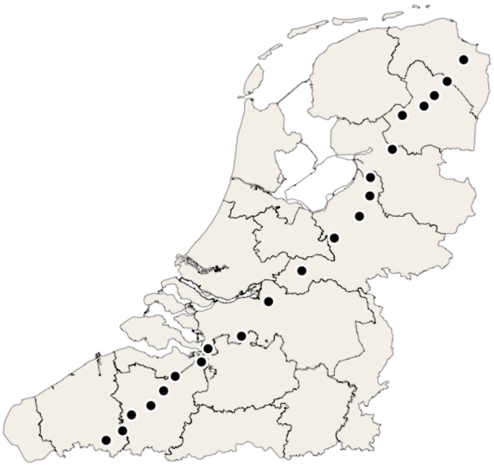

Fig. 4 The selection of 21 dialect locations used in the regression analyses [5]

6. Syntactic variation in context

Syntax versus Perception

The SAND1 MDS map in Figure 2 is shown next to the Daan and Blok dialect map in Figure 1. This view puts the geographic distribution of syntactic variation into a perceptual perspective. The objective SAND1 dialect area classification based on a syntactic measure looks quite similar to Daan and Blok's subjective dialect area classification based on subjective judgements. The similarities are even more remarkable when taken into account the fact that the colours used in the Daan and Blok dialect map were chosen more or less intuitively, although corresponding to a gradually increasing divergence from Standard Dutch (Goeman 2000).

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

Syntax versus Pronunciation

The SAND1 MDS map in Figure 2 is shown above the Heeringa MDS dialect map based on pronunciation differences in Figure 3. This view illustrates the geographic distribution of syntactic variation in comparison to pronunciation. The pronunciation dialect map shows a smooth dialect continuum except for the Frisian city dialect islands in the blue Frisian area. These varieties are symbolised with diamonds to indicate that they do not belong to the group in which they are found geographically. Apart from the general observation that the SAND1 MDS map shows a less smooth color continuum overall, the most interesting discrepancy between these two maps is arguably the complete absence of the Frisian city dialect islands in the SAND1 MDS map. Upon closer examination, however, only three out of thirteen Frisian dialect islands on the map in Figure 3 also occur as dialect locations in the SAND.[7] This mismatch of locations already explains most of the discrepancy between Figures 2 and 3, since city dialect islands are by definition of a local and isolated nature.

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

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

Syntax versus Geography

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

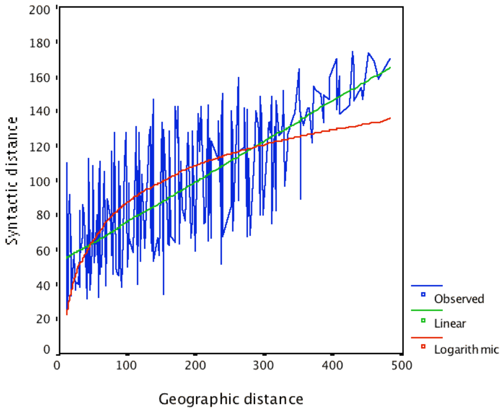

Fig. 5 Geographic distances versus syntactic distances using the subset of 21 dialect locations shown in Figure 4

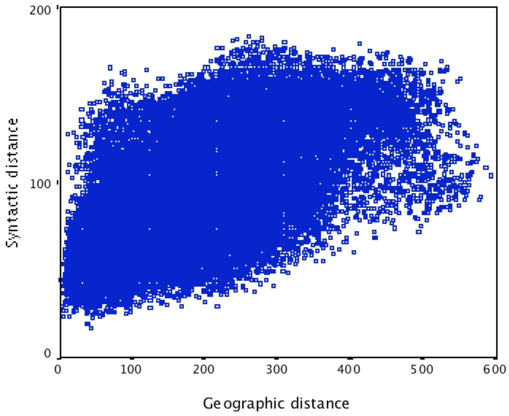

Fig. 6 Geographic distances versus syntactic distances using all 267 dialect locations

