The signal and the noise in Dutch verb clusters

A quantitative search for parameters

Jeroen van Craenenbroeck
KU Leuven/CRISSP
jeroen.vancraenenbroeck@kuleuven.be

December 18, 2014

Abstract

This paper combines quantitative-statistical and formal-theoretical approaches to language variation. I provide a quantitative analysis of word order variation in verb clusters in 267 dialects of Dutch and map the results of that analysis against hypotheses extracted from the theoretical literature on verb clusters. Based on this new methodology, I argue that variation in verb cluster ordering in Dutch dialects can be largely reduced to three grammatical parameters.

1 Introduction

Typological studies into linguistic phenomena typically reveal a bewildering amount of variation and flexibility on the one hand combined with (near-)universal rigidity on the other. For instance, when we consider the possible orderings of demonstratives, numerals, adjectives, and nouns inside the noun phrase, the languages of the world use no less than 14 different orders as their neutral word order, but at the same time this means that of the 24 (four factorial) theoretically possible orders, 10 are universally unattested (Greenberg (1963); Cinque (2005)). The job of the comparative linguist, then, is to separate that which is fixed and necessary—the principles, in generative parlance—from that which is variable and contingent—the parameters. Those parameters can be seen as the smallest units of linguistic variation. Kayne (2000) has argued that a detailed comparison of large numbers of closely related languages or language varieties presents a new research tool towards

I would like to thank Sjef Barbiers, Marijke De Belder, Hans Bennis, Jonathan Bobaljik, Lotte Hendriks, Frans Hinskens, Dany Jaspers, Richard Kayne, Ad Neeleman, Marc van Oostendorp, Ian Roberts, Koen Roelandt, Martin Salzmann, Tanja Temmerman, Guido Vanden Wyngaerd, and audiences at *Mapping Methods* (Tartu, May 2014), *What happened to Principles & Parameters?* (Arezzo, July 2014), *Methods in Dialectology XV* (Groningen, August 2014), *Maps and grammar* (Amsterdam, September 2014), and *Dialect syntax: the state of the art* (Frankfurt am Main, December 2014) for discussion, comments, and references.

uncovering linguistic parameters, the idea being that such a comparison is the closest real-world alternative to a controlled laboratory experiment: one tries to keep all non-relevant variation constant so as to be able to examine the effect of minute changes to the phenomenon under investigation. In this paper I follow up on Kayne's lead by examining in detail a specific linguistic phenomenon (word order in clause-final verb clusters) in 267 varieties of Dutch. I argue that we can distill grammatical parameters from this large dataset by looking for statistical patterns in the data and mapping those against the insights gleaned from the theoretical literature on verb clusters. The resulting picture is one in which quantitative-statistical and formal-theoretical approaches to linguistics go hand in hand, mutually benefiting from one another.

The paper is organized as follows. The next section introduces the data that will form the basis for the analysis. As will become clear, even if we restrict ourselves to a relatively confined empirical domain such as verb cluster ordering, the amount of attested variation is considerable and a statistical approach quickly becomes appealing. Section 3 introduces and details my methodology, which I call Reverse Dialectometry, because it reverses the perspective typically taken in dialectometric work (see e.g. Nerbonne and Kretzschmar Jr. (2013) and references mentioned there). While such work examines differences and similarities between dialect locations based on their linguistic profile, I focus on the differences and similarities between linguistic phenomena—more specifically, verb cluster orders—based on their geographical distribution. Section 4 presents the results of this analysis. I show that the dialect Dutch verb cluster variation can be reduced to three main dimensions and indicate for each of those dimensions how well they align with the theoretical literature on verb clusters. Section 5 then interprets these results from a formal-theoretical point of view. I propose to identify the three relevant dimensions as grammatical parameters and sketch the outlines of a parametric account of verb cluster ordering. Section 6 concludes the paper and discusses some prospects.

2 The data: variation in Dutch verb clusters

As is well-known, verbs tend to cluster at the right-hand side of the clause in Dutch and other (mainly) West-Germanic languages (see Wurmbrand (2005) for extensive discussion and references). Consider in this respect a simple example of a two-verb cluster in (1).

(1) a. *dat hij heeft gelachen.*
that he has laughed
'that he has laughed.'
b. *dat hij gelachen heeft.*
that he laughed has
'that he has laughed.'

The perfective auxiliary *heeft* 'has' can either precede or follow the past participle it selects, in this case *gelachen* 'laughed'. The two examples mean exactly the same thing and for most—if not all—speakers of Standard Dutch the choice between them is more or less

optional.[1] Let us pretend, however, for the sake of the argument, that the data in (1) stem from two different dialects, with dialect A only allowing the order auxiliary–participle and dialect B only the opposite one. One could then postulate a parameter that captures this difference and argue that dialect A and dialect B have a different setting for this parameter. For instance, starting out from a head-initial base structure, one could argue that in dialect B the participle has moved across its auxiliary, while in dialect A it has stayed put. This would yield the following parameter setting:

(2) a. dialect A: [−MoveParticipleAcrossAux]
b. dialect B: [+MoveParticipleAcrossAux]

Needless to say, actual linguistic data are nowhere near as clear-cut or black-and-white as this hypothetical example. In order to appreciate this, it suffices to include three-verb clusters into the discussion. An example is given in (3).

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

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. As is customary in the literature on verb clusters, I will use number combinations to refer to the various cluster orders. The cluster in (3) 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'.[2] In three-verb 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 (3)—i.e. modal-modal-infinitive—only four out of those six orders are actually attested:

(4) 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 (4-a)–(4-d) are well-formed. Quite the contrary, there is a substantial amount of variation when it comes to which dialect allows which subset of these four cluster orders. For example, while

[1] See De Sutter (2009) for a detailed analysis of the different factors that influence the word order of two-verb clusters.

[2] Similarly, the clusters in (1-a) and (1-b) display the orders 12 and 21 respectively.

in the dialect of Midsland (illustrated in (5)) only 132 and 321 are well-formed, Langelo Dutch (shown in (6)) only allows for 123 and 312.

(5) *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)

(6) *Langelo Dutch*

a.	*dat iedereen moet 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 (4-a)–(4-d).[3] 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.

It should be clear that a data pattern such as this one is not straightforwardly amenable to the type of (overly) simple parameter account outlined above. Looking at the combinations in table 1, it is not obvious which parameters are responsible for this variation or even how to go about trying to identify those parameters. Things get even worse when we expand our empirical viewpoint further and consider *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.[4] They are briefly described in (7).

(7)
a. three questions about two-verb clusters of the type auxiliary-participle
b. one question about two-verb clusters of the type modal-infinitive
c. one question about three-verb clusters of the type modal-modal-infinitive
d. one question about three-verb clusters of the type modal-auxiliary-participle

[3] 15 if we exclude the option whereby none of the orders is allowed in the dialect in question. That would be a dialect in which three-verb clusters of the type modal-modal-infinitive simply do no occur. As far as I know, no such dialect exists in Dutch.

[4] There were a number of related questions having to do with cluster interruption and the IPP-effect, which I do not take into consideration here.

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

e. one question about three-verb clusters of the type auxiliary-auxiliary-infinitive
f. one question about three-verb clusters of the type auxiliary-modal-infinitive

The cluster types in (7-a) and (7-c) were already illustrated above (see examples (1) and (3) respectively). For the four remaining ones I provide representative examples below.

(8) a. *dat jij het niet* ***mag zien.***
that you it not may see
'that you're not allowed to see it.' (modal-infinitive)

b. *dat hij haar* ***moet hebben gezien.***
that he her must have seen
'that he must have seen her.' (modal-auxiliary-participle)

c. *dat hij* ***is gaan zwemmen.***
that he is go swim
'that he went for a swim.' (auxiliary-auxiliary-infinitive)

d. *dat hij mij* ***had kunnen roepen.***
that he me had can call
'that he could have called me.' (auxiliary-modal-infinitive)

Together, the eight questions listed in (7) yielded a total of 31 clusters 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. Put differently, when considering the verb cluster order data from the SAND-questionnaire, we can discern 137

different dialect types.[5]

What does this mean for parameter theory? In its purest form, the theory of parameters assumes that any and every observable morphosyntactic difference between two languages should be reducible to a different setting for at least one parameter. Applied to the case at hand, this would mean that each of the 137 different dialect types differs from all the others in at least one parameter setting. Alternatively, it could be that some of the variation found in the SAND-database is due to noise, i.e. to extra-grammatical factors ranging from sociolinguistic variation due to gender, register, social, or geographical norms, over cases of dialect mixing or influence from the standard language all the way to recording, transcription, or even speech errors. What we should try to do, then, is separate the signal from the noise and determine which portions of the variation are due to grammar proper and hence should follow from grammatical theory. Barbiers (2005) takes this approach in his analysis of verb clusters. With respect to the modal-modal-infinitive clusters introduced in (4), he proposes that the grammar rules out the 231- and the 213-order (see below, section 3.3, for details), but that the four remaining orders are grammatical in all varieties of Dutch. Any interdialectal differences that we find with respect to these orders—such as the contrast between Midsland Dutch and Langelo Dutch in (5) and (6)—is then due to sociolinguistic factors such as "geographical and social norms as well as considerations of register and context" (Barbiers (2005, 234–235)).

