PROJECT DESCRIPTION

Quantity and quality in linguistics: reverse dialectometry

Jeroen van Craenenbroeck

1 Introduction

This project seeks to bridge the gap between formal-theoretical and quantitative-statistical linguistics. It combines a quantitative methodology with hypotheses taken from the formal-theoretical literature. On the one hand it investigates to what extent theoretical analyses of language phenomena can narrow down the hypothesis space created by quantitative-statistical analyses of large digital collections of language data, but on the other hand the quantitative findings also serve as a testing ground for the linguistic hypotheses. The project thus tries to establish a fruitful bidirectional collaboration between two subdisciplines of linguistics that traditionally do not communicate or collaborate much with one another. The research domain of the project will be verb clusters in Dutch dialects.

This project description is organized as follows. Section 2 introduces the research domain of this project, i.e. verb clusters. Section 3 describes the theoretical state of the art with respect to the proposal. It focuses both on the quantitative-statistical approach to dialectal variation (dialectometry, see subsection 3.2) and on the formal-theoretical literature on verb clusters (subsection 3.3). Section 4 introduces the main research questions, hypotheses, and objectives of the current project, while section 5 outlines the methodology. A key ingredient in that methodology will be a reversal of the dialectometric approach, so as to allow formal-theoretical hypotheses to be included in the quantitative analysis. Section 6 describes the work plan and work packages of the project. The research described here is designed to be carried out by a PhD-student as part of a four-year PhD-track.

2 Research domain: verb clusters

The research domain of this project is word order variation in clause-final verb clusters in varieties of Dutch. As will become clear in this and the following sections, verb clusters constitute the ideal topic for a combination of qualitative and quantitative research. On the one hand, the phenomenon has been extensively discussed and investigated in the formal-theoretical literature (see subsection 3.3 below), while on the other hand, recent dialect projects have unearthed a substantial amount of interdialectal variation which at present has not been analyzed yet and which seems to necessitate a more quantitative approach. To get a feel for the empirical richness of this domain, consider the example in (1).

(1) *Ik vind dat iedereen moet kunnen zwemmen.*
I find that everyone must can swim
'I think everyone should be able to swim.'

The embedded clause in this example contains a verb cluster consisting of three verbs: the main verb *zwemmen* 'swim' is selected by the modal *kunnen* 'can', which is in turn selected by *moet* 'must'. All three verbs cluster at the end of the clause, with the linear order reflecting the selectional hierarchy: the most deeply embedded verb is also rightmost in the cluster.[1] In three-verb

[1] As is customary in the literature on verb clusters, I use number combinations to refer to the various cluster orders. The cluster in (1) for example displays a 123-order, whereby '3' refers to the most deeply embedded verb of this three-verb cluster (i.e. *zwemmen* 'swim'), '2' refers to *kunnen* 'can', and '1' to *moet* 'must'.

clusters, there are six (three factorial) theoretically possible orders. However, a large-scale dialect investigation in 267 Dutch dialects in Belgium, France, and the Netherlands (the SAND-project, see Barbiers et al. (2005) and Barbiers et al. (2008)) has revealed that for the cluster type illustrated in (1)—i.e. modal-modal-infinitive—only four out of those six orders are actually attested:

(2) a. Ik vind dat iedereen moet kunnen zwemmen. (✓123)
b. Ik vind dat iedereen moet zwemmen kunnen. (✓132)
c. Ik vind dat iedereen zwemmen moet kunnen. (✓312)
d. Ik vind dat iedereen zwemmen kunnen moet. (✓321)
e. *Ik vind dat iedereen kunnen zwemmen moet. (*231)
f. *Ik vind dat iedereen kunnen moet zwemmen. (*213)

Moreover, it is not the case that in every one of those 267 dialects the orders in (2)a-d are well-formed: some allow one, some two, some three, and some four. In other words, there is a substantial amount of variation when it comes to which dialect allows which subset of these four cluster orders. For example, while in the dialect of Midsland (illustrated in (3)) only 132 and 321 are well-formed, Langelo Dutch (shown in (4)) only allows for 123 and 312.

