Applying automatically parsed corpora to the study of language variation

Jelke Bloem **Arjen Versloot** **Fred Weerman**

Amsterdam Center for Language and Communication
University of Amsterdam
1012 VB Amsterdam, Netherlands
{j.bloem, a.p.versloot, f.p.weerman}@uva.nl

Abstract

In this work, we discuss the benefits of using automatically parsed corpora to study language variation. The study of language variation is an area of linguistics in which quantitative methods have been particularly successful. We argue that the large datasets that can be obtained using automatic annotation can help drive further research in this direction, providing sufficient data for the increasingly complex models used to describe variation. We demonstrate this by replicating and extending a previous quantitative variation study that used manually and semi-automatically annotated data.

We show that while the study cannot be replicated completely due to limitations of the existing automatic annotation, we can draw at least the same conclusions as the original study. In addition, we demonstrate the flexibility of this method by extending the findings to related linguistic constructions and to another domain of text, using additional data.

1 Introduction

There are many examples of linguistic variation that are not easily explained in terms of rules. One may find two grammatically correct constructions that can be used to express similar meanings, yet speakers still use both of them. A well-known example in English is the dative alternation, where a transitive verb such as *give* can be phrased as a double object construction (1) or as a prepositional dative (2):

(1) He gave his friend the ticket.

(2) He gave the ticket to his friend.

Studies of such phenomena tend to find that there are multiple variables that may influence whether a speaker chooses one or the other construction. This has prompted various multivariate studies by quantitative linguists to analyze instances of such variation in language corpora, starting with Gries (2001), or Bresnan et al. (2007) for a study on the dative alternation. The multivariate statistical models that these studies employ can quantify the contribution of each variable to the variation in probabilistic terms, rather than examining them in isolation.

We show the benefits of using automatically annotated corpora for the study of language variation by replicating a previous, manual multivariate study on Dutch verbal cluster variation (De Sutter, 2009), and extending it to fit more types of clusters. We also show that the same variables are explanatory in at least two different text domains, Wikipedia text and European Parliament proceedings. The larger scale of our investigation allows us to generalize the claims of the previous study to Dutch two-verb cluster variation in general. This topic makes for a good methodological case study for the use of automatically annotated corpora, as verbal clusters are a widely studied phenomena in Dutch syntactic literature (Evers, 1975; Den Besten and Edmondson, 1983; Haegeman and van Riemsdijk, 1986; Zwart, 1996; Wurmbrand, 2004), and the optionality in verbal cluster order has received particular attention in some recent dissertations (De Sutter, 2005; Coussé, 2008; Arfs, 2007). In addition, a methodologically sound quantitative study, which did not make use of automatically annotated data, already exists to compare to (De Sutter, 2009).

This particular case of variation allows for a lot of optionality. The verbal clusters found at the ends of Dutch clauses allow for almost free order variation when there are two verbs. For example, in two-verb clusters the auxiliary verb can be positioned before or after the main verb:

(3) Ik denk dat ik het **begrepen heb**
I think that I it understood have
'I think that I have understood it.'

(4) Ik denk dat ik het **heb begrepen**
I think that I it have understood

As in the dative alternation example, the variation in these two-verb clusters seems to be influenced by multiple variables, beyond the constraints of grammatical rules. Therefore, we consider the multivariate model by De Sutter (2009) to be the most accurate model of Dutch verbal clusters developed so far. Unfortunately, it is also too limited and does not cover all of the constructions that are generally considered to be two-verb clusters. The author claims to have done this for reasons of methodological rigor. However, in a multivariate model the contribution of each variable can be studied independently. If an additional verbal cluster construction is added and it is marked with a variable as being a different construction, it should not make much of a difference for the other variables, assuming that the same set of variables is involved for all verbal cluster constructions. We have included these additional cluster types to create a model with a larger scope, and show that the effects of de Sutter's smaller model are still present. We also compare to a smaller model of our own, created from the large corpus but without the additional cluster types, to verify that our results aren't just an effect of including the additional constructions.

Another reason for excluding the other types of verbal clusters might have been the annotation effort involved in finding corpus examples of them. We avoid this issue by using an automatically annotated corpus, and can extract large samples of various types of constructions simply by defining what counts as a cluster in the syntactic annotation.

In section 2 we briefly discuss Dutch verbal clusters and the variation found in them. We then discuss previous work on modeling of Dutch verbal clusters, including the model by De Sutter (2009) in section 3. Section 4 describes the automatically annotated corpus that we used. In section 5 we discuss the model we created from this data and compare it to that of De Sutter (2009). We also compare models created from two different text types. Section 6 discusses the implications of these results.

2 Verbal cluster variation

In this section, we will briefly summarize how verbal clusters are formed, and discuss the extent of the variation they exhibit. To refer to the two verbal cluster orders, we will follow terminology introduced by Stroop (1970), where construction (3) is called the 2-1 order and construction (4) is 1-2. This is because the finite auxiliary is considered to be the verb that is highest in the syntactic tree, while the main verb is the lowest. This fact lets us number the verbs.

