

Lexical preferences in Dutch verbal cluster ordering

Jelke Bloem

This study discusses lexical preferences as a factor affecting the word order variation in Dutch verbal clusters. There are two grammatical word orders for Dutch two-verb clusters, with no clear meaning difference. Using the method of collocation analysis, I find significant associations between specific verbs and word orders, and argue that these associations must be encoded in the lexicon as lexical preferences. In my data, the word orders also show some semantic associations, indicating that there might be a meaning difference after all. Based on these findings, I conclude that both word orders are stored in the lexicon as constructions.

1. Introduction

In some contexts, a speaker of a language can choose between different constructions to express a similar meaning. A well-known example of two such near-synonymous constructions in English is the dative alternation: [S *gave* DO *to* IO] or [S *gave* IO DO]. The Dutch language has a word order variation where two orders express what appears to be the same meaning: the word order in clusters of two verbs. In this study, I tested for lexical associations with a particular word order using collocation analysis, and discuss whether the two word orders should be viewed as different constructions.

In Dutch, as in other Germanic languages, grammatical categories such as tense and aspect can be expressed with auxiliary verbs. These verbs are grouped together with the main verb (a lexical verb) into verb clusters, in which the order of the verbs may vary. Unlike other Germanic languages, Dutch exhibits this variation even in clusters of two verbs. Both possible two-verb cluster orders are grammatical, leading to optionality:

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

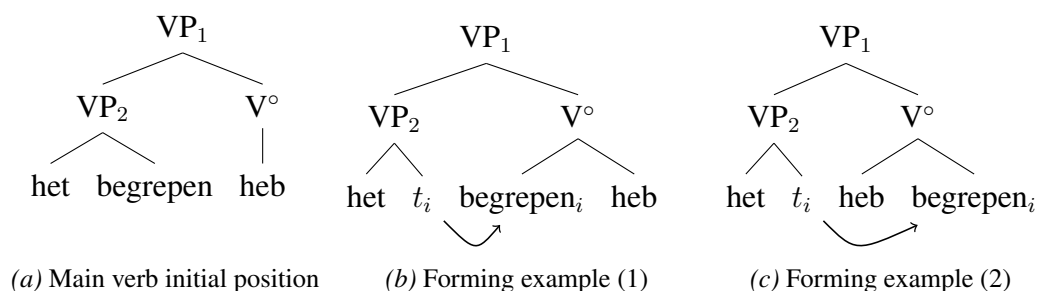


Figure 1: Verb raising to generate the verbal clusters of examples 1 and 2, following the analysis of Evers (1975).

Speakers may produce either order in similar contexts. A notable aspect of this optionality is that the difference in word order does not appear to correspond to a meaning difference. However, when speakers choose between constructions in these situations, they do not do so randomly.

Example 1 shows the 2-1 order, so called because the syntactically higher head verb (referred to as 1) comes after the lower lexical verb (referred to as 2). This is the order that is used in German. Example 2 shows the 1-2 order, where the head verb comes first. This order is used in English. In Dutch, there are no such grammatical restrictions and speakers may produce either order. A notable aspect of this optionality is that the difference in word order does not appear to correspond to a meaning difference (e.g. Bader et al. 2009), although it has been argued that there is a difference in interpretation (Pardoën 1991).

The word order variation found in verb clusters is commonly analyzed as the result of a purely syntactic process, in which one order is the base form, and others are derived from it. These analyses rely on various syntactic movements to account for the observed word orders (e.g. Evers 1975; Barbiers 2008). The original proposal of Evers (1975) states that verbal clusters are formed through the mechanism of verb raising. In verb raising, the main verb is generated as the complement of the head verb of the cluster and then moves up to join the head verb. Figure 1 illustrates this process of cluster formation. The main verb *begrepen* is raised to attach to the governing verb *heb*, forming a complex head. It can attach on either side of the head verb, resulting in either a 2-1 (Figure 1b) or 1-2 (Figure 1c) ordered verbal cluster.

Lexical-semantic considerations generally do not come into play in these purely syntactic approaches. For example, Barbiers (2008) states that semantic interpretation takes place with the entire verb cluster and its syntactic structure as an ‘atomic’ unit and that it only gets its word order during phonological interpretation (PF), which comes after semantic interpretation (LF) in minimalist grammar. Under this view, lexical semantics and verb cluster word order cannot interact.

However, when speakers choose between grammatical verb cluster orders, they do not choose randomly. Corpus studies have shown that a large portion of the variation can be statistically explained using multifactorial statistical models that incorporate a variety of linguistic factors. De Sutter (2005) and Bloem et al. (2014) found that morphological and semantic factors correlate with the variation, as well as processing-related properties such as sentence length and word frequency. This finding indicates that the syntactic mechanisms that produce verb clusters do

not operate independently from the non-syntactic factors discussed in these studies. At the same time, these models do not account for all of the variation. Bloem et al. (2014) report a concordance index of $c = 0.7649$ over the constructions investigated in the present work. The c-index measures the predictive power of a model — a value of 0.5 corresponds to chance level, and 1 indicates perfect prediction. This means there are other factors that correlate with the choice between verb clusters, but are not included in these statistical models.