(3) *Midsland Dutch*
a. *dat elkeen mot kanne zwemme.
that everyone must can swim
'that everyone should be able to swim.' (*123)
b. dat elkeen mot zwemme kanne. (✓132)
c. *dat elkeen zwemme mot kanne. (*312)
d. dat elkeen zwemme kanne mot. (✓321)
e. *dat elkeen kanne zwemme mot. (*231)
f. *dat elkeen kanne mot zwemme. (*213)

(4) *Langelo Dutch*
a. *dat iedereen mot kunnen zwemmen.*
that everyone must can swim
'that everyone should be able to swim.' (✓123)
b. *dat iedereen mot zwemmen kunnen. (*132)
c. dat iedereen zwemmen mot kunnen. (✓312)
d. *dat iedereen zwemmen kunnen mot. (*321)
e. *dat iedereen kunnen zwemmen mot. (*231)
f. *dat iedereen kunnen mot zwemmen. (*213)

More generally, there are 16 (two to the fourth power) possible subsets or combinations of word orders that a dialect can select from (2)a-d. Out of those 16 options, 12 are attested in the SAND-data. They are listed in Table 1, each accompanied by a sample dialect in which this particular combination occurs.

The amount of the variation increases further if we take into consideration *all* cluster orders that were part of the SAND-questionnaires. There was a total of eight questions in the questionnaire that dealt exclusively with verb cluster order. In combination these eight questions represent 31 possible cluster orders. If we now list, for each of the 267 SAND-dialects, which dialect has which combination of those 31 cluster orders, we arrive at 137 different verb cluster order patterns. It is precisely these types of data that will form the empirical basis for the current project. The PhD-student will extract from the raw SAND-data all information related to verb cluster ordering. This includes not only the eight questions referred to above, but also those sections of the questionnaires having to do with cluster interruption, *Infinitivus pro Participio*, the placement of *te* 'to', and the morphological shape of the past participle. The result will be a extensive and highly variable data set which will then be subjected to both a quantitative and a qualitative analysis. The next section sketches the state of the art in these two subfields of linguistics.

sample dialect	**123**	**132**	**321**	**312**
Beetgum	✓	✓	✓	✓
Hippolytushoef	✓	✓	✓	*
Warffum	✓	✓	*	*
Oosterend	✓	*	*	*
Schermerhorn	✓	✓	*	✓
Visvliet	✓	*	✓	✓
Kollum	✓	*	✓	*
Langelo	✓	*	*	✓
Midsland	*	✓	✓	*
Lies	*	*	✓	*
Bakkeveen	*	*	✓	✓
Waskemeer	*	✓	*	*

Table 1: Word order combinations in modal-modal-infinitive clusters in the SAND-dialects

3 Theoretical state of the art

3.1 Introduction

This section describes the state of the art in the two subdisciplines that form the backbone of the current research project: (i) the quantitative-statistical approach to language variation (subsection 3.2), and (ii) the formal-theoretical literature (subsection 3.3). As will become clear, the two approaches at present represent largely non-intersecting areas of research with little or no communication or collaboration between them. Forging a bridge between the two will be one of the main objectives of the current research project (cf. section 4).

3.2 Quantitative-statistical approaches

Like many of the social sciences, linguistics has seen a considerable surge in the use of mathematical and statistical methods in recent years. Although the quantitative approach to linguistics has a venerable tradition that dates back at least to the early twentieth century (see Zipf (1935) for a famous example), in the last two or three decades the volume of this type of research has increased dramatically. This quantitative shift has also affected traditional dialectology. Starting with the work of Jean Séguy (Séguy, 1973a,b,c) and Hans Goebl (Goebl, 1982, 1984), a new subdiscipline of linguistics was born which has come to be known as dialectometry. In a nutshell, it is an approach to language variation whereby computational and quantitative techniques are applied in dialectological research (see Nerbonne and Kretzschmar Jr. (2013) and Heeringa and Nerbonne (2013) for recent overviews and references). The key difference between dialectometry and traditional dialectology is the fact that dialectometric research aggregates over large numbers of linguistic features rather than focusing on individual ones, as is common in traditional dialectology. This allows for a much more nuanced and detailed view on the degree of difference or similarity between various dialect regions as well as the transition zones connecting them. As an illustration of how the dialectometric approach works, let us go through a simplified example.[2] Suppose we are looking at six dialect locations and ten linguistic features. For each dialect location and each linguistic feature we can indicate whether or not that feature occurs in that location. For our hypothetical example this can be represented as in Table 2, where '1' signals that a particular linguistic feature is present in a dialect location, while '0' signals its absence.