In generative literature, the formation of these clusters is described as a verb movement process known as verb raising, where the main verb is moved upwards in the syntactic tree from its phrase to be joined with the auxiliary verb. This explains the common observation that verbal clusters cannot be interrupted (Evers, 1975), though there are some instances of cluster interruptions, particularly in Flemish Dutch (Evers, 2003). A broad overview of verb raising across Germanic languages is provided in Wurmbrand (2006).

There are various types of two-verb clusters that exhibit order variation:

Auxiliary cluster Examples (3) and (4) show two-verb clusters with auxiliary heads. Following De Sutter (2009) we consider this to be any cluster that is headed by the auxiliaries *hebben* "to have", *zijn* "to be" and *worden* "to be".

Modal cluster A modal verb (*willen* "want", *kunnen* "can") may also be used as a cluster head. Modal clusters are generally treated as a different construction in the literature, as different grammatical rules may apply to it, particularly in other Germanic languages. In Dutch we can observe that the 2-1 order is far more common in this construction, and some authors even say there is no optionality here (i.e. Zuckerman (2001)). We observe that in the Wikipedia part of the Lassy Large corpus, modal clusters occur in the 1-2 order only 0.5% of the time. However, they are considered to be grammatical.

Clusters with other verbs There are verbs such as *staan* 'to stand' and *helpen* 'to help' that can also be raised to form a verbal cluster in certain contexts. This list includes causal verbs, aspectual verbs, and some harder to classify ones, and seems too diverse to be grouped together. For brevity and due to their relative rarity (together they form about 5.5% of the clusters) we did not include these constructions in our study, though it would certainly be possible to explore them in a large corpus study and perhaps contribute to their classification.

Te-infinitival clusters In the cluster types we discussed so far the auxiliary verb was finite, but there are infinitival clusters as well, where both the auxiliary and the main verb are infinite, and the main verb is marked by the infinitival marker *te*. These clusters are uncommon (2.2% of our dataset) and have not been the focus of any study on variation, though since these clusters form a clear group, we have included them. They occur with both auxiliary and modal heads.

Main clause cluster Verbal clusters can occur in main clauses as well, though in a different form. Stroop (2009) states that three-verb clusters in main clauses are comparable to two-verb clusters in subordinate clauses. In that study, he discusses various cluster types in a corpus of spoken Dutch and observes the distributions between 1-2 and 2-1 orders. While there are three verbs in these main clause clusters, only the last two verbs have free order variation, due to the V2-effect present in Germanic languages:

(5) De fuut **kan** in alle wateren van enig formaat **aangetroffen worden** .
The grebe can in all waters of some size found be .
'The grebe can be found in all bodies of water of substantial size.'

(6) Wegwerpbatterijen **kunnen** niet **worden opgeladen** .
Disposable batteries can not be charged .
'Disposable batteries cannot be charged.'

The finite verb must always be in verb-second position.

Stroop furthermore observes, when looking at larger clusters, that variations of three-verb subordinate clause clusters are distributed similarly to variations of four-verb main clause clusters (that have three verbs with varying order). This observation holds for both Dutch and Flemish data, even though the frequencies of orders are different between the languages. Factors that influence order variation seem to be able to affect main clause and subordinate clause clusters in the same way, as Stroop demonstrated for the regional factor. We are therefore convinced that main clause clusters should be included in studies on verbal cluster variation.

The rules and mechanisms discussed in generative literature allow for a lot of optionality, as discussed above, and thus mainly outline the constructions in which order variation can occur. These accounts generally left open the question of the variation found in the surface order. It appears that syntacticians did not concern themselves with explaining it and considered it to be an effect of a non-syntactic process. This is evident in the analysis adopted by Haegeman and van Riemsdijk (1986).

The issue of explaining the factors that influence the choice between two variants was later picked up by other researchers who were interested in non-syntactic effects as well. Coussé et al. (2008) provide a summary of recent work on verbal cluster variation, in particular, three dissertations on the topic (De Sutter, 2005; Coussé, 2008; Arfs, 2007). A diverse set of variables that may influence the use of 1-2 and 2-1 cluster orders has been found, and they group them into four categories: contextual factors (region and mode of communication), rhythmic factors (adherence to the standard stress pattern of Dutch), semantic factors, and discourse factors (mainly the syntactic priming effect).

From this, Coussé et al. (2008) conclude that the choice of verbal cluster order is influenced by a complex set of interacting factors. Therefore, any model representing this phenomenon would need to take many factors into account. The multivariate modeling technique used by De Sutter (2005) seems to fit this criterion.

We will now discuss some related studies where multivariate modeling has been applied to the study of language variation, as well as a proposal to involve automatically annotated data in linguistic studies, and then discuss De Sutter's multivariate study on verbal cluster order that we are replicating.

3 Multivariate modeling of language variation

In corpus linguistics, linguistic phenomena examined over larger sets of data have often been found to be too complex to model in terms of a single independent variable. In this case, rather than running one statistical test for each variable, it is considered best practice to test for all variables in a single test. The statistical power of such a test will be greater than that of running several tests and applying corrections, which increases the chances of erroneously rejecting the null hypothesis with each test. Starting with Gries (2001), this methodology has been applied to the study of language variation. A well-known example is the dative alternation study of Bresnan et al. (2007). In these studies, multivariate statistical models are used to quantify the effect size of each variable, indicating their relative importance. Being corpus studies, these quantitative studies generally also emphasize evidence from larger samples of language and operationalize their variables in a precise way.