In this paper I follow the same general principle as Barbiers—i.e. I assume that the SAND-data contain a certain amount of noise—but with a different methodology and a different conclusion. I show how a statistical analysis of the SAND-data on verb clusters can be mapped against the findings from the formal-theoretical literature on this phenomenon and conclude that the majority of the variation we find can be reduced to the interaction between three grammatical microparameters. The next section describes my methodology in more detail.

3 Methodology: Reverse Dialectometry

3.1 Introduction

This section outlines the methodology I adopt in analyzing the SAND-data on verb cluster order. I call it Reverse Dialectometry, because it reverses the perspective taken in much dialectometric work (see e.g. the references mentioned in Nerbonne and Kretzschmar Jr. (2013)). A dialectometric analysis tries to model, visualize, and analyze similarities and differences between dialect locations based on the linguistic phenomena that are attested in those locations (see e.g. Spruit (2008); Nerbonne (2010); Szmrecsanyi (2012)). Typi-

[5]Not every question from the questionnaire was asked in every dialect location, i.e. the data table contains a number of gaps (see section 3.2 below for detailed discussion). Those gaps were not taken into consideration when counting the number of cluster order patterns, which means that 137 is a conservative estimate. If those gaps turn out to conceal even more variation, the actual number of patterns could be as high as 207.

cal dialectometric research questions include the distinction between dialect regions and dialect continua, or the correlation between linguistic distance and (various measures of) geographical distance.[6] This paper takes a different approach: my analysis is focused on modeling and understanding the differences and similarities between *linguistic constructions*—cluster orders to be precise—based on their geographical distribution. In a nutshell, cluster orders that occur in the same locations will be assumed to be more alike than those that have a different geographical distribution. Note that in this setup it is largely irrelevant whether or not those dialect locations form a contiguous region. The locational data are merely used as binary variables that sketch a detailed empirical picture, which can then subsequently be matched and mapped against additional variables taken from the theoretical literature on verb clusters.

This section is organized as follows. In the next subsection I describe how the raw data from the SAND-questionnaires were preprocessed so as to make them amenable to a statistical analysis. Subsection 3.3 outlines the difference between active (locational) and supplementary (linguistic) variables and makes clear how the latter were extracted from the theoretical literature and how they were operationalized. Subsection 3.4 describes the Multiple Correspondence Analysis I performed on the data set. All calculations were carried out using the `FactoMineR`-package (Husson et al. (2014)) in `R` (R Core Team (2014)). The results of the analysis are presented in section 4.

3.2 Preparation of the data

All data discussed and analyzed in this paper come from the SAND-project. As pointed out above, this four-year dialect atlas project (2000–2004) investigated the variety of Dutch spoken in 267 dialect locations in Belgium, France, and the Netherlands, and has yielded two atlases (Barbiers et al. (2005) and Barbiers et al. (2008)). The SAND-data stem from three sources: a written questionnaire, a series of oral dialect interviews, and a additional set of telephone interviews (see Cornips and Jongenburger (2001) for a detailed description of the SAND-methodology). The analysis carried out in this paper uses the raw data from the oral dialect interviews, which also form the basis for the maps in Barbiers et al. (2005) and Barbiers et al. (2008).[7] The data files contain a list of all the data points contained in the two atlases, including information about the type of phenomenon under investigation, the number of the map and atlas, the element listed in the legenda of the map, and of course the dialect location where the phenomenon was attested. I have converted these data into a 31×267-matrix, whereby each verb cluster order occupies a row and each dialect location a column. Cells are filled by *yes* when that cluster order is attested in that dialect location, *no* when it is absent, and *NA* when the question pertaining to that cluster order was not asked in that dialect location. A small sample of my data table—the upper left-hand corner—can be found in table 2.

[6] Well-known in this respect 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.

[7] Many thanks to Jan Pieter Kunst of the Meertens Institute for giving me access to these data.

	Midsland	**Lies**	**West-Terschelling**	**Oosterend**	...
AUX1(be.sg)-PART2	no	no	no	NA	...
PART2-AUX1(be.sg)	yes	yes	yes	NA	...
AUX1(have.sg)-PART2	no	no	no	no	...
PART2-AUX1(have.sg)	yes	yes	yes	yes	...
AUX1(have.pl)-PART2	no	no	no	no	...
PART2-AUX1(have.pl)	yes	yes	yes	yes	...
MOD1(sg)-INF2	no	no	yes	no	...
INF2-MOD1(sg)	yes	yes	yes	yes	...
MOD2-INF3-MOD1(sg)	no	no	no	no	...
MOD1(sg)-MOD2-INF3	no	no	no	yes	...
MOD1(sg)-INF3-MOD2	yes	no	no	no	...
...	...	...	...	...	...

Table 2: Upper left-hand portion of the data table used for the analysis

The first row of this table contains data pertaining to a two-verb cluster consisting of a singular form of *to be* used as an auxiliary followed by a participle, i.e. a 12-order. As the values in the subsequent cells indicate, this order is not attested in the varieties spoken in Midsland, Lies, and West-Terschelling, while the question pertaining to this order was not part of the dialect interview in Oosterend. The second line provides the data for the 21-order of that same cluster, and the remaining rows provide similar information for other verb cluster orders.

Two aspects of the conversion from raw data to the table partially represented above are worth commenting on. The first concerns concerns the question methodology used in the oral dialect interviews of the SAND-project. The data pertaining to verb clusters are based on two types of questions: translation tasks and elicitation questions. In the former, the informants were given a Standard Dutch sentence and were asked to translate it into their dialect, while the latter involved (pre-recorded) oral versions of dialect sentences for which they had to provide a grammaticality judgment (see Cornips and Jongenburger (2001) for a more detailed description of the various question methodologies used in the SAND-project).[8] For verb clusters this means that while in elicitation questions every possible cluster order was presented to the informants and explicitly judged by them,[9] in

[8]Of the 31 cluster orders that were tested, 16 are based on translation tasks, corresponding to 3 of the cluster types listed in (7), i.e. (7-a), (7-b), and (7-f).

[9]This is not strictly true for all clusters: for a number of them, the extensive written questionnaire and dialect literature review that preceded the dialect interviews had turned up systematic gaps in verb cluster ordering. Orders that were unattested in any variety of Dutch were not included in the oral interviews, not even in elicitation questions. A well-known example is the 213-order. See Barbiers (2005) for a detailed overview.

translation tasks they were presented with one single order in Standard Dutch, which they translated into their dialect with the same order, a different one, or—in the case of multiple responses—a combination of orders. Accordingly, some of the *no*'s in table 2 are based on informants explicitly rejecting a particular order, while others are a reflection of the absence of this order in the informants' translation of the sentence that was offered to them. As far as I was able to ascertain, however, this methodological split has had no impact on the results. Moreover, the Standard Dutch sentences that were used as basis for the translation tasks showed a fair amount of word order variation, and informants regularly spontaneously offered more than one possible order. That being said, the difference between elicitation questions and translation tasks is a methodological complication that should be kept in mind when interpreting the results of the analysis.

The second point concerns the *NA*'s in table 2. As was pointed out above, not every question from the oral questionnaire was asked in every dialect location. This means that for particular combinations of cluster order and dialect location, information about whether that order occurs in that location is missing.[10] Given that the Multiple Correspondence Analysis described in subsection 3.4 cannot be applied to a data table containing missing values, I first imputed the missing data using the `imputeMCA`-function of the `R`-package `missMDA` (Husson and Josse (2013)). This function uses the iterative MCA algorithm to impute missing values in a categorical data table (see Husson and Josse (2013) for more details and Josse et al. (2012) for general discussion of imputing missing data), which allowed me to perform the analysis on a complete 31×267 data table. That being said, I also performed the same analysis as described in the following section based on a partial data tabel, leaving out the 82 dialects for which one or more data points was missing, and found no significant differences in the results. Given that the method based on imputing missing values has a greater empirical coverage, that is the one I adopt in the rest of this paper.

3.3 Active and supplementary variables

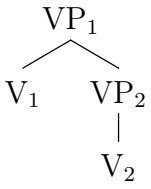
As was pointed out in section 2, the ultimate goal of the analysis presented here is to determine which—if any—combination of grammatical parameters can best account for the observed word order variation in Dutch verb clusters. In order to do so, I need to include in the analysis insights and results from the theoretical literature on verb clusters. In this subsection I describe how these theoretical analyses can be operationalized for the quantitative analysis.