[2] I focus on one particular dialectometric technique here (mutidimensional scaling), as this is the one that will play a key role in the remainder of this project proposal. See Heeringa and Nerbonne (2013) and references mentioned there for a more complete overview of possible dialectometric methods.

dialect location	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}
Veurne	1	0	1	0	1	0	1	0	1	0
Gistel	1	0	1	0	0	1	1	0	1	0
Poelkapelle	1	0	1	0	1	1	0	0	1	1
Roeselare	1	1	1	1	0	1	0	0	0	0
Ieper	0	1	1	0	1	0	1	0	0	0
Brugge	0	1	1	1	0	1	0	0	1	1

Table 2: Hypothetical overview of 10 linguistic features in 6 dialect locations

In a next step of the analysis, the linguistic features listed in this data table are used to represent the degree of similarity or difference between the various dialect locations. The more linguistic features two locations have in common, the more alike they are, while if they share only a small subset (or even none) of the linguistic features, they are characterized as being very different. Formally, the data table is converted into a distance matrix, whereby each dialect location is compared to each other dialect location and a score (between 0 and 1) is assigned to each pair to indicate how dissimilar the two dialects are are. Applied to Table 2, this yields the distance matrix in Table 3.

	Veurne	**Gistel**	**Poelkapelle**	**Roeselare**	**Ieper**	**Brugge**
Veurne	0	0.333	0.428	0.75	0.5	0.777
Gistel	0.333	0	0.428	0.571	0.714	0.625
Poelkapelle	0.428	0.428	0	0.625	0.75	0.5
Roeselare	0.75	0.571	0.625	0	0.714	0.428
Ieper	0.5	0.714	0.75	0.714	0	0.75
Brugge	0.777	0.625	0.5	0.428	0.75	0

Table 3: Distance matrix based on the data in Table 2

The values in the various cells provide a measure for the degree of similarity—or rather, dissimilarity—between each pair of dialects.[3] For example, it makes precise the intuition that Veurne shares more dialect features with Gistel (distance: 0.333) than it does with Brugge (distance: 0.777). The next step in a typical dialectometric analysis involves the application of multidimensional scaling (MDS) to this distance matrix. MDS is a mathematical technique for reducing a multidimensional distance matrix to a low dimensional space in which each point represents an object from the distance matrix, and distances between points represents, as well as possible, dissimilarities between objects (Cox and Cox, 2001; Borg and Groenen, 2005). If we apply MDS to the six-dimensional distance matrix in Table 3 and reduce it to a two-dimensional Euclidean space, the result is the plot in Figure 1.

Each of the six dialect locations is represented on this two-dimensional plane. The closer two locations are together, the more linguistic features they share, i.e. Euclidean distance in this plot represents dissimilarity in the distance matrix in Table 3. As is clear from the plot, the six dialects can be roughly split into three groups with respect to the (hypothetical) linguistic features under consideration here: Roeselare and Brugge pattern together, as do Gistel and Poelkapelle, and—to a slightly lesser extent—Veurne and Ieper. The final step of the dialectometric analysis involves projecting these dialect locations back onto a geographical map. The idea behind this approach is Nerbonne and Kleiweg (2007)'s so-called Fundamental Dialectological Postulate, which states that geographically proximate varieties tend to be more similar (linguistically) than distant ones. In other words, based on the plot in Figure 1 we would expect Roeselare and Brugge

[3] As is clear from Table 3, a distance matrix is symmetrical across the diagonal (because the distance between dialect A and dialect B is identical to the distance between B and A) and contains only zeroes on the diagonal (because every dialect is non-distinct from itself).

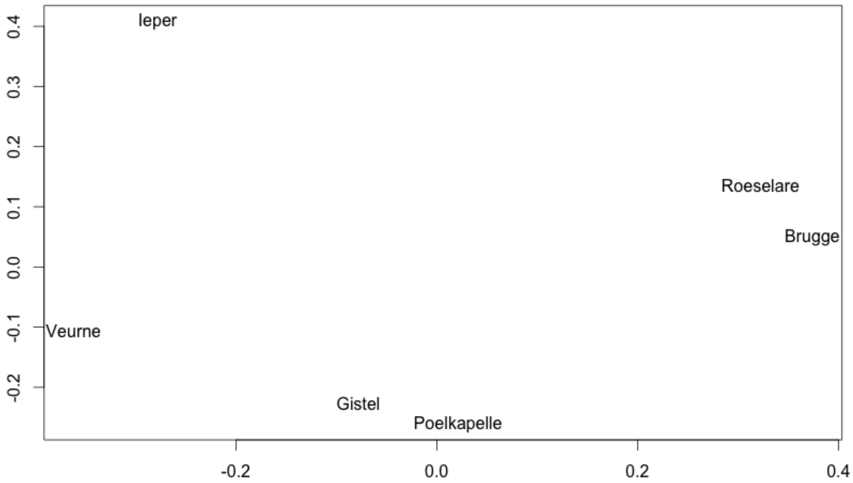


Figure 1: Two-dimensional MDS-representation of the distance matrix in Table 3

to be geographically far apart from Ieper and Veurne, with Gistel and Poelkapelle occupying an intermediate position. As the map in Figure 2 shows, this is indeed the case.