Gries (2001) discusses the case of English transitive phrasal verbs (*to pick up the book* / *to pick the book up*), and explains 84% of the variation based on a multivariate model containing many variables from previous work. He also critiques previous work on language variation. Firstly, studies often relied on introspective analysis and made-up examples, which can be too subjective, not representative, and in the case of acceptability judgements, not necessarily output of the human language system. Secondly, when only a single variable is examined at a time, many other possible variables may influence the result, even when seemingly minimal pairs are used. Lastly, the provided models cannot be used to predict variation in natural discourse situations. Variables are not weighed, and if two variables have conflicting preferences, the possibility of a prediction is already ruled out.

Despite the methodological precision, the data of the study consisted of only 403 sentences in total, about 200 for each construction. Furthermore, they were chosen manually, introducing some subjectivity into the study. When the statistical tendencies of around 30 variables are studied, 403 sentences do not provide a lot of detail. We believe that these quantitative studies can be further improved by using data from automatically annotated corpora.

3.1 Using automatically annotated data

Corpora that have been annotated by an automatic parser rather than manually, will contain far more data, allowing for larger sample sizes of particular constructions to be found. These samples can be extracted automatically as well. To do so, an exact definition of what constitutes the construction must be formulated at the level of syntax — for example, two verbs that are adjacent and in the same subordinate clause, with one being the head of the other. All of the sentences that match this definition can then be extracted from the corpus. This process avoids subjectivity beyond the definition of the construction. However, it limits the variables that can be used in the study to the ones that are, or can be, automatically annotated in a corpus. It also comes at the cost of accuracy, though in most cases, it is expected that the larger sample size makes up for any random parsing errors. Systematic errors (for example, constructions that the parser consistently fails to annotate) may skew the results however, so care should be taken that the parser is able to annotate the constructions of interest at all.

Some automatically annotated corpora have become available in recent years, though they have not yet been widely used for linguistic study. Nevertheless, I will discuss a few studies that have applied automatically annotated corpora for the purposes of language variation research. The use of automatically parsed corpora as a linguistic resource has been discussed by van Noord and Bouma (2009). They argue that parsing technology has advanced enough to be incorporated into other language technology that can build upon its results. This allowed for the creation of very large corpora of parsed sentences that are sufficiently large to compensate for any parsing error 'noise'. Several applications of such very large corpora can be found in the article.

For the Dutch language, the potential uses of the 500 million word automatically annotated LASSY Large corpus have been discussed (van Noord, 2009). This paper also mentions that some natural language processing tasks, such as learning selection restrictions of specific verbs, cannot be performed successfully using smaller corpora because the sample size would be too small. This issue is likely to apply to any linguistic phenomenon at the level of open-class lexical items (i.e. nouns, verbs or combinations thereof), most of which are rare. For the case of verbal clusters, this shows that using such a large corpus would be required to study the effect of specific main verbs on the order variation.

Automatically annotated data has already been used to study language variation for the famous case of the dative alternation, discussed in the introduction (examples (1), (2)). Lehmann and Schneider (2012) used a 580 million word dependency-parsed corpus of English to study the influence of specific lexical types on this alternation. These types consist of 'triplets' of words: a ditransitive verb, a direct object head and an indirect object head. These three slots of the triplet are all filled with open-class words, therefore requiring vast amounts of data to study: "We indeed find that 580 million words are barely enough data to yield results for full lemma triplets". Unfortunately, the study is a monovariate study where lexical type was the only variable investigated. This does not exclude the possibility of other underlying factors, which the lexemes may happen to correlate with, influencing the variation.

The only other similar variational study we are aware of is on the optionality of the Dutch *om*-complementizer, a variation similar to the optionality of English *that* for relative clauses. In Dutch, *om* as a head of a *to*-infinitival clause is optional in many, though not all cases:

(7) Anna probeerde (om) een bom onschadelijk te maken
Anna tried (CMP) a bom harmless to make
'Anna tried to defuse a bomb.'

Bouma (2013) provides further examples, and creates a model of this variation using data from part of the LASSY Large corpus. This model is multivariate, and includes both lexical effects and other variables, such as clause length. It is implemented as a mixed-effects logistic regression model, in which the verbal governor (i.e. the verb *proberen* in example (7)) is a random effect, and the other non-lexical variables are fixed effects. The other variables are all related to processing complexity, as it is theorized that *om*-insertion reduces processing load by reducing ambiguity. He finds that the best model includes both semantic and complexity features and report a concordance score of 0.809, indicating that the model has modest predictive qualities.

While there are not many examples of such studies, using automatically annotated sources of data appears to be fruitful for studying complex phenomena where many factors may play a role, and a large sample size is desirable. We consider Dutch verbal cluster variation to be such a phenomenon.