The first step involves decomposing the theoretical accounts into their constitutive parts. Let me illustrate how this works based on the analysis of Barbiers (2005). He starts out from a uniformly head-initial base structure. This means that two cluster orders are base-generated in his account and do not involve any additional syntactic operations: 12 and 123.[11] This is illustrated in the trees in (9) and (10).

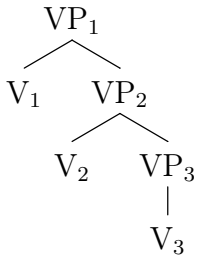
[10] In total, there are 538 cells that contain *NA*. Out of a total of 8277 (=31×267) cells, this represents 6.49%. Missing values occur in 82 of the 267 dialect locations.

[11] Here and throughout this paper I am setting aside clusters containing four or more verbs, as these

(9)

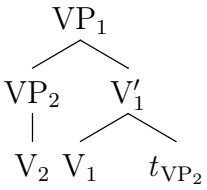


(10)

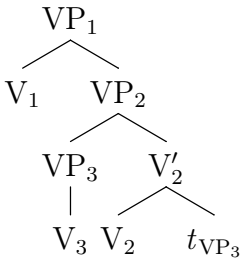


Additional cluster orders are derived via syntactic movement in Barbiers's analysis, more specifically VP-intraposition. Applied to the base structures in (9) and (10), this yields the orders 21, 132, 312, 231, and 321:

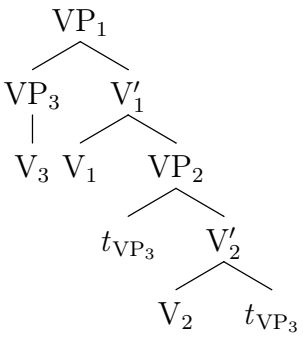
(11)



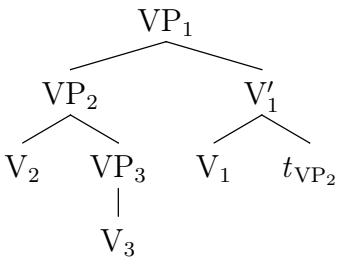
(12)



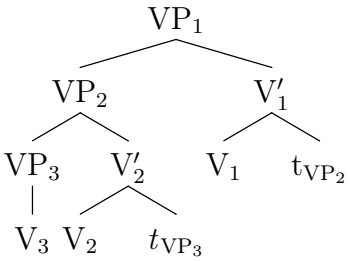
(13)



(14)



(15)



Two comments are in order with respect to these derivations. The first one concerns the 213-order. As pointed out by Barbiers (2005, 248), this order cannot be derived in his system. The 213-order would involve movement of VP_2 to specVP_1 to the exclusion of VP_3, but given that VP_3 is a subpart of VP_2, movement of the latter necessarily entails

data were not part of the SAND-questionnaire.

movement of the former.[12] The second comment pertains to the derivation in (15), i.e. the 321-order. As discussed in Barbiers (2005, 250–253), this order can come about in two ways. Either the movement operation of VP_2 to specVP$_1$ is the result of a feature checking requirement between V_2 and V_1 (comparable to what we see in the 231-order in (14)), or it is the result of a feature checking relation between V_3 and V_1 (like in the 312-order in (13)), whereby VP_3 pied-pipes VP_2. In this latter case, the derivation contains a specific type of pied-piping, whereby the pied-piper sits in the specifier of the pied-piped element. The difference between the 312-order in (13) and the 321-order in (15) then boils down to the absence or presence of this type of pied-piping.

As was already hinted at in the previous paragraph, the various VP-movements in (11)–(15) are feature-driven. The feature system these movements are based on can be summarized as follows:

(16)
a. main verbs: [*i*Event]
b. modals and aspectual auxiliaries: [*u*Event]
c. perfective auxiliaries: [*u*Perfective]
d. perfective participlces: [*i*Perfective]

The first two of these feature specifications encode the intuition that while modals and aspectual auxiliaries modify and hence need to combine with (a structure expressing) an event, they do not themselves express such an event; this role is reserved for main verbs. The last two feature specifications express a similar intuition with respect to perfectivity: a perfective participle can express perfectivity on its own (e.g. as in *the signed letter*), but a perfective auxiliary cannot. Instead, such an auxiliary necessarily has to combine with a perfective participle in order to be interpretable. The way this feature system is set up has important repercussions for the word order in certain verb clusters. Consider for example a cluster consisting of two modal verbs and a main verb. In the feature system outlined in (16), this cluster can be represented as follows:

(17) $[_{VP_1}\ \text{modal}_{[u\text{Event}]}\ [_{VP_2}\ \text{modal}_{[u\text{Event}]}\ [_{VP_3}\ \text{infinitive}_{[i\text{Event}]}\ ]]]$

As indicated by the arrows, there is a feature checking relation between each of the modals and the main verb. However, there is no such relation between the two modals. While both of them bear an [Event]-feature, neither has the interpretable version of this feature and so the two modals cannot enter into a checking relation with one another. This means that there can be no movement of VP_2 into specVP$_1$ and hence, any orders that rely on this movement operation should be excluded on Economy grounds: they contain a movement operation that is not triggered by any feature checking relation. As is clear from the derivations in (12)–(15), the only order that crucially depends on VP_2 moving

[12]Note that this conclusion holds only if one assumes that VP_3 cannot be subextracted out of VP_2 prior to the movement of VP_2 to specVP$_1$. In Barbiers's system this follows from the combination of the following assumptions: (i) VP-intraposition necessarily targets specVP, (ii) VPs have only one specifier, and (iii) bar-levels (e.g. V'_2) cannot move.

into specVP$_1$ is 231.[13] Barbiers's feature-based account thus predicts that the 231-order should be excluded in the case of modal-modal-infinitive clusters. A similar line of reasoning applies to auxiliary-modal-infinitive clusters. Consider their featural representation in (18).

(18) $[_{\text{VP}_1}\ \text{auxiliary}_{[u\text{Perfective}]}\ [_{\text{VP}_2}\ \text{modal}_{[i\text{Perfective},u\text{Event}]}\ [_{\text{VP}_3}\ \text{infinitive}_{[i\text{Event}]}\ ]]]$

Recall that perfective auxiliaries do not necessarily entertain a feature checking relation with the main verb, but rather with the verb they immediately govern (and which typically appears as a perfective participle). In (18), this verb is the modal in VP$_2$, which accordingly carries an interpretable [Perfective]-feature.[14] Thus, there are checking relations between V$_1$ and V$_2$, and between V$_2$ and V$_3$, but not between V$_1$ and V$_3$. Barbiers thus predicts that in the case of auxiliary-modal-infinitive clusters, the 312-order (which crucially depends on movement of VP$_3$ into specVP$_1$) should be unattested.

This concludes my summary of Barbiers (2005)'s analysis of verb cluster ordering in Dutch: it starts out from a uniformly head-initial base order, derives additional orders through VP-intraposition (possibly accompanied by pied-piping via the specifier), and excludes a number of verb cluster combinations through a feature checking account. From this account we can now distill the grammatical parameters that are inherent to it, i.e. those aspects of the analysis with respect to which cluster orders can differ and which are thus possible points of interdialectal variation. They are summed up in (19).

(19)
a. [±base-generation]: can the order be base-generated?
b. [±movement]: can the order be derived via movement?
c. [±pied-piping]: does the derivation of the order involve pied-piping?
d. [±feature-checking violation]: does the order involve a feature checking violation?

I have represented Barbiers's parameters as binary choices. Each of the four parameters in (19) splits verb cluster orders into two mutually exclusive subsets: those with a positive setting for this parameter and those with a negative one. More specifically, the 31 cluster orders under investigation here can be coded in terms of the grammatical parameters in (19). This means that table 2 can be expanded with columns representing not locational, but linguistic information:[15]

In total, I have added 70 linguistic variables to the data table, representing not just the analysis of Barbiers (2005), but also those of Barbiers and Bennis (2010), Abels (2011),

[13] Recall that while the 321-order *can* involve this movement, it need not, as VP$_2$ can also end up to the left of V$_1$ by virtue of being pied-piped by VP$_3$.

[14] Note that the analysis here does not take into account that the modal surfaces as an infinitive rather than a participle due to the IPP-effect.

[15] Based on this table one might get the impression that the parameters [±base-generation] and [±movement] are mirror images of one another, thus making one of them superfluous. Recall, however, that the 213-order cannot be derived in Barbiers's system, neither via base-generation nor via movement. As a result, it has a negative setting for both parameters and thus provides a differentiation between them.

	base-generation	movement	pied-piping	feature-checking violation
AUX1(be.sg)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(be.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.sg)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(have.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.pl)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(have.pl)	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD1(sg)-INF2	yesBase	noMvt	noPiedP	noFeatCheckFail
INF2-MOD1(sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD2-INF3-MOD1(sg)	noBase	yesMvt	noPiedP	yesFeatCheckFail
MOD1(sg)-MOD2-INF3	yesBase	noMvt	noPiedP	noFeatCheckFail
MOD1(sg)-INF3-MOD2	noBase	yesMvt	noPiedP	noFeatCheckFail
...	...	...	...	...