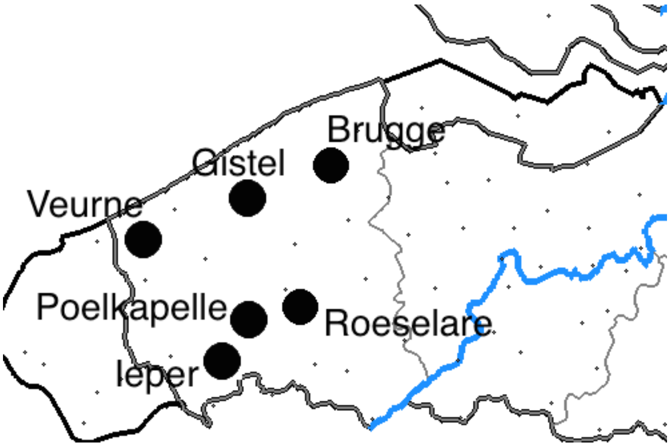


Figure 2: Geographical representation of the locations in the MDS-plot in Figure 1

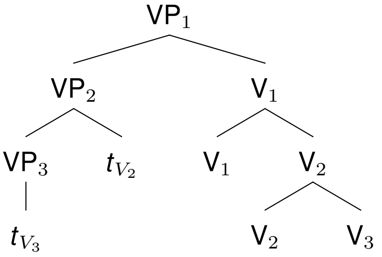
Although this example is purely hypothetical and based on non-existing data, it serves the purpose of highlighting some of the central characteristics of dialectometric analyses. What most of them have in common is that they use a statistical analysis of large numbers of linguistic features to quantify the degree of similarity or difference between dialect locations. This allows them to identify different dialect areas, dialect continua, or transitional zones. However, the formal-theoretical description of these linguistic features plays no role whatsoever in the analysis. The linguistic properties of the dialect locations under investigation are treated as binary categorical variables and their treatment in the theoretical literature does not enter into the discussion. This is where the current project will diverge from classical dialectometry (see section 5). Moreover, quantitative research that deals explicitly with word order variation in verb clusters is extremely rare. A notable contribution is the work of Gert De Sutter (see in particular De Sutter (2009)), but he only focuses on two-verb clusters in written Standard Dutch. Spruit (2008, 129) mentions the SAND-data on verb clusters in passing and provides one MDS-map, but no analysis, while Augustinus and Van Eynde (2014) only focus on verb cluster interruption in (written and spoken) Standard Dutch. The research project described here will thus provide the first ever quantitative analysis of a wide range of verb cluster variation data in Dutch.

3.3 Formal-theoretical approaches

Verb clusters have taken center stage in the generative literature on Germanic for the past four decades (see in particular the overview in Wurmbrand (2005)). Ever since Evers (1975) it became clear that the theoretical analysis of this phenomenon has far-reaching consequences for the theory of grammar in general. The most straightforward illustration of this concerns headedness. As was discussed in detail by Evers, a simple 123-order such as the one in (1) (repeated below as (5)) poses a considerable challenge for the hypothesis that Dutch is head-final: if complements precede their selecting heads in Dutch, the main verb *zwemmen* 'swim' should precede the modal *kunnen* 'can', which should in turn precede *moet* 'must'. In other words, typological considerations lead us to expect a 321-order in three-verb clusters, but while this order is common in German, it is not in Standard Dutch. This led Evers to propose a theory of head movement, whereby lower verbs raise to higher ones, thus forming complex verbal heads. His analysis is represented in (6).

(5) *Ik vind dat iedereen moet kunnen zwemmen.*
I find that everyone must can swim
'I think everyone should be able to swim.'