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

7. Feature Variables

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

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

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

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

	personal	reflexive	possessive	ownness	focus
	"hem"	"zich"	"zijn"	"eigen"	"zelf"
hem	+				
hemzelf	+				+
zich		+			
zichzelf		+			+
zijn			+		
zijn zelf			+		+
zijn eigen			+	+	
zijn eigen zelf			+	+	+

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

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

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

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

Abstracting away from occurring atomic variables to represented feature variables has several advantages when measuring dialect distances. For example, a measure based on atomic variables cannot differentiate between the variables *zich* and *zichzelf* on the one hand and *zich* and *zijn eigen* on the other hand. Both are assigned a distance value of one because in both cases the two variables are different. A measure based on feature variables also assigns a distance value of one between the variables *zich* and *zichzelf* because they share the *reflexive* feature variable but differ with respect to the *focus* feature variable as shown in Table 6. However, the distance between the variables *zich* and *zijn eigen* is assigned a distance value of three because the three underlying features for these variables do not match at all. The atomic variable *zich* reflects the *reflexive* feature variable and the atomic variable *zijn eigen* represents the *possessive* and *ownness* feature variables.

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

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

8. Atomic Variables versus Feature Variables

The measurement results using either atomic variables or feature variables have been compared with respect to the SAND1 data in the reflexives subdomain. The geographic distributions turn out to be nearly identical after application of the MDS procedure. The measure using atomic variables consisted of 75 comparisons between each pair of dialects, and application of the MDS procedure results in a three-dimensional solution which correlates highly with the original distance matrix (r = 0.93).[9] The measure using feature variables included 61 comparisons between each pair of dialects and results in a correlation of 0.94. These correlations indicate that both atomic variables as well as feature variables can be used to faithfully illustrate syntactic variation in three dimensions. Furthermore, both maps correspond to a reasonable extent to the descriptive Dutch area classification with respect to reflexives in Barbiers and Bennis (2004). This description distinguishes 5 main dialect areas in the geographic distribution of variation with respect to reflexives. Contours of these generalisations can also be found on the MDS maps.

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

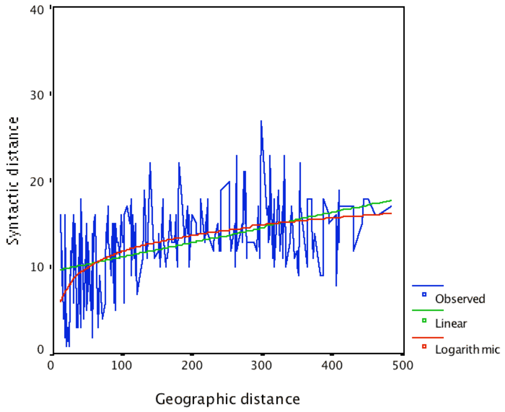

Fig. 7 Geographic distances versus syntactic distances with respect to reflexives using atomic variables

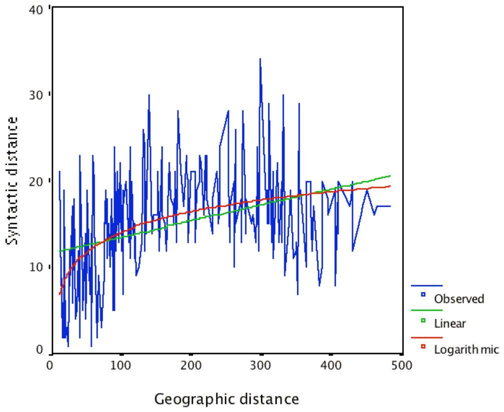

Fig. 8 Geographic distances versus syntactic distances with respect to reflexives using feature variables

Regression analyses were performed to correlate the syntactic measure using atomic variables with the measure using feature variables with respect to reflexives. A regression analysis using the same selection of 21 dialect locations as shown in Figure 4 results in a correlation coefficient of 0.93. A regression analysis using all 266 dialects leads to a correlation value of 0.92.[10] This means that there is a strong correlation between the syntactic measure using atomic variables and the measure using feature variables with respect to reflexives.