3.2 Verbal cluster variation

Going back to the topic of variation in Dutch verbal clusters, we believe that the most methodologically sound study can be found in the dissertation of De Sutter (2005), summarized in De Sutter (2009). In this study, various variables from previous work on verbal clusters were modeled using a multivariate model. Like in Gries (2001), and unlike in Bouma (2013), the starting point was not to create the most optimal model, but to determine the effect of each variable from previous linguistic work. Verbal clusters were extracted semi-automatically from a part of the De Standaard CONDIV corpus, which contains texts from a Flemish newspaper spanning a time period of a few months. By choosing this part of the corpus, the author controlled for regional, register and diachronic variation.

As well as limiting the source data, De Sutter also limited cluster types. Only clusters in complement clauses, introduced by the complementizer *dat* 'that' (no main clause clusters or other subordinates), containing a participle main verb (no infinitival clusters), and only clusters with the non-modal auxiliaries *hebben* "to have", *zijn* "to be" and *worden* "to be" have been included. Only two-verb clusters were considered. These criteria resulted in 2.390 two-verb clusters, 1.601 (66.99%) of which were in the 2-1 order. The data were then annotated for 10 variables, which were mostly extracted from the data manually, and the operationalization of the variables was carefully considered in order to be as objective as possible. The statistical model is then used to reveal the contribution of each variable towards either an 2-1 or 1-2 order choice.

We have tested the same variables, which were identified in previous literature, on our data set as far as they could be operationalized. We will summarize the data and methodology that we used in our study in the next section, and compare it to that of De Sutter (2009). The results will be discussed in section 5.

4 Method and data

We create a multivariate logistic regression model for explaining verbal cluster variation much like that of De Sutter (2009), but based on automatically annotated data. In some cases where we had to choose between creating an optimal model and creating a comparable model, we chose comparability. For example, it is generally best practice to use the most frequent value of a categorical variable as the reference value to compare the effects of the other values to. However, to maintain comparability of the effect sizes of both models we chose to use the same reference values as De Sutter. To demonstrate some benefits of automatically annotated data, we did aim to include more types of two-verb verbal clusters that exhibit optionality, including the major constructions left out by De Sutter. As mentioned, the only constructions exhibiting optionality that we did not include, are the cluster with 'other' verbs instead of auxiliary or modal verbs.

We used the Lassy Large corpus as our source of automatically annotated Dutch language data (van Noord et al., 2013). It contains texts from various written sources annotated with full syntactic dependency trees. The sentences have been parsed automatically by the Alpino parser for Dutch (van Noord et al., 2006). This parser is currently the state of the art, and an evaluation over different types of text shows an average concept accuracy (in terms of correct named dependencies) of 86.52% (van Noord, 2009). For our main comparison, we used the Wikipedia part of this corpus, which consists of the entirety of the Dutch version of the freely editable online encyclopedia Wikipedia on the 4th of August, 2011 (about 145 million words). From this data, 411.623 two-verb verbal clusters were extracted (71.65% of which were in the 1-2 order). We chose to use this part of the corpus as we believe it to be a good representation of ‘average’ standard Dutch. While Wikipedia texts have been written and edited by many speakers from different parts of the Dutch-speaking world, and probably by second language learners of Dutch as well, non-standard language is likely to be edited out by other editors. The accuracy of the Alpino parser on this text type was 88.38% in van Noord (2009), better than average but not as good as newspaper texts.

We do recognize that languages can not really be averaged, and a model based on such data will not be able to account for regional diversity or individual differences. Significant regional differences have been observed in the usage of verbal clusters (i.e. the data of Barbiers et al. (2006)), so it would be interesting, and possible, to study this using an automatically annotated corpus where authorship metadata is available. ‘Region’ could then be added as a variable to the model to explain some of the variation. In this study we won’t address this, however, we address the issue of language diversity by comparing the results from the Wikipedia part of the corpus to a model created from the Europarl part of the corpus (containing European parliament texts) in section 5.1.

The verbal clusters were extracted from the corpus using XPath 2.0 queries via the DACT command line tools. These queries precisely define what constitutes a verbal cluster, and every word group matching one of these definitions was extracted. Contextual information necessary to determine the value of the independent variables of each cluster was extracted in a similar way.

In defining and operationalizing the variables discussed by De Sutter (2009), we were limited to the information available in the annotation of the Lassy Large corpus, at least without doing any manual annotation. Some variables had to be operationalized in a different way, or could not be extracted at all. We now briefly summarize our operationalizations in comparison to those of De Sutter. For a more detailed description and motivation for each of the variables, we refer to De Sutter (2009).

The variables we operationalized are listed in the results table (1). To operationalize the TYPE OF THE AUXILIARY VERB, De Sutter divided the auxiliary verbs up into five grammatical classes: *zijn* ”to be” as a copulative verb, *zijn* as a passive auxiliary, *zijn* or *hebben* ”to have” as temporal auxiliaries, *worden* ”to be”, and unclassifiable. He developed an algorithm to identify them, which is a complex three-stage pipeline. It is described in De Sutter (2005, p. 205-230), involving 5 syntactic, 5 morphological and 2 semantic criteria. We did not try to re-create it because we would prefer to work with readily available corpus resources as much as possible for methodological demonstration purposes. The algorithm is also not perfect, hence the ‘unclassifiable’ class. Instead, we categorized the auxiliary verbs at the lexical level. We did group the modal verbs together (which De Sutter did not include in his study), as they appeared to behave very similarly in preliminary checks.