(6)



Evers's analysis has by no means remained uncontested, however. In the wake of Kayne (1994)'s antisymmetric framework, various authors have proposed that in spite of first appearances, Dutch is a head-initial language. This means that the 123-order in (5) can be base-generated—thus obviating the need for the various movement operations in (6)—but it introduces new complications in other cluster types (see Zwart (1997) and Barbiers (2005) for discussion). In addition, some authors have proposed that both orders can be base-generated (Barbiers and Bennis, 2010), while still others have put forward the hypothesis that any order that respects the hierarchical representation can be base-generated (Abels, 2011).

The theory of headedness is not the only aspect of the morphosyntax of Dutch—and Germanic more generally—that verb clusters have shed new light on. Over the years it became clear that cluster formation in West-Germanic is intimately connected with other aspects of the morphosyntax of the languages making up this subfamily. Well-known examples are the *Infinitivus pro Participio*-effect, whereby an infinitive occurs in lieu of a past participle inside a cluster, see Zwart (2007), cluster interruption (also known as Verb Projection Raising, Haegeman and van Riemsdijk (1986); Augustinus and Van Eynde (2014)), the placement and possible deletion of *te* 'to' (IJbema, 2001), ordering restrictions on PP-complements and -modifiers in the Dutch VP (Barbiers, 2008), and even typological generalizations about the word order of DP-internal nominal modifiers (Abels, 2011).

In short, the formal-theoretical literature on verb clusters is vast and covers many aspects of this phenomenon in great detail. At the same time, however, there is no theoretical literature dealing with the range of variation that was outlined in section 2. While recent years has seen an increasing number of papers on verb cluster data from non-standard language varieties (see e.g. Wurmbrand (2005) and references cited there), there are no proposals to date that attempt to provide a parametric theoretical analysis of verb cluster ordering in a wide range of West-Germanic languages or dialects. Similarly lacking in the theoretical literature are the insights and methodologies of quantitative-statistical approaches to language variation (see the previous subsection).

4 Research questions, hypotheses, and objectives

The previous section has shown there to be a potentially fruitful complementarity between formal-theoretical and quantitative-statistical approaches to language variation. While the latter use linguistic features as mere categorical variables and do not incorporate the insights and hypotheses from theoretical analyses, the former are methodologically ill-equipped to deal with large amounts of variation and very often start out from a simplified empirical picture. The present project aims to bring these two approaches together. Its central research question can be formulated as follows:

(7) **Central research question**
Can quantitative-statistical and formal-theoretical approaches to language variation be fruitfully combined into a single analysis, and if so, how?

As should be clear from the preceding discussion, this project starts out from the hypothesis that the question in (7) can be answered affirmatively. More specifically, I hypothesize that the two approaches can benefit from one another in two specific ways:

(8) **Central hypotheses**

1. Theoretical analyses of specific linguistic phenomena can guide the interpretation of quantitative results by winnowing down the space of hypotheses.
2. Statistical analyses of large digital language collections can serve as a touchstone for theoretical analyses of specific linguistic phenomena.

As for the first hypothesis, it is a well-known characteristic of quantitative-statistical approaches that they yield a large number of significant results, and that the most challenging part of the analysis is interpreting and cutting down those results. As an illustration, consider Spruit (2008), who provides a quantitative, dialectometric analysis of the first half of the data collected in the SAND-project. When looking for associations between 485 syntactic variables—associations of the type 'If a dialect has property A, what are the odds that it also had property B?'—he finds no less than 10,730 with an accuracy of 90 percent or higher,[4] and when either the antecedent or the consequent is allowed to contain a disjunction, that number even goes up to fifty-six million. Spruit's conclusion when faced with these results is telling: "From a statistical perspective many more linguistically interesting variable associations can be expected to surface upon closer investigation. (..) However, every approach will require extensive consultation with syntactic theorists to meaningfully interpret the data." (Spruit, 2008, 106) This project intends to follow up on this suggestion, by incorporating theoretical insights into the quantitative-statistical analyses (see below, section 5).

The second hypothesis in (8) starts out from the opposite perspective. It examines to what extent theoretical analyses can benefit from quantitative results and methodologies. Many theoretical analyses are designed with only a relatively limited set of data in mind, which in the past has earned them the—sometimes justified—criticism of oversimplifying the empirical situation. To the extent that such analyses are formally sufficiently explicit, though, it should be possible to quantitatively test them against large data collections or even to compare one theory against another.

Against the background of the research question in (7) and the hypotheses in (8), the objectives of the project are threefold. They are listed in (9).

(9) **Central objectives:**

[4] Spruit (2008, 98) defines accuracy as the number of dialects where both phenomena occur divided by the number of dialects where only the first one occurs.