Table 3: Encoding of the SAND-data according to four grammatical parameters taken from Barbiers (2005)

Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004).[16] Moreover, I have included a head-initial head movement analysis, a head-final head movement analysis, a head-initial XP-movement analysis, and a head-final XP-movement analysis (all as described in Wurmbrand (2005)). Finally, I added 17 additional variables based on the theoretical literature, but not tied to a specific analysis. For example, I encoded for every cluster order whether or not it involves IPP.

In the statistical analysis described in the next subsection, these 70 linguistic variables are treated as *supplementary variables*. Unlike active variables (in our case the locational data extracted from the SAND-database), supplementary variables do not contribute to the construction of the principal components. Instead, they serve to interpret or illustrate those components (see Husson et al. (2011, 20–24)). The next subsection provides more details about this.

3.4 The analysis

Multiple Correspondence Analysis is a principal component method that can be applied to tables containing individuals as rows and categorical variables as columns. In the (partial) data tables I have presented so far, the individuals are verb cluster orders and the variables

[16] One aspect of Schmid and Vogel (2004)'s analysis I was not able to implement is the effect of focus/stress on verb cluster ordering, as this feature was neither tested nor transcribed in the SAND-questionnaires.

are the geographical and linguistic characterizations of those cluster orders. The analysis proceeds in three steps: first, the raw data table is transformed into a distance matrix, then the number of dimensions of that distance matrix is reduced, and finally the result of that dimension reduction is matched against the supplementary (linguistic) variables. I now proceed to describe these steps in more detail.[17]

In a first step, the raw data table is converted into a distance matrix. This is a 31×31 table which has the verb cluster orders from the SAND-data both as rows and as columns. Each cluster order is compared pairwise with each other cluster order and a numeric value (between 0 and 1) is assigned to that comparison to indicate how distinct these two cluster orders are from one another.[18] This distance is determined by looking at the active variables in the data table, i.e. the locational data. Concretely, the more two cluster orders either occur or are absent in the same dialect locations, the smaller the distance between them will be. If two cluster orders have the exact same distribution, the distance between them is 0, while if they have a completely complementary distribution, their distance would be 1. What we get, then, is a measure of the degree of similarity or difference between the various cluster orders based on their geographical distribution. It is worth making explicit at this point that the notion of 'geographical distribution' is not dependent on those dialect locations forming a contiguous dialect region. Rather, the geographical data are merely used as binary variables to determine which cluster orders typically go together and which ones do not. This is one of the crucial differences between Reverse Dialectometry as outlined in this paper and Classical Dialectometry, where the correlation between linguistic and geographical distance plays a central role (see above, section 3.1, for discussion and references).

The second step of the analysis involves dimension reduction. As in principal component analysis, the goal of MCA is to reduce a (typically large) set of possibly correlated variables to a smaller group of linearly uncorrelated ones. Put differently, in the distance matrix mentioned above, each cluster order is situated in a 31-dimensional space, and in order to be able to visualize and interpret the verb cluster data, we need to reduce the dimensionality of this space. For example, a two-dimensional representation of the verb cluster data under investigation here is given in figure 1.

[17] As mentioned above, all calculations were carried out in `R` (R Core Team (2014)) using the `FactoMineR`-package (Husson et al. (2014)), in particular the `MCA`-function in that package. For precise details about the algorithms and math underlying that function I refer to Husson et al. (2011).

[18] Given that the distance between a cluster order A and a cluster order B is identical to that between B and A, the distance matrix is symmetrical across the diagonal. Moreover, given that each cluster order is identical to itself, the diagonal of the distance matrix only contains zeroes.

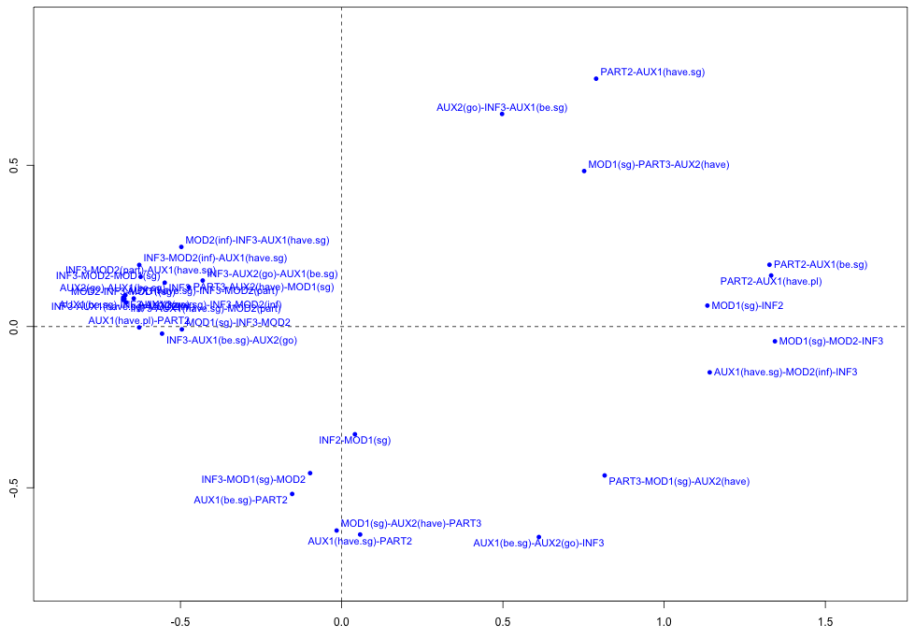


Figure 1: Two-dimensional representation of the SAND verb cluster data

In this graph, each of the 31 cluster orders is situated on a two-dimensional plane. When two cluster orders are close together (e.g. PART2-AUX1(be.sg) and PART2-AUX1(have.pl) on the right-hand side of the graph) this means that they have a highly similar geographical distribution, while when two orders are far apart (like PART3-MOD1(sg)-AUX2(have) in the lower right quadrant and MOD2(inf)-INF3-AUX1(have.sg) in the upper left one), they typically do not co-occur in the same dialect locations. In other words, figure 1 offers a visual representation of the degree of similarity between the 31 cluster orders.

A dimension reduction of this nature always involves a trade-off between explaining as much of the variance as possible that was in the original data set on the one hand, and keeping the number of dimensions as small as possible for easy visualization and interpretation on the other. In order to determine the appropriate cutoff point, we can make use of a so-called scree plot. This two-dimensional graph represents the dimensions on the x-axis and indicates on the y-axis for each of them what percentage of the variance in the original data set is explained by that dimension. The scree plot for our verb cluster data is represented in figure 2. What this graph shows is that the first dimension explains more than 50% of the variance found in the SAND verb cluster data. Dimension 2 adds another 13% and dimension 3 roughly 10%. After that, however, there is a sharp drop: the fourth dimension accounts for less than 5% of the variance. Together, the first three dimensions represent 78.46% of the variance. In other words, roughly 80% of the variation in verb cluster ordering in Dutch can be ascribed to three variables. The goal of the third step of the analysis is to determine what those variables are.

Consider again the graph in figure 1. It shows which verb cluster orders typically cluster