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

All in all, the results based on a measure using either atomic variables or feature variables are quite similar. An explanation may be found in the shape of the regression curves shown in Figures 7 and 8. Both regressions start from a relatively steep angle until the syntactic distance levels off to a fairly flat level in relation to the geographic distance, suggesting that measuring syntactic distances between distant dialect locations no longer reliably reflects linguistic dissimilarity. This assumption may be confirmed using the *local incoherence* validation method described in Nerbonne and Kleiweg (2005). Local incoherence is a numerical probe to compare distance matrices with respect to the degree to which they reflect local geography faithfully. Lower local incoherence scores indicate that a given distance matrix better reflects local conditioning of dialect differences. Application of this method to the distance matrix based on atomic variables results in a local incoherence value of 10.3. The matrix based on feature variables results in a local incoherence value of 10.7. This means that the measure using atomic variables brings about slightly better results than the measure using feature variables, which confirms the results of the regression analysis.

9. Conclusions

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

10. Future Research

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

Notes

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

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

2. Syntactic variables are referred to as syntactic features in Spruit (2005).

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

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

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

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

7. The three Frisian city dialect islands in Figure 3 which also occur in the SAND are Midsland, Heerenveen and Kollom.

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

9. The MDS map visualising syntactic distances with respect to reflexive and reciprocal pronouns is printed in Spruit (2005:186).

10. No data is available with respect to reflexives for the dialect of Morbecque.

References

Barbiers, S., H. Bennis, M. Devos, G. de Vogelaer, M. van der Ham, eds., (2005). Syntactic Atlas of the Dutch Dialects, volume 1. Amsterdam University Press, Amsterdam.

Barbiers, S. and H. Bennis (2004). 'Reflexieven in dialecten van het Nederlands. Chaos of structuur?' In J. de Caluwe & G. de Schutter & M. Devos et al., eds., *Schatbewaarder van de taal. Johan Taeldeman. Liber Amicorum.* Academia Press Gent en Vakgroep Nederlandse Taalkunde Universiteit Gent, Gent, 43-58.

Bree, C. van (1992). 'The stability of language elements, in present-day eastern Standard-Dutch and eastern Dutch dialects'. In: J.A. van Leuvensteijn en J.B. Berns, Dialect and Standard language [...] in the English, Dutch, German and Norwegian language areas, Amsterdam enz., 178-203.

Bree, C. van (1994). 'The development of so-called Town Frisian'. In P. Bakker and Maarten Mous, eds., Mixed Languages. 15 Case Studies in Language Intertwining, Studies in Language and Language Use 13, IFOTT Amsterdam, 69-82.

Cornips, L. and W. Jongenburger (2001). 'Elicitation techniques in a Dutch syntactic dialect atlas project'. In H. Broekhuizen and T. van der Wouden, eds., Linguistics in the Netherlands 2001, John Benjamins, 53-63.

Daan, J. and D.P. Blok (1969). Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde, volume XXXVII of Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Goebl, H. (1982). Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie, volume 157 of Philosophisch-Historische Klasse Denkschriften. Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of W.-D. Rase and H. Pudlatz.

Heeringa W., and J. Nerbonne (2001), 'Dialect Areas and Dialect Continua'. In: David Sankoff, William Labov and Anthony Kroch, eds., Language Variation and Change 13, 2001, 375-400.

Heeringa, W. J. (2004). Measuring Dialect Pronunciation Differences using Levenshtein Distance, PhD thesis Rijksuniversiteit Groningen, Groningen.

Nerbonne, J. and P. Kleiweg (2005), 'Toward a Dialectological Yardstick', accepted to appear in: Journal of Quantitative Linguistics.

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

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

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika 17, 401-419.

Appendix

The following two tables show the mapping from atomic variables to feature variables related to reciprocals and one-pronominalisation in the reflexives subdomain as described in Section 7.

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

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

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

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