1. to further our theoretical understanding of word order variation in verb clusters in West-Germanic in general and Dutch more specifically;
2. to develop a quantitative-statistical methodology in which insights and hypotheses from the theoretical literature can be successfully integrated, compared, and tested;
3. to lay the foundation for a rapprochement between two subfields of linguistics that typically do not communicate or collaborate much with one another.

These are ambitious objectives—especially the third one—but what little literature there is that tries to combine and integrate quantitative and qualitative approaches (e.g. Spruit (2008); Wieling and Nerbonne (2010); Sauerland and Bobaljik (2013)) suggests that it is an endeavour worth pursuing. The next section describes the methodology that will be adopted towards achieving these objectives.

5 Methodology: reverse dialectometry

As became clear in section 3.2, the dialectometric method as it stands is ill-suited to incorporate insights from formal-theoretical analyses. The present project wants to tackle this problem by making two changes to this approach. The first one involves a reversal of perspective: rather than using linguistic phenomena to quantify the degree of similarity or difference between dialect locations, it uses dialect locations—or more specifically, the geographic distribution of linguistic phenonema across various dialects—as a means to study similarities and differences between linguistic phenomena. In order to make this more precise, let us revisit the hypothetical example from section 3.2 from this reverse dialectometric point of view. The data are now presented not as being about dialect locations, but about linguistic properties:

linguistic property	Veurne	Gistel	Poelkapelle	Roeselare	Ieper	Brugge
L_1	1	1	1	1	0	0
L_2	0	0	0	1	1	1
L_3	1	1	1	1	1	1
L_4	0	0	0	1	0	1
L_5	1	0	1	0	1	0
L_6	0	1	1	1	0	1
L_7	1	1	0	0	1	0
L_8	0	0	0	0	0	0
L_9	1	1	1	0	0	1
L_{10}	0	0	1	0	0	1

Table 4: Reverse overview of 10 hypothetical linguistic features in 6 dialect locations

Just like the data in Table 2, these facts can be converted into a distance matrix, cf. Table 5. However, the main difference with the distance matrix in Table 3 is that in this case linguistic properties are compared to linguistic properties. In other words, in the reverse dialectometric approach it is the geographical information that is used as binary categorical variables, while the linguistic constructions themselves become the main focus of the analysis. This distance matrix can now serve as the input for an analysis in terms of multidimensional scaling, which yields the plot in Figure 3. This plot provides a visual representation of the degree of similarity or difference between the ten (hypothetical) linguistic features in Table 4: the closer two features are together, the more they co-occur in the same dialect locations.

This analysis will now form the input for the second methodological innovation of this project, i.e. the link between the quantitative results and the formal-theoretical analysis. To the extent that theoretical analyses of L_1-L_{10} are on the right track, they should predict the clustering depicted in Figure 3. Put differently, linguistic constructions that are theoretically alike (e.g. that are generated by one and the same parameter setting), should also cluster together geographically. To make this more precise, let us turn again to verb clusters. Recall that this phenomenon has played and

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}
L_1	0	0.833	0.333	0.8	0.6	0.4	0.6	1	0.4	0.8
L_2	0.833	0	0.5	0.333	0.8	0.6	0.8	1	0.833	0.75
L_3	0.333	0.5	0	0.666	0.5	0.333	0.5	1	0.333	0.666
L_4	0.8	0.333	0.666	0	1	0.5	1	1	0.8	0.666
L_5	0.6	0.8	0.5	1	0	0.833	0.5	1	0.6	0.75
L_6	0.4	0.6	0.333	0.5	0.833	0	0.833	1	0.6	1
L_7	0.6	0.8	0.5	1	0.5	0.833	0	1	0.6	1
L_8	1	1	1	1	1	1	1	0	1	1
L_9	0.4	0.833	0.333	0.8	0.6	0.4	0.6	1	0	05
L_{10}	0.8	0.75	0.666	0.666	0.75	0.5	1	1	0.5	0