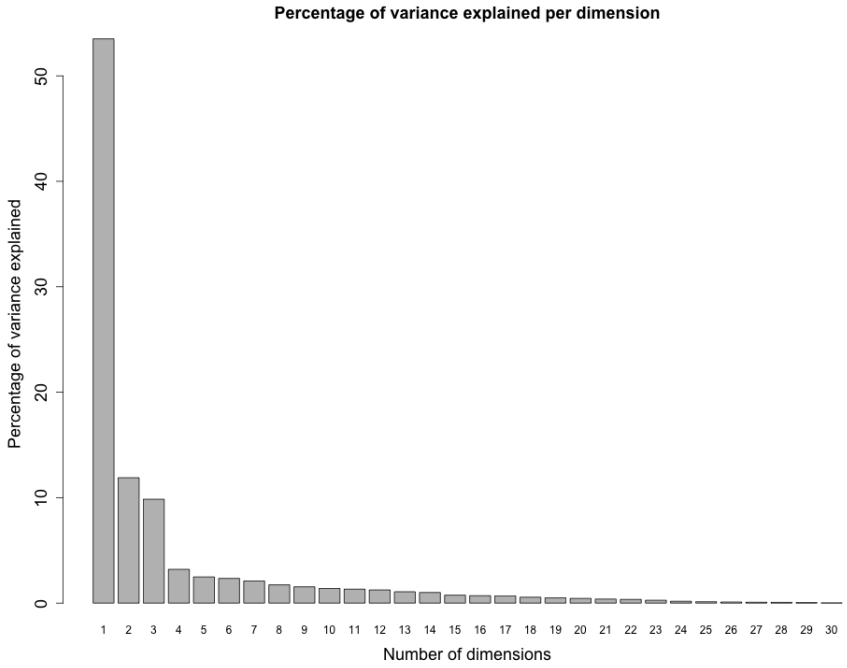


Figure 2: Scree plot for the MCA-analysis of the SAND verb cluster data

together, and which ones do not. If the microvariation we find in dialect Dutch verb cluster ordering is to be reduced to grammatical parameters (see the discussion in section 1), then we expect the pattern in figure 1 to be determined (at least in part) by such parameters. Put differently, cluster orders that are close together in the graph should be the result of the same or a highly similar parameter setting, while orders that are further apart should have fewer parameter settings in common. Grammatical parameters thus create *natural classes* of verb cluster orders. This is where the supplementary (linguistic) variables come in: we use them to interpret the (in our case: three) dimensions that were retained in step two of the analysis. This is achieved by mapping or matching the first three dimensions of the MCA-analysis against those variables. In a nutshell, we are trying to see if or to what extent the data spread in figure 1 aligns with some theoretical property of the clusters in question. There are two basic ways of testing this. The first is to color-code the plot in figure 1 according to (the values of) a linguistic variable and to see if cluster orders that have the same color (i.e. that share some linguistic property) are also close to one another on the graph (i.e. have a similar geographical distribution). The second is to calculate, for each combination of linguistic variable and MCA-dimension, the squared correlation ratio (η^2), which provides a measure for the proportion of variance on that particular dimension that is explained by that linguistic variable. The value of η^2 is between 0 and 1, and the

higher the number, the stronger the link between the dimension and the linguistic variable.

This concludes the methodological section of this paper. I have made explicit how both the raw data from the SAND-project and the theoretical linguistic literature on verb clusters were operationalized for a quantitative, statistical analysis that takes the form of a Multiple Correspondence Analysis. This method allows us to reduce the multi-dimensional variational space to a three-dimensional one. In the next section I examine for each of those three dimensions to what extent they align with the supplementary, linguistic variables, and in section 5 I interpret those results from a formal linguistic point of view.

4 Results

4.1 Introduction

In this section I present the results of the analysis outlined in the previous section. More specifically, for each of the three dimensions retained in the analysis I indicate which of the linguistic variables correlates most strongly with that dimension. The section will be highly descriptive in nature; for a further refinement and a linguistic interpretation of the results presented here, I refer the reader to section 5.

Before proceeding with the discussion of the results, I need to make one preliminary remark. As pointed out by Richardson (2011), the value of η^2 for the combination of a dimension and a particular categorical variable is sensitive to the number of values that variable can have: the higher the number of possible values, the higher the value of η^2.[19] This means that when evaluating the results, we should be wary of variables that have a high η^2-value merely (or mostly) because they have many different values. Accordingly, in what follows I will mainly concentrate on two- or three-valued variables when discussing the results of the MCA-analysis.

4.2 Dimension 1

Table 4 below lists which of the supplementary (linguistic) variables in the MCA-analysis described above had the highest squared correlation ratio for the first dimension. In light of the preliminary remark made above, I also list for each variable how many values it has.

The variable LightHeavyOrdering is inspired by Abels (2011) and Bobaljik (2004), who suggest that cluster ordering might be sensitive to the 'morphological size' of the verb forms involved in the cluster, the idea being that participles are "smaller" than infinitives (see Abels (2011, 24)). In order to test the effect of the morphological shape on cluster ordering, I encoded the 31 cluster orders in terms of their morphological make-up. For example, the cluster *is gestorven* ('has died', lit. is died) was encoded as FinPart (a finite verb followed by a participle), and *zwemmen kunnen moet* ('must be able to swim', lit. swim can must) as InfInfFin. Given that this method of encoding yielded 13 different

[19] One way of clearly demonstrating this is by introducing a fake variable into the data set, which assigns a different value to each of the 31 cluster orders. Such a variable has a 'perfect' η^2-value of 1.

variable	η^2	**number of categories**
LightHeavyOrdering	0.621	13
ClusterOrder	0.517	8
BarBenNominalInfinitive	0.425	3
BaderVMod	0.398	3

Table 4: The highest η^2-values for dimension 1

values,[20] however, its high η^2-value is arguably an artefact, which seems corroborated by the fact that this same variable also has a high η^2-value for the second and third dimension (see below). Accordingly, I will set this variable aside in the remainder of the discussion.

A similar fate befalls the variable ClusterOrder, which is the second-highest ranked one in table 4. It simply encodes all clusters in terms of their basic cluster order (i.e. 12, 123, 321, etc.), regardless of the verbs making up the cluster. As a result, it has 8 possible values (two orders for two-verb clusters plus six orders for three-verb clusters), which arguably explains its high position in this ranking. This variable will also make a reappearance in the η^2-ranking of the other dimensions (see below), suggesting that it is the number of values that accounts for its high ranking, not its intrinsic degree of correlation with the variance along dimension 1.

Things get more interesting, though, when we consider the next two variables in table 4. The variable BarBenNominalInfinitive is taken from Barbiers and Bennis (2010). They propose that whenever an infinitive is spelled out to the left of its selecting verb, that infinitive is actually nominalized (and hence technically not part of the cluster). I have encoded this in my data table as a three-valued variable: it is set to 'yes' when an infinitive precedes its selecting verb, to 'no' when it follows its selecting verb, and to 'dna' when the cluster contains no infinitive. As indicated in table 4, this variable has a squared correlation ratio of 0.425.[21] The color coded plot in figure 3 provides a visual representation of this.

[20] FinInf, FinInfInf, FinInfPart, FinPart, FinPartInf, InfFin, InfFinInf, InfFinPart, InfInfFin, InfPartFin, PartFin, PartFinInf, and PartInfFin.

[21] It is hard to find absolute measures for η^2 to determine the size of the effect. Some authors cite Cohen (1962), in which case an η^2-value of 0.0099, 0.0588, and 0.1379 would correspond to a small, medium, and large effect respectively, but see Richardson (2011) for critical discussion.

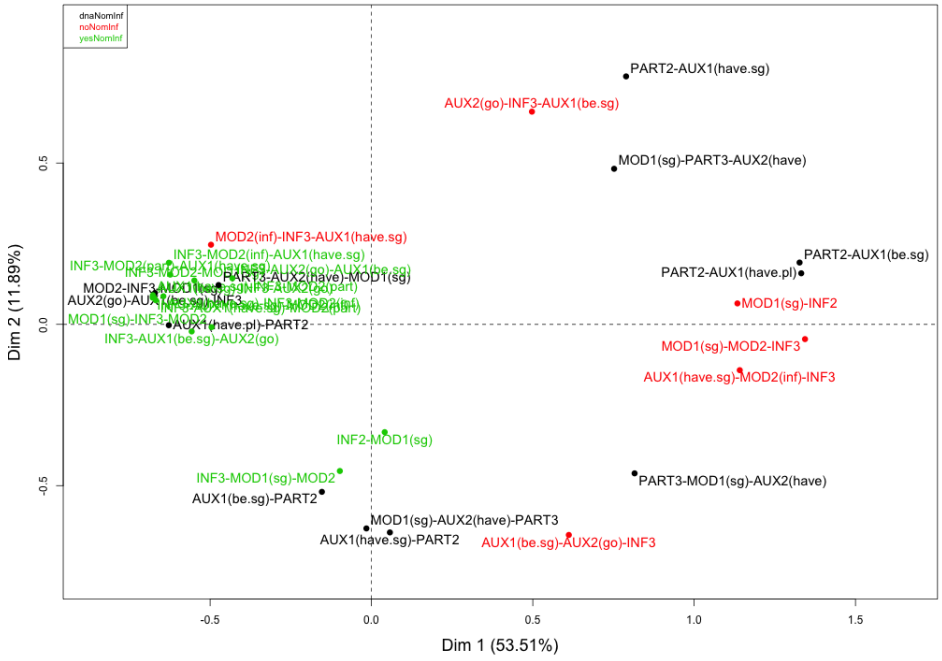


Figure 3: Verb cluster plot color coded according to the variable BarBenNominalizedInfinitive

The thing to focus on in this graph is the extent to which the distribution of points along the x-axis (i.e. dimension 1) correlates with the color coding, in particular the contrast between green (a positive setting for the variable) and red (a negative setting). The black points are cluster orders where the variable does not apply and hence these points are not informative (though see the next section for further discussion). As is clear from the graph, the red-green divide corresponds to a very large degree with positive vs. negative values on the x-axis. The only clear exception is the cluster MOD2(inf)-INF3-AUX1(have.sg), i.e. a modal followed by a main verb followed by a singular form of the verb *have*. An example is given in (20).