MORPHOLOGICAL STRUCTURE OF THE MAIN VERB encodes whether the main verb is separable or not. Separable particle verbs such as ‘wash up’ are written as a single word in Dutch, unless the particle is separated from the verb, which may happen even if it interrupts a verbal cluster. LENGTH OF THE MIDDLE FIELD is simply the number of words in the clause before the verbal cluster. Next, there are two variables relating to the word before the verbal cluster. INFORMATION VALUE is operationalized as the openness of this word’s class, for which there are three classes: **high**ly informational (nouns, verbs, numerals), **intermediate**ly informational (adjectives and adverbs) and **low** informational (pronouns, conjunctions and prepositions). INHERENCE refers to multi-word units (MWUs), which is a complex concept with no clear definition, but generally describes some sort of collocation of words. The corpus includes annotation on MWUs, and we make use of this annotation to decide whether the preverbal word is part of one.

EXTRAPOSED CONSTITUENTs are constituents that come after the verbal cluster, for which there are three ways to attach to the rest of the sentence: none (no constituent), attached to the main verb of the cluster, or attached somewhere higher up in the tree than the cluster (preverbally). We had to operationalize this variable differently. De Sutter made a distinction between adjuncts and complements, grouping adjuncts with ‘none’, but we could not extract this distinction from the corpus, hence the difference. The FREQUENCY OF THE MAIN VERB was estimated by counting the number of occurrences of the verb’s root in the entire Lassy Large corpus. Lastly, we added two variables to distinguish the new cluster types discussed in section 2 that De Sutter didn’t include: FINITENESS OF THE HEAD to mark infinitival clusters, and CONTEXT CLAUSE TYPE for subordinate versus main clause clusters.

There were also two variables that we could not include. Firstly, the DISTANCE BETWEEN ACCENTS, which relates to the hypothesis that order variation may occur to match the stress pattern of Dutch (Schutter, 1996). Our corpus does not contain stress or accent information, and we are not aware of a method for automatic annotation. This variable turned out to have almost no effect in De Sutter’s model. Secondly, we left out SYNTACTIC PERSISTENCE, which refers to a priming effect of a previous construction on the next. We decided not to include this variable, as we are mainly using a corpus based on Wikipedia. We cannot make sure that the writer wrote or even read the previous verbal cluster in the text. This variable did have an effect in De Sutter’s model (OR=3.28).

5 Results

Table 1 shows the comparison to the model of De Sutter (2009). The table lists all of the explanatory variables used in our logistic regression model, along with their effect size in our model and in the model of De Sutter. Based on these variables, the model can predict the dependent variable of verbal cluster word order, which is expressed as a binary variable representing either 1-2 or 2-1 word order.

For each explanatory variable in the table, one of the possible values was taken as the reference value, or baseline. The baselines were selected to be the same as those of De Sutter, in order to have comparable models. In the cases where variables had to be operationalized differently (where there are gaps in the table), the baselines are of course also different, though we tried to pick the most similar values as the reference. The effect size of each variable for both studies is given as an odds ratio. An odds ratio further from 1 indicates a stronger effect. Odds ratios > 1 indicate an association with the 1-2 word order, odds ratios < 1 indicate 2-1 order. An exception is the MAIN VERB FREQUENCY variable — it is a continuous variable, where an odds ratio cannot be interpreted in the same way. Instead, we show the β estimate of effect size (representing an effect on the dependent variable, cluster order, in terms of standard deviations), and its average standard error.

We did not perform statistical tests to assess whether the effect sizes of the two models differ significantly. Some differences are to be expected, considering the different data sources (Flemish newspapers versus Wikipedia) or other uncontrolled variables. We would mainly like to see whether the same categories show substantial effects relative to their reference values, and whether the effects are in the same direction (either the 1-2 or 2-1 order) in both models. In cases where variables were operationalized in different ways between the two studies, this is indicated by leaving the missing operationalizations blank. The reference value for each variable and study is listed as **1.00**, as the reference value can obviously not have an effect compared to itself, and 1 is the neutral value.

As a first observation for the results in table 1, we can see that the directions and size of the effects are generally similar, except where the variables had to be operationalized differently. For the variables that were operationalized the same — MORPHOLOGICAL STRUCTURE OF THE PARTICIPLE, LENGTH OF MIDDLE FIELD, INHERENCE and MAIN VERB FREQUENCY we observe similar relative effects as De Sutter (2009), and in the same directions. The INFORMATION VALUE of the preverbal constituent shows an effect in the same direction, but far smaller, and not in the same order — Intermediate is more strongly associated with the 1-2 order than High. Perhaps the larger sample size caused this. We have checked that it is not due to the additional cluster types — a model with only subordinate, finite, nonmodal clusters shows an even smaller information value effect size.