Table 5: Distance matrix based on the data in Table 4

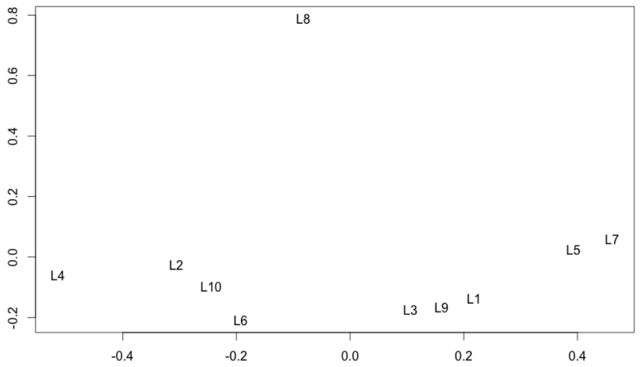


Figure 3: Two-dimensional MDS-representation of the distance matrix in Table 5

continues to play an important role in the debate on headedness in Dutch. Interestingly, different theories of headedness predict different groupings of verb cluster orders. Consider in this respect the overview in Table 6. It shows which three-verb cluster orders can be base-generated under which theories.

			mixed$_1$	**mixed**$_2$
cluster order	**head-final**	**head-initial**	(Barbiers and Bennis, 2010)	(Abels, 2011)
123	no	yes	yes	yes
132	no	no	no	yes
231	no	no	no	yes
213	no	no	no	no
321	yes	no	yes	yes
312	no	no	no	no

Table 6: Overview of cluster orders that can be base-generated under four theories of headedness

A head-final theory predicts the 321-order to be ‘special’, i.e. to be set apart from all the others, as this is the only one that does not require any reordering operations. Suppose, for the sake of the argument, that this order corresponds to L_8 in Figure 3. In that case, the quantitative-statistical analysis of the variation data would provide a strong argument in favor of the head-final approach to headedness, as the order that is theoretically ‘special’ turns out to be singled out empirically as well. A head-initial theory on the other hand would fare rather poorly in this respect, as it would predict the 123-order to stand out. Similarly, the mixed theory of Barbiers and Bennis (2010) expects 123 and 321 to be grouped together and set apart from the other orders, while in the mixed theory of Abels (2011) 312 and 213 should go together, as these are the only orders that can*not* be base-generated. More generally, though, what this project proposes to do—and this is the second methodological innovation alluded to above—is encode verb cluster orders not just in terms of their geographical distribution (as schematically illustrated in Table 4), but also in terms of their formal-theoretical analysis as in Table 6. The output of the geographical analysis (cf. Figure 3) can then be mapped or matched against the theoretical characterization, thus allowing us to attain the twofold goal outlined in section 4: on the one hand, theoretical analyses can be used to interpret and understand MDS-plots such as the one in Figure 3, while on the other hand, the statistical results can be used to test and evaluate theoretical hypotheses (as was schematically illustrated with respect to the headedness data in Table 6).

6 Work plan

As was pointed out in the introduction, the research described here is designed to be carried out by a PhD-student as part of a four-year PhD-track. This means that the various steps and work packages that are part of this project will all be carried out by that PhD-student, though it should be clear that every aspect of the project will proceed in close collaboration with and under direct supervision of the PhD supervisor. The various work packages are listed below and represented schematically in the Gantt-diagram in Figure **??**.

1. **Year 1 (October 2015–September 2016):**
   (a) literature review on verb clusters
   (b) PhD-training in syntax, morphology, statistics
   (c) data collection and first statistical analyses
2. **Year 2 (October 2016–September 2017):**
   (a) continuation of statistical analyses
   (b) operationalization of the theoretical literature
   (c) first test of the theoretical analyses
   (d) formulation of first hypotheses regarding word order in verb clusters
   (e) presentation of first results at national conferences or workshops (e.g. BKL or LIN)
3. **Year 3 (October 2017–September 2018):**
   (a) further development of theoretical analysis
   (b) further testing of the theory based on quantitative results and analyses
   (c) presentation of results at international conferences and workshops (QITL, Methods in Dialectology, CGSW, NELS, GLOW)
   (d) publication of results in (inter)national journals
   (e) writing of the first chapters of the dissertation
4. **Year 4 (October 2018–September 2019):**
   (a) completion of the first draft of the dissertation
   (b) organization of a workshop on the interaction between quantitative and qualitative linguistics
   (c) finalizing the dissertation & preparation for the defense
   (d) public defense of the dissertation

References

Abels, Klaus. 2011. Hierarchy-order relations in the germanic verb cluster and in the noun phrase. *GAGL* 53:1–28.

Augustinus, Liesbeth, and Frank Van Eynde. 2014. Looking for cluster creepers in Dutch treebanks. Dat we ons daar nog kunnen mee bezig houden. Ms. KU Leuven (to appear in the proceedings of CLIN 24).