(20) *Vertel maar niet wie zij kunnen roepen had.*
tell PRT not who she can call had
'Don't tell me who she should have called.'

The final variable from table 4 is taken from Bader (2012). It refers to his 'Mod-V Constraint', which states that "The complement of a modal verb precedes the modal verb."[22] As should be clear, this variable will yield largely the same data distribution as Barbiers and Bennis (2010)'s nominalized infinitive, the only difference being infinitives selected by verbs other than modals, which fall outside the purview of Bader's constraint, but not that of Barbiers & Bennis.

[22]Bader's account is OT-based, hence the categorical nature of this statement.

This concludes my discussion of dimension 1. I have outlined which of the linguistic variables have the highest η^2-values with respect to this dimension. In section 5, I will further refine these results and try to find a linguistic interpretation for them, but before doing so, I first turn to the other two dimensions.

4.3 Dimension 2

The linguistic variables with the highest squared correlation ratio for the second dimension are listed in table 5 below.

variable	η^2	**number of categories**
LightHeavyOrdering	0.575	13
ClusterOrder	0.464	8
Slope	0.422	4
SchmiVoMAPhc	0.379	3
SchmiVoMAPlrVfunc	0.348	4
SchmiVoMAPlrV	0.334	4
BarbiersBaseGeneration	0.309	2

Table 5: The highest η^2-values for dimension 2

The first two variables on this list have already been discussed in the previous subsection and will be set aside for the reasons outlined there. The third is one of the additional variables that was added based on the linguistic literature but not linked to a specific analysis. It concerns the question of whether the cluster is ascending or descending, a factor which is known to play a role in, for example, cluster penetrability (see Salzmann (2013) for recent discussion). Given that three-verb clusters are not necessarily uniformly ascending or descending, the variable has four possible values: ascending, descending, ascending-descending (e.g. 132), and descending-ascending (e.g. 312). This variable has a squared correlation ratio of 0.422, and its effect can be graphically represented as in figure 4.

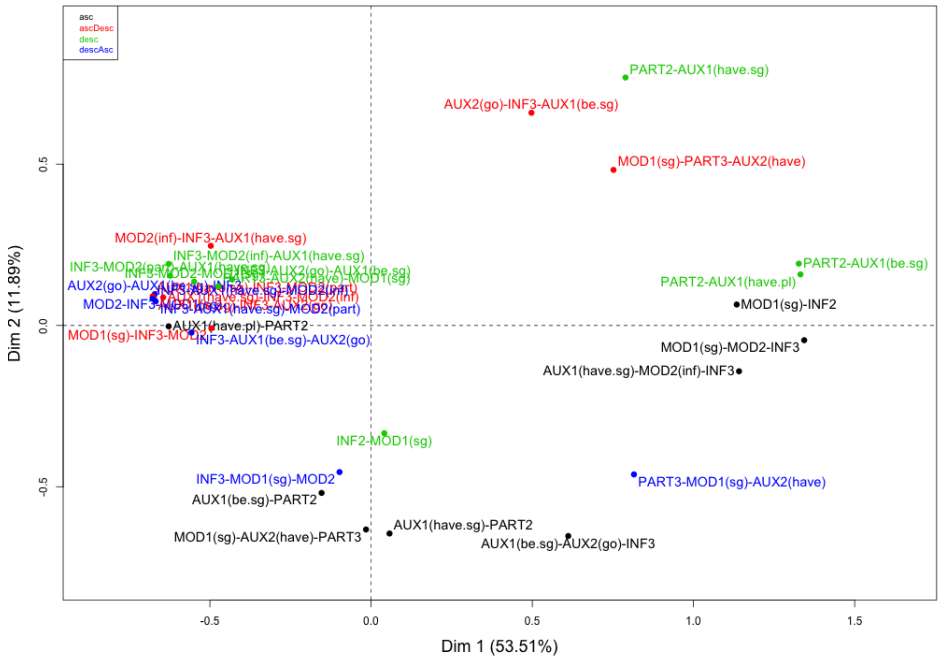


Figure 4: Verb cluster plot color coded according to the variable Slope

This time we are focusing on the distribution of the points along the y-axis, and the extent to which that distribution matches the different colors. Although the picture is perhaps less clear than the one in figure 3, there does seem to be a tendency for blue and black cluster orders (i.e. ascending and descending-ascending) to be below the x-axis, and for red and green orders (i.e. descending and ascending-descending) to be above the x-axis.

The next three variables are taken from Schmid and Vogel (2004). Like Bader's, their account is OT-based, and these three variables correspond to three constraints in their analysis. The constraints in question are given in (21)-(23).

(21) **MAPhc**
If A and B are sister nodes at LF, and A is a head and B is a complement, then the correspondent of A precedes the one of B at PF.

(22) **MAPlrVfunc**
If A is a functional verb (or a verb containing functional features) that asymmetrically c-commands at LF another verb B that belongs to the same extended projection, then the correspondent of A precedes that of B at PF.

(23) **MAPlrV**
The heads of an extended projection of V are linearized in a left-to-right fashion, i.e., if head A asymmetrically c-commands head B at LF, then the PF correspondent of A precedes the one of B at PF.

As is clear from these definitions, all three these constraints set head-initial orders apart from non-head-initial ones. I have encoded the SAND cluster orders in terms of the number of violations they incur in Schmid & Vogel's OT-analysis. This leads to a three-valued variable in the case of SchmiVoMAPhc (zero, one, or two violations), and four-valued ones in the case of SchmiVoMAPlrVfunc and SchmiVoMAPlrV (because some orders violate this constraint three times, see the original paper for details). A visual representation of how MAPhc aligns with the second dimension is given in figure 5. Note how the green orders (i.e. those that violate the constraint once) tend to have a positive value on the y-axis, while the red ones (i.e. those that do not violate the constraint) tend to have a negative one. The status of the black points (= orders that violate the constraint twice) is somewhat unclear.

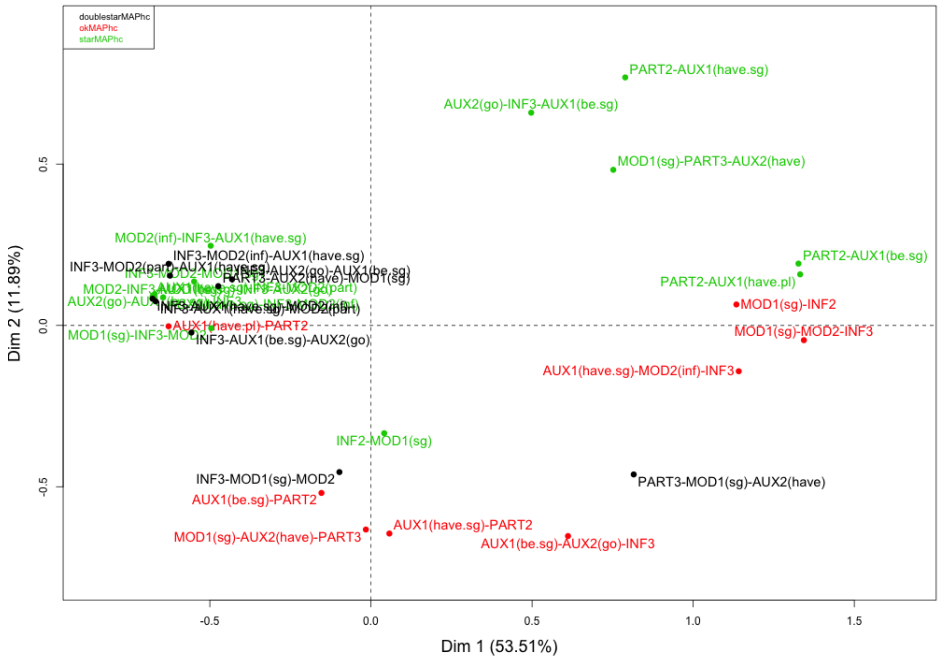


Figure 5: Verb cluster plot color coded according to the variable SchmiVoMAPhc

The final variable in table 5 is taken from Barbiers (2005) (see also table 3 above). It sets apart the orders that can be base-generated in Barbiers's account from those that cannot. Given that his analysis starts out from a head-initial base structure, it should come as no surprise that this variable yields comparable results to the constraints in (21)-(23).

4.4 Dimension 3

The third and final dimension under discussion here yields the η^2-values listed in table 6.