We do see differences in the two variables that had to be operationalized differently in our study. For the variable GRAMMATICAL RELATION OF EXTRAPOSITION TO HEAD we interestingly find a very strong effect, though the directions of the effects are reversed compared to de Sutter. This is a complex syntactical property, so we cannot guarantee that it was implemented in exactly the same way besides the noted difference. Either way, we can still conclude that this variable has a strong effect on verbal cluster order. The most striking difference is in the variable TYPE OF AUXILIARY. This is due to the available annotation — as discussed in the last section, a complex additional procedure would be needed to identify the grammatical classes of De Sutter. We still find somewhat of a lexical effect with our operationalization. The auxiliary verbs have a lot of grammatical ambiguity, so it is smaller than the grammatical effect found by De Sutter. We do note a strong tendency of clusters with modal verbs to occur in the 1-2 order. This confirms previous observations that modal clusters have a strong preference for the 1-2 order (Wurmbrand, 2006).

De Sutter does not provide a value of overall model fit that can be compared across models, however, we can look at the concordance index (c-index). This value is an indication of the predictive power of a model. A c-index of 0.5 corresponds with chance level, and 1 indicates perfect prediction. Like de Sutter, we report the c-index after 100 bootstrap repetitions to compensate for overfitting. He reports c = 0.803, and our model has a concordance index of 0.8635. However, it should be noted that these two c-indexes cannot be directly compared between different models, since the variables are somewhat different. These values do indicate that the models are good enough for prediction tasks. They are also similar to the c-score Bouma (2013) reported. We can furthermore look at the intercept of our model, which is 0.6035, and represents the odds of predicting an 1-2 outcome in the case where all the variables have their default value, an indication of the difficulty of the task. Clearly the model predicts better than that. However, it should be noted that this is not a typical predictive task. There is no 100% gold standard as in a parsing task for example, and both orders are grammatically correct. It might very well be that a large amount of the variation is random, depends on extralinguistic factors, or factors not captured by the annotation scheme. The focus here is mainly on finding out how much can be explained by the linguistic variables under discussion, and the effect sizes reported in table 1 are the main measure of that.

Variable	**Categories**	Odds ratio **De Sutter (2009)**	Odds ratio **This study**
Type of auxiliary	Copular *zijn*	**1.00**	
	Auxiliary of time	18.30 ***	
	Passive *zijn*	7.82 ***	
	zijn		**1.00**
	worden	11.73 ***	1.19 ***
	hebben		2.19 ***
	Modal verb		132.42 ***
Morphological structure of the main verb	Non-separable	**1.00**	**1.00**
	Separable	3.87 ***	4.92 ***
Length of middle field	0-2 words	**1.00**	**1.00**
	3-5 words	2.03 ***	2.42 ***
	6-8 words	2.29 ***	3.23 ***
	9-11 words	2.29 ***	3.34 ***
	12-14 words	2.57 **	3.33 ***
	>14 words	1.98	3.15 ***
Information value	Low	**1.00**	**1.00**
	Intermediate	1.41	1.21 ***
	High	1.94 ***	1.11 ***
Inherence	No fixed expression	**1.00**	**1.00**
	Fixed expression	2.26 ***	2.10 ***
Grammatical relation of extraposition to head	Adjunct/no extraposition	**1.00**	
	Complement of main verb	0.47 ***	
	Complement of preverbal head	1.21	
	No extraposition		**1.00**
	Comp/adj of main verb		51.44
	Comp/adj of preverbal head		0.44
Main verb frequency		$\beta = 2.44^{E-06}$ ASE=7.74^{E-07}**	$\beta = 3.73^{E-08}$ ASE=1.05^{E-08}***
Finiteness of the head	Finite head		**1.00**
	Infinite head		0.03
Context clause type	Subordinate clause		**1.00**
	Main clause		0.34

Table 1: Comparison of the size of the effect of the variables on verbal cluster variation for the two studies

One could make the objection that our model is not comparable to that of de Sutter, because it contains different constructions (main clause, infinitival and modal clusters). We controlled for these construction types with variables, but just to verify that this works, we also created a model that excludes all of these additional construction types. The observed effects were very similar to the full model. For reasons of space we cannot provide the entire table, but some of the effects are: *worden* = 2.34 (was 2.19 in the main model) *hebben* = 1.21 (was 1.19), SEPARABLE = 5.01 (was 4.92). We do note a lower c-score here of 0.7649, it appears more difficult to predict verbal cluster order when the clusters are all of the same type. This is somewhat lower than De Sutter's c = 0.803, likely due to the variables we could not include. More interestingly, we were also able to create models for specific construction types, i.e. main clause verbal clusters only, which has not been done before. For reasons of space we cannot list or elaborate on the results here, but we find that the same variables also affect main clause cluster variation. There are some differences in effect direction, mainly for the auxiliary verbs, which makes sense as it is a different construction.

5.1 Europarl corpus

In section 4 we discussed our choice for the Wikipedia part of the Lassy Large corpus. However, this corpus consists of other kinds of sources as well, and now that we have a highly automated way of building the model, it is relatively easy to test it on a different part of the corpus to see whether the same variables hold in a different domain of text. The Dutch Europarl part of the corpus consists of the translated proceedings of the European

Parliament. These texts have been written in a rather formal register by translators of the European Union. This part consists of 37 million words, from which 467.521 verbal clusters were extracted. Interestingly, that's 55.898 more clusters than were extracted from the larger (145 million words) Wikipedia corpus, indicating far more complex syntax in the Europarl corpus.