Barbiers, Sjef. 2005. Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics. In *Syntax and variation. Reconciling the biological and the social*, ed. Leonie Cornips and Karen P. Corrigan, volume 265 of *Current issues in linguistic theory*, 233–264. John Benjamins.

Barbiers, Sjef. 2008. Werkwoordclusters en de grammatica van de rechterperiferie. *Nederlandse Taalkunde* 13:160–197.

Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet van der Ham. 2008. *Syntactische atlas van de Nederlandse dialecten. Deel II*. Amsterdam: Amsterdam University Press.

Barbiers, Sjef, and Hans Bennis. 2010. De plaats van het werkwoord in zuid en noord. In *Voor Magda. Artikelen voor Magda Devos bij haar afscheid van de Universiteit Gent*, ed. Johan De Caluwe and Jacques Van Keymeulen, 25–42. Gent: Academia.

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos, and Margreet van der Ham. 2005. *Syntactische atlas van de Nederlandse dialecten. Deel I*. Amsterdam: Amsterdam University Press.

Borg, Ingwer, and Patrick J.F. Groenen. 2005. *Modern Multidimensional Scaling. Theory and applications.*. Springer, 2nd edition.

Cox, Trevor F., and Michael A.A. Cox. 2001. *Multidimensional scaling*. Boca Raton: Chapman & Hall.

Evers, Arnold. 1975. *The transformational cycle in Dutch and German*. Bloomington, Indiana.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, volume 157 of *Philosophisch-Historische Klasse Denkschriften*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Goebl, Hans. 1984. *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, volume 191–193 of *Beihefte zur Zeitschrift für romanische Philologie*. Tübingen: Max Niemeyer Verlag.

Haegeman, Liliane, and Henk van Riemsdijk. 1986. Verb projection raising, scope, and the typology of verb movement rules. *Linguistic Inquiry* 17:417–466.

Heeringa, Wilbert, and John Nerbonne. 2013. Dialectometry. In *Language and Space. An International Handbook of Linguistic Variation. Volume 3: Dutch*, ed. Frans Hinskens and Johan Taeldeman, volume 30 of *Handbooks of Linguistics and Communication Science*, 624–645. De Gruyter.

IJbema, Aniek. 2001. Grammaticalization and infinitival complements in Dutch. Doctoral Dissertation, Leiden University.

Kayne, Richard S. 1994. *The antisymmetry of syntax*. Cambridge, Massachusetts: MIT Press.

Nerbonne, John, and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14:148–166.

Nerbonne, John, and William A. Kretzschmar Jr. 2013. Dialectometry++. *Literary and Linguistic Computing* 28:2–12.

Sauerland, Uli, and Jonathan Bobaljik. 2013. Syncretism distribution modeling: accidental homophony as a random event. In *Proceedings of GLOW in Asia 9*, ed. Nobu Goto, Otaki Koichi, Atsushi Sato, and Kensukei Takita, 31–53. Tsu, Japan: University of Mie.

Séguy, Jean. 1973a. *Atlas linguistique de la Gascogne; complément du volume VI*. Paris: Centre national de la recherche scientifique.

Séguy, Jean. 1973b. *Atlas linguistique de la Gascogne; volume VI.*. Paris: Centre national de la recherche scientifique.

Séguy, Jean. 1973c. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique Romane* 37:1–24.

Spruit, Marco René. 2008. Quantitative perspectives on syntactic variation in Dutch dialects. Doctoral Dissertation, Universiteit van Amsterdam.

de Sutter, Gert. 2009. Towards a multivariate model of grammar: the case of word order variation in Dutch clause final verb clusters. In *Describing and modeling variation in grammar*, ed. Andreas Dufter, Jörg Fleischer, and Guido Seiler. Mouton de Gruyter.

Wieling, Martijn, and John Nerbonne. 2010. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25:700–715.

Wurmbrand, Susanne. 2005. Verb clusters, verb raising, and restructuring. In *The Blackwell Companion to Syntax*, ed. Martin Everaert and Henk van Riemsdijk, volume V, chapter 75, 227–341. Oxford: Blackwell.

Zipf, George K. 1935. *The psychobiology of language*. Boston: Houghton-Mifflin.

Zwart, C. Jan-Wouter. 1997. *Morphosyntax of verb movement: A minimalist approach to the syntax of Dutch*. Dordrecht: Kluwer Academic Publishers.

Zwart, Jan-Wouter. 2007. Some notes on the origin and distribution of the IPP-effect. *Groninger Arbeiten zur Germanistischen Linguistik* 45:77–99.