The first two variables I set aside like in the previous two subsections. The variable Slope was also already introduced above, but whereas there it mainly competed with other three-

variable	η^2	**number of categories**
ClusterOrder	0.777	8
LightHeavyOrdering	0.710	13
Slope	0.707	4
SchmiVoMAPch	0.701	3
HaegRiemsBaseOrder	0686	2
BaderBaseOrder	0686	2
HFinHmvtBaseOrder	0686	2
HFinXPmvtBaseOrder	0686	2

Table 6: The highest η^2-values for dimension 3

or four-valued variables, in this dimension there are a number of two-valued variables with very high η^2-values, so it speaks to reason to focus primarily on those (though see below for why Slope ranks high with respect to this dimension). The constraint SchmiVoMAPch is once again taken from Schmid and Vogel (2004) and it is essentially the mirror image of SchmiVoMAPhc (cf. (21)):

(24) **MAPch**
If A and B are sister nodes at LF, and A is a head and B is a complement, then the correspondent of B precedes the one of A at PF.

In other words, this constraints enforces a strictly head-final order. Just like SchmiVoMAPhc, it is three-valued in that it divides verb cluster orders into those that violate the constraint zero, one, or two times. More interestingly, however, is that for dimension 3 there are two-valued variables—i.e. variables where there is no risk of artifically inflating η^2 through the number of values—that have a very high value for η^2. The variables HaegRiemsBaseOrder, BaderBaseOrder, HFinHmvtBaseOrder, HFinXPmvtBaseOrder are taken from Haegeman and Riemsdijk (1986), Bader (2012), a head-final head movement analysis, and a head-final XP-movement analysis (cf. Wurmbrand (2005)) respectively. What these accounts share is that they start out from a head-final base order, i.e. all four these variables set apart strictly descending orders (= orders that can be base-generated under a head-final base hypothesis) from all others.[23] The effect is quite pronounced, as can be seen not just on the basis of the high η^2-value, but also through the color coded plot in figure 6.

[23]Note that this is probably why the variable Slope shows up in this ranking as well.

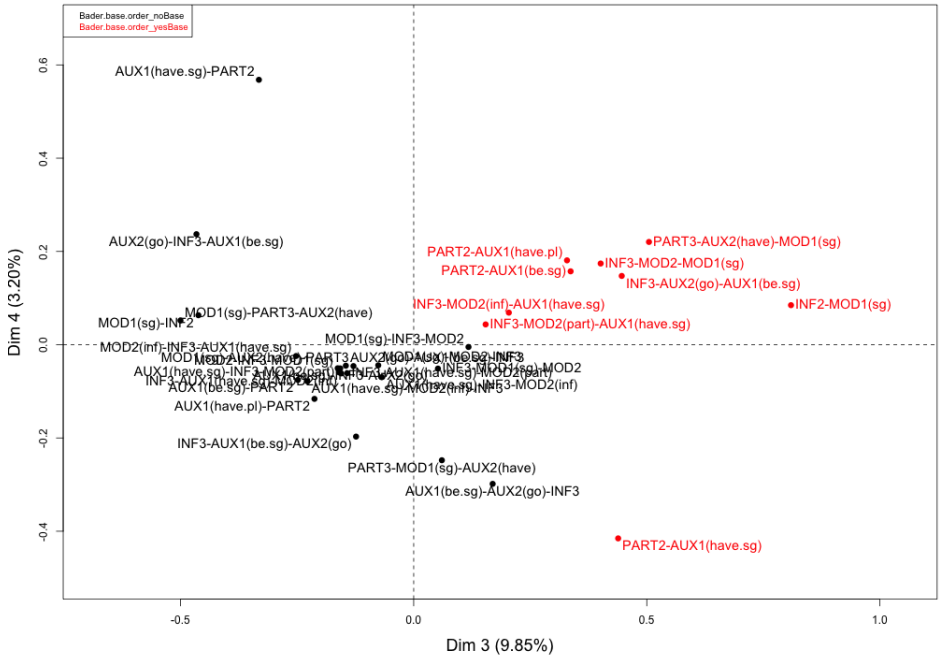


Figure 6: Verb cluster plot color coded according to the variable BaderBaseOrder

Unlike the plots in figures 1-5, this one does not map the first against the second dimension, but the third (x-axis) against the fourth (y-axis).[24] As is clear from the color coding, the third dimension very strongly correlates with head-finality: the red orders (the strictly head-final ones) are clearly separated from the black orders (the non-head-final ones).

4.5 Conclusion

This concludes my presentation of the results of the MCA-analysis. For each of the three dimensions I have listed which of the linguistic variables correlates most strongly with that dimension. In the next section I attempt to provide a linguistic interpretation of those results.

5 Interpretation: dimensions as parameters

In this section I move beyond the description of the results from the previous section and connect the statistical analysis to the discussion in section 2, i.e. the question to what extent variation in dialect Dutch verb cluster ordering can be reduced to the interaction

[24]The choice of the fourth dimension is arbitrary: we could just as easily have plotted the third against the first or second dimension.

between various (micro)parameters. The central hypothesis I want to put forward is that the three dimensions uncovered by the quantitative analysis correspond to three parameters. In working out this idea, I will in some cases need to further refine and elaborate upon the results from the previous section. On the basis of these three parameters it will be possible to sketch the outlines of a parametric account of verb cluster ordering in the dialects of Dutch. The goal of this exercise is not to uncover 'the ultimate analysis' of verb clusters, but rather to show in what way the results of a purely quantitative approach of a particular linguistic phenomenon can inform and guide theoretical accounts of that phenomenon.

Let us start by returning to the third dimension of the MCA-analysis (see subsection 4.4 above). Recall that it set apart strictly head-final orders from all others. A possible linguistic interpretation of this is that we are seeing the effect of the underlying base order. More specifically, suppose we reinterpret dimension 3 as the paramater in (25).

(25) **Parameter #1: Movement**
A dialect {does/does not} diverge from its underlying head-final base order.

Connecting this back to the plot in figure 6, this means that the red (strictly head-final) orders have a negative setting for this parameter, while the black (non-strictly head-final) ones have a positive one. With this in mind, let us reconsider the second dimension. Recall from table 5 and figure 4 that it seemed related to the 'slope' of the cluster, i.e. it set apart ascending and descending-ascending orders on the one hand from descending and ascending-descending ones on the other. As a first step towards a deeper understanding of these contrasts I convert the four-valued parameter Slope into a two-valued one. It can be defined as in (26) and divides up the cluster orders as in table 7.

(26) **FinalDescent:**
Set to 'yes' if the cluster ends in a descending order and set to 'no' if the cluster ends in an ascending order.

FinalDescent_yes	**FinalDescent_no**
21	12
321	123
231	312
132	213

Table 7: Classification of verb cluster orders according to the variable FinalDescent

This new variable has a η^2-value of 0.382 and its degree of alignment with the second dimension can be represented as in figure 7.

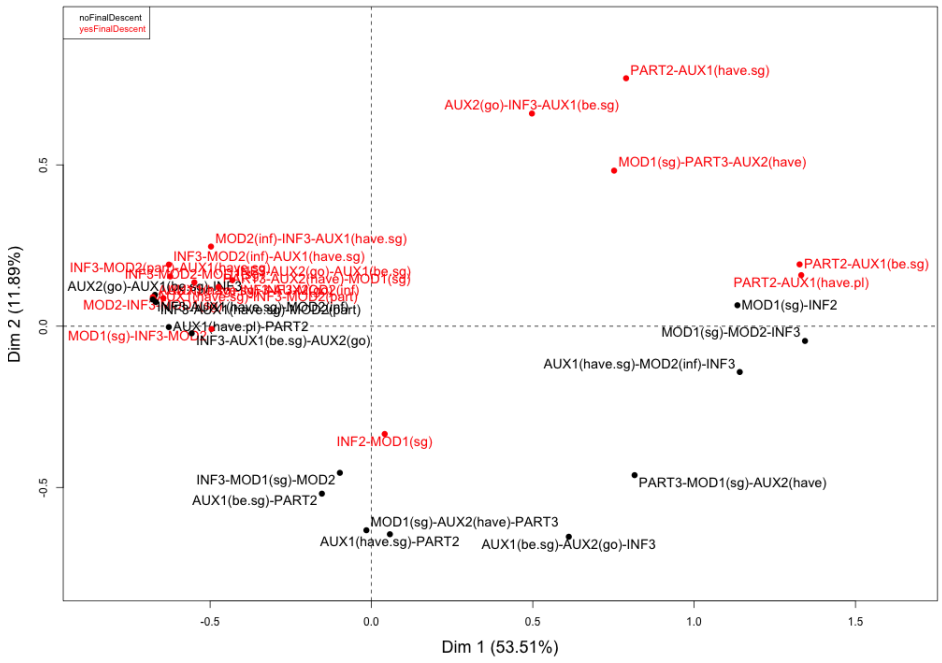
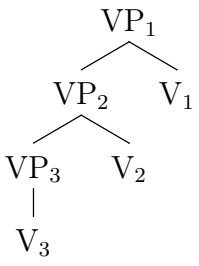


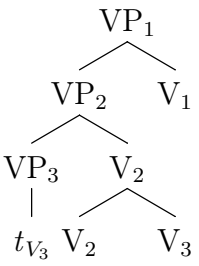
Figure 7: Verb cluster plot color coded according to the variable FinalDescent

As it stands, however, the variable FinalDescent offers little insight into the possible linguistic analysis of verb cluster ordering. In particular, while the slope of a cluster frequently features in the theoretical literature, it is usually presented as an epiphenomenon—e.g. the result of a particular movement operation or base-generated structure—rather than as an explanatory factor in and of itself. In light of our reanalysis of dimension 3 as (divergence from) a base-generated head-final order, though, a possible linguistic interpretation of FinalDescent emerges. Abstracting away from the two-verb clusters for the moment, the cluster orders in the left-hand column of table 7 are either head-final or one movement step removed from head-final. This is illustrated in (27)-(29).