We find that 86.78% of the clusters is in the 1-2 order, a far higher percentage than is generally reported and higher than the Wikipedia corpus (71.65%). Higher proportions of 1-2 orders are generally associated with more formal registers and with editing guidelines — prescriptivists in the past have considered it to be the Dutch order, while the 2-1 order was nonstandard or German (Coussé et al., 2008). Again, we will not produce the entire table of effect sizes here — it suffices to say that all of the effect directions are the same, and the sizes are also very similar, even for variables that vary between constructions such as TYPE OF AUXILIARY. We may conclude from this that the previously discussed findings are not domain-specific. It is also a demonstration of the flexibility of using automatically annotated data.

6 Discussion

Using an automatically annotated corpus, we have shown that verbal cluster order variation is influenced by the various language-internal variables identified by De Sutter (2009) for all two-verb cluster types with order variation. While the (relative) importance of these variables for non-modal, finite clusters with a complementizer-marked subordinate clause was already established, this study shows that they apply to other types of two-verb clusters as well, even when testing on a larger and more varied dataset. We furthermore showed that these findings can be extended into another domain of text, and are not domain-specific.

Our main contribution is to show that using automatically annotated corpora is an excellent source for obtaining more data for language variation studies. Although some variables (stressed syllables) could not be measured due to a lack of automatically annotated data, advances in other language technology, such as word sense disambiguation, is likely to open up more possibilities for additional kinds of annotation to be applied to huge corpora automatically. It would also be possible to obtain more annotation by combining information from other sources. For example, to estimate syllable stress, one could consider scraping this information from phonetic transcriptions in dictionaries and adding that information to the corpus. Using these huge, automatically annotated resources seems like a natural extension of the recent trend of using large multivariate models, and allows the creation of explanatory models for uncommon linguistic constructions. We have also demonstrated the flexibility of the automatic approach — a model can easily be tested on a different dataset, provided that the right annotation is available.

De Sutter (2009) draws several conclusions, which could also be drawn from our study. Firstly, that the variation appears to be affected by 8 variables simultaneously (7 in our case), and can be predicted well enough by the model. Secondly, there are various methodological conclusions: ”We have shown that syntactic variation research needs a rigorous quantitative, corpus-based approach ...”. We can only add to this conclusion by stating that this rigorous definition of variables aids in automatic extraction of samples, which lets us retrieve all relevant constructions from large automatically annotated corpora, to the extent that the annotation allows. This, in turn, opens up options for more detailed analysis as outlined in this paper.

6.1 Linguistic interpretation

Even though we have found patterns in the variation and associations with variables that were hypothesized to be related to the phenomenon, this in itself is not an explanation of verbal cluster variation in terms of linguistic theory. We will briefly address this by referring to the hypothesis of De Sutter (2005), who states that the choice between 1-2 and 2-1 order may be related to processing difficulty. He assumes the 2-1 order to be easier to process because he considers the 1-2 order to be a prestige option, implying that the 2-1 order is the default. However, this explanation can go both ways. Recent evidence from child language acquisition supports a default 1-2 order (Meyer and Weerman, 2014). They theorize that children learn about verb raising (which forms clusters) when they acquire the 1-2 order. The most common view is that both 1-2 and 2-1 clusters come from verb raising (Evers, 1975), in which the main verb is moved up the syntactic tree to join the head. Before verb raising, the head verb is in final position, following the base Object-Verb order of Dutch. Raising the verb to form a 2-1 cluster is therefore a vacuous movement, the surface order will be the same as the underlying structure (the head comes last), and provides no evidence of any sort of special verb raising mechanism to child learners of the language. On the other hand, raising the verb to the right to create a 1-2 order violates the base word order of Dutch, and is therefore more straightforward to notice and learn. 2-1 orders can simply be interpreted as Object-Verb orders, until the learner figures out the mechanism of verb raising from the 1-2 order evidence (Meyer and Weerman, 2014). In this theory, 1-2 orders would be the earliest form of verb raising, and therefore more entrenched and easier to process.

Either way, in linguistic contexts that are more difficult to process, speakers are expected to be more likely to use the more entrenched order that is easier to process, whichever it may be. The model may be inconclusive on this matter. De Sutter argues for the 2-1 order by looking at the MAIN VERB FREQUENCY variable, stating that higher-frequency items are easier to access, and the model shows that higher-frequency words are more associated with the 2-1 order. Here, a **less** difficult context is associated with the 2-1 order. However, we can make the opposite argument for the variable LENGTH OF MIDDLE FIELD. It seems plausible that longer clauses are **more** difficult to process, and longer clauses are also associated with the 2-1 order. To invoke the processing hypothesis here, one would have to assume that the 1-2 order is the default and easier to process. Given these two opposites predicted by Meyer and Weerman (2014) and De Sutter (2005), it would be interesting to look for additional variables that are related to processing difficulty and can be extracted from the corpus automatically, such as syntactic complexity. These can then be added to the model to test which order occurs in contexts with higher processing load.

6.2 Future work

We consider several other directions for future work. The model can be extended to other domains that have been discussed in the literature, such as spoken language data from the Corpus of Spoken Dutch which uses a similar annotation scheme. As discussed earlier, regional variation is an interesting topic that could also be modeled using large amounts of data, for example by using the SoNaR corpus which includes metadata on the authors of many of its texts and adding a region variable to the model. We have yet to investigate clusters with more than two verbs, to which the automatic approach is uniquely suited. Larger verbal clusters are less frequent, and thus the best place to find rare constructions is in the largest available corpus. Now that large samples of data are easily available, a method such as collostructional analysis may be used to explore the association between particular main verbs and the 1-2/2-1 order, providing more detail on possible semantic factors.

In this study, we have aimed to follow the methodology of De Sutter (2009) closely, but this also had several downsides. It would also be possible to aim for creating the best possible model over the dataset, though differences that might arise from this (for example, different reference levels) would then make comparisons more difficult. This would allow testing of some potential methodological improvements. Multilevel modeling may be used as in Bouma (2013) to model the effects of individual lexical items. A form of Principal Component Analysis could be applied to generalize over the variables and try to reduce their number. Choosing verbal cluster order could also be viewed as a two-class classification problem, for which many other modeling methods exist.

Acknowledgements

We thank Gosse Bouma and the anonymous reviewers for their useful comments, and Gertjan van Noord and Daniël de Kok for helping to provide efficient access to the corpus data.

References

Mona Arfs. *Rood of groen? De interne woordvolgorde in tweeledige werkwoordelijke eindgroepen met een voltooid deelwoord en een hulpwerkwoord in bijzinnen.* Göteborg University, 2007.

Sjef Barbiers, H Bennis, G De Vogelaer, M Devos, and M van der Ham. Dynamische syntactische atlas van de Nederlandse dialecten (DynaSAND). *Amsterdam, Meertens Instituut. URL: http://www.meertens.knaw.nl/sand*, 2006.

Gosse Bouma. Om-omission in Dutch verbal complements. *Manuscript in preparation*, 2013.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, R Harald Baayen, et al. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.

Evie Coussé. *Motivaties voor volgordevariatie. Een diachrone studie van werkwoordvolgorde in het Nederlands.* Universiteit Gent, 2008.

Evie Coussé, Mona Arfs, and Gert De Sutter. Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep. Een taalgebruiksgebaseerd perspectief op de synchronie en diachronie van de zgn. rode en groene woordvolgorde. *Taal aan den lijve. Het gebruik van corpora in taalkundig onderzoek en taalonderwijs*, pages 29–47, 2008.

Gert De Sutter. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen.* University of Leuven: PhD thesis, 2005.

Gert De Sutter. Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. *Describing and modeling variation in grammar*, 204:225–254, 2009.

Hans Den Besten and Jerold A Edmondson. The verbal complex in continental West Germanic. *On the formal syntax of the Westgermania*, 3:155–216, 1983.

Arnold Evers. *The transformational cycle in Dutch and German*, volume 75. Indiana University Linguistics Club Bloomington, 1975.

Arnold Evers. Verbal clusters and cluster creepers. *Amsterdam studies in the theory and history of linguistic science, Series 4*, pages 43–90, 2003.

Stefan T Gries. A multifactorial analysis of syntactic variation: particle movement revisited. *Journal of quantitative linguistics*, 8(1):33–50, 2001.

Liliane Haegeman and Henk van Riemsdijk. Verb projection raising, scope, and the typology of rules affecting verbs. *Linguistic inquiry*, pages 417–466, 1986.

Hans Martin Lehmann and Gerold Schneider. Syntactic variation and lexical preference in the dative-shift alternation. *Language and Computers*, 75(1):65–75, 2012.

Caitlin Meyer and Fred Weerman. Cracking the cluster: The acquisition of verb raising in dutch. *Manuscript in preparation*, 2014.

G de Schutter. De volgorde in tweeledige werkwoordelijke eindgroepen met voltooid deelwoord in spreek-en schrijftaal. *Nederlandse taalkunde*, 1:207–220, 1996.

Jan Stroop. Systeem in gesproken werkwoordsgroepen. *Taal en tongval*, 22:128–147, 1970.

Jan Stroop. Twee- en meerledige werkwoordsgroepen in gesproken Nederlands. *Fons verborum*, pages 459–469, 2009.

Gertjan van Noord. Huge parsed corpora in lassy. *Proceedings of TLT7. LOT, Groningen, The Netherlands*, 2009.

Gertjan van Noord and Gosse Bouma. Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 33–39. Association for Computational Linguistics, 2009.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-30909-0. doi: 10.1007/978-3-642-30910-6_9. URL `http://dx.doi.org/10.1007/978-3-642-30910-6_9`.

Gertjan van Noord et al. At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, 2006.

Susi Wurmbrand. West Germanic verb clusters: The empirical domain. *Verb clusters: A study of Hungarian, German, and Dutch*, pages 43–85, 2004.

Susi Wurmbrand. Verb clusters, verb raising, and restructuring. *The Blackwell companion to syntax*, pages 229–343, 2006.

Shalom Zuckerman. *The acquisition of "optional" movement*. PhD thesis, University Library Groningen [Host], 2001.

Jan-Wouter Zwart. Verb clusters in continental West Germanic dialects. *Microparametric syntax and dialect variation*, pages 229–258, 1996.