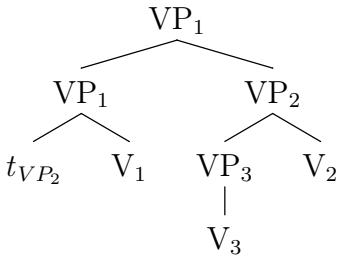
(27)



(28)

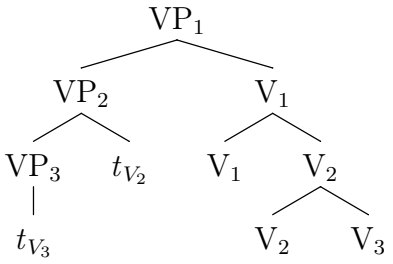


(29)

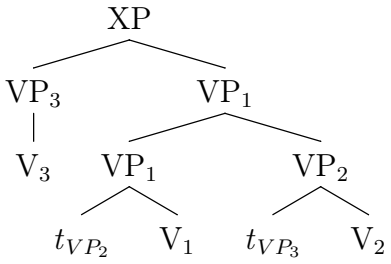


The 231-order requires head movement of V_3 to V_2 (cf. (28)), while the 132-order involves VP_2-extraposition to the right of V_1, as shown in (29). The three-verb clusters in the right-hand column of table 7 on the other hand all require two movement operations if we start out from a head-final base structure. This is shown in the following derivations:

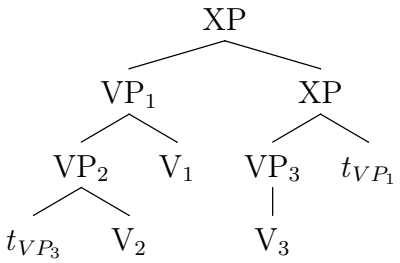
(30)



(31)



(32)



The 123-order involves two head movement steps (first V_3 moves to V_2 and then the complex V_2-head moves to V_1, cf. (30)), the 312-order combines extraposition of VP_2 with leftward movement (to a position agnostically labeled XP in (31)) of VP_3, and the 213-order—to the extent that one wants to derive it at all, cf. Wurmbrand (2005)—requires leftward movement of both VP_3 and the remnant VP_1, as shown in (32). With this much as backgroud, we can reinterpret the second dimension as a grammatical parameter that is sensitive to the number of movement operations involved in deriving a particular cluster order. In the spirit of Chomsky (1995) this can be formulated as an Economy requirement:

(33) **Parameter #2: Economy of Movement**
A dialect {does/does not} allow more than one movement operation in the derivation of its verb clusters.[25]

[25] As pointed out above, this does not apply to two-verb clusters, where it is always possible to get to the other order in a single movement step. At present, I have nothing insightful to say about how to integrate these cluster types into the second parameter.

This leaves the first dimension. Recall that it set apart dialects that consistently place participles and infinitives on opposite sides of their selecting verb (participles to the left, infinitives to the right) from dialects that don't. Arguably this too is the overt manifestation of a more abstract linguistic property. For instance, whether or not a participle precedes its auxiliary could be due to it being adjectival or verbal in nature (see Barbiers and Bennis (2010) and much literature preceding it), and similarly the position of infinitives could be a sign of their being nominalized (in which case they would occupy the regular preverbal object position). In order to fully explore these hypotheses, however, I would need more data than I currently have at my disposal (e.g. the use of certain adverbial modifiers in the case of adjectival participles, or the possiblity of scrambling across an intervening adverb in the case of a nominalized infinitive), and doing so would take me far beyond the scope of the current paper. For this reason, I will simply frame the first dimension as a specific instantiation of a classical head parameter:

(34) **Parameter #3: Verbal Head Parameter**
A dialect {does/does not} linearize participles to the left and infinitives to the right of their selecting verbs.

What emerges from the discussion, then, is a rough, parametrized account of verb cluster ordering in Dutch dialects: we have identified the major points at which dialects can differ and have integrated those points into a coherent theoretical account. As pointed out above, the goal of this exercise has not been to argue that the analysis of verb clusters laid out here is superior to existing accounts. Rather, it has served as a proof of concept, to illustrate that the methodology developed in this paper can inform and guide our theoretical thinking about this phenomenon.

6 Conclusions and prospects

This paper is situated at the intersection of quantitative and qualitative linguistics. It uses quantitative-statistical methods to further our theoretical understanding of variation in verb cluster ordering in Dutch dialects. In so doing, it harnasses and combines the strenghts of both approaches: quantitative linguistics has sophisticated means of dealing with large and highly varied data sets, while hypotheses and analyses from qualitative linguistics can be used to guide and narrow down the interpretation of the statistical results. For the case at hand—verb clusters—I have shown how the 137 dialect types that were manifested in the raw data can be largely whittled down to the interaction between three grammatical parameters. The method thus allows for a way to separate the signal (i.e. that part of the variation that is due to grammar proper) from the noise (all extragrammatical factors, ranging from sociolinguistic variation all the way to simple speech errors).

The research presented here can be extended in various ways. For instance, the data set can be expanded to include not just (verb cluster variation in) varieties of Dutch, but also other Germanic languages and language varieties. Similarly, the MCA-analysis

used in this paper can be supplemented by other statistical techniques such as association rule mining (see e.g. Spruit (2008)) or various clustering techniques (e.g. Heeringa and Nerbonne (2013)). More generally, though, I hope this paper has shown the viability and mutual benefits of an increased collaboration between formal-theoretical and quantitative-statistical approaches to language variation.

References

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

Bader, Markus. 2012. Verb-cluster variations: a harmonic grammar analysis. Handout of a talk presented at "New ways of analyzing syntactic variation", November 2012.

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, 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.

Bobaljik, Jonathan. 2004. Clustering theories. In *Verb clusters. A study of Hungarian, German and Dutch*, ed. Katalin É. Kiss and Henk van Riemsdijk, 121–145. Amsterdam: John Benjamins.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Massachusetts: MIT Press.

Cinque, Guglielmo. 2005. Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry* 36:315–332.

Cohen, J. 1962. The statistical power of abnormal–social psychological research. *Journal of Abnormal and Social Psychology* 65:145–153.

Cornips, Leonie, and Willy Jongenburger. 2001. Het design en de methodologie van het sand-project. *Nederlandse Taalkunde* 3:215–232.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of language*, ed. Joseph Greenberg. Cambridge, Massachusetts: MIT Press.

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.

Husson, Francois, and Julie Josse. 2013. *missmda: Handling missing values with/in multivariate data analysis (principal component methods)*. URL `http://CRAN.R-project.org/package=missMDA`, r package version 1.7.2.

Husson, Francois, Julie Josse, Sebastien Le, and Jeremy Mazet. 2014. *Factominer: Multivariate exploratory data analysis and data mining with r*. URL `http://CRAN.R-project.org/package=FactoMineR`, r package version 1.26.

Husson, François, Sébastien Lê, and Jérôme Pagès. 2011. *Exploratory multivariate analysis by example using R*. Boca Raton/London/New York: CRC Press.

Josse, Julie, Marie Chavent, Benot Liquet, and François Husson. 2012. Handling missing values with regularized iterative Multiple Correspondence Analysis. *Journal of Classification* 29:91–116.

Kayne, Richard. 2000. *Parameters and universals*. Oxford University Press.

Nerbonne, John. 2010. Mapping aggregate variation. In *Language and space. international handbook of linguistic variation. vol. 2, language mapping*, ed. A. Lameli, R. Kehrein, and S. Rabanus, 476–495. Berlin: Mouton de Gruyter.

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.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Richardson, John T.E. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review* 6:135–147.

Salzmann, Martin. 2013. New arguments for verb cluster formation at PF and a right-branching VP. Evidence from verb doubling and cluster penetrability. *Linguistic Variation* 13:81–132.

Schmid, Tanja, and Ralf Vogel. 2004. Dialectal variation in German 3-Verb clusters. *The Journal of Comparative Germanic Linguistics* 7:235–274.

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.

Szmrecsanyi, Benedikt. 2012. *Grammatical variation in british english dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

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.