In this study, I look for evidence for such factors by focusing on the usage patterns of main verbs in verb clusters in a corpus of Dutch. Previous corpus studies on verb clusters have not included main verbs, possibly due to the need for a very large corpus to find enough instances of specific verbs in verb clusters. Main verbs are the most likely source of variation that has not been studied yet, besides factors that cannot be studied using written corpora of standard language. By studying the main verbs that speakers prefer to use in each order, we can learn more about the choice between the two word orders.

I will obtain this information by computing associations between a specific verb and a specific word order using a statistical measure of association, i.e. verbs that occur significantly more often in the 1-2 order and verbs that occur significantly more often in the 2-1 order. I will then compare the two orders. These associations will be referred to as ‘lexical associations’, a measure of association in a corpus. This procedure is known as a distinctive collexeme analysis (Gries & Stefanowitsch 2004).

Lexical associations with an open class of words such as verbs or nouns, with no selectional restrictions, are not often investigated. Most likely, this is because there are so many possible words that could be used, and most of them are relatively infrequent. For example, in a 1.7 million word sample from the 700 million word Dutch Lassy Large corpus (van Noord et al. 2013), we can find 2.570 instances of two-verb verbal clusters in the 2-1 order, but to model lexical associations, one needs to find enough instances of the word order with particular verbs. The most frequent main verb in this sample occurred 129 times and the fifth most frequent verb occurred only 31 times, and this number strongly decreases for less frequent verbs following Zipf’s law. A sample of this size would give us insufficient data for all but the most frequent of words despite the overall size of the corpus.

With the increasing availability of large, automatically annotated corpora, it is now possible to obtain large samples of constructions using the same word, making it possible to compute lexical associations and study lexical preferences. I am aware of one previous study taking this approach. Lehmann & Schneider (2012) used a 580 million word dependency-parsed corpus of English to study the influence of specific lexical types on the English dative alternation. These types consist of ‘triplets’ of words: a ditransitive verb, a direct object head and an indirect object head — these slots are all filled with open-class words, making it a more difficult problem than main verbs in verb clusters. The authors note that ‘We indeed find that 580 million words are barely enough data to yield results for full lemma triplets’.

In this study, I use the Wikipedia section of the Lassy Large corpus, which is currently the largest corpus for Dutch, to automatically extract all 411,623 two-verb verbal clusters of the type investigated in this study. The size of this data set is sufficient to study lexical associations for a single open slot (the main verb in a verb cluster). I am not aware of any other studies applying collostructional analysis methods to a large dataset from an automatically annotated corpus.

Lexical associations can emerge for various reasons, and are not easily accounted for. Firstly,

observed lexical associations can be a consequence of more general factors, such as the ones investigated in the studies just mentioned. It might be the case that the construction prefers a larger class of verbs, and that particular verb is just one of them. For example, if a particular verb is significantly associated with the 1-2 word order, and it is morphologically complex, it could be because the general factor ‘morphological complexity’ is associated with the 1-2 order. Because these effects are already known, I have not studied them in detail, but I have tried to control for them. Secondly, lexical associations can indicate a lexical preference of the word order, and lastly, they can indicate a semantic preference of the word order. When controlling for more general factors, lexical and semantic preferences are two factors we might expect to find by investigating the usage patterns of main verbs. In the remainder of the introduction, I will discuss these factors in more detail.

1.1. Lexical association and preference

Because human language learning appears to make use of statistical learning abilities (e.g. Rebuschat & Williams 2012), it might be expected that the specific verbs speakers use play a direct role in the choice between grammatical verb cluster orders. A word that is more often heard in one of the two possible word orders by a speaker for whatever reason, may also be produced more often in that order, preserving the association. This phenomenon has been called lexical association (Hindle & Rooth 1993), and is also known as ‘lexical idiomaticity’ (e.g. Speelman et al. 2009).¹ When this association cannot be attributed to other, more general factors, it must be a lexical preference. A lexical association can be observed, but to establish that it is a lexical preference, other factors must be excluded. Conversely, if lexical preferences do not exist, one would expect that no statistically significant associations will be found, after controlling for other known factors that affect verb cluster word order and are properties of the main verb.

Such a preference has to be stored in the mental lexicon in some way, for example, as features of the word in the lexicon (in lexicalist grammars), or as a weighted link between the word and the word order construction in a constructional network (in a construction grammar). There are three logical options: the association is stored as a preference of the verb, as a preference of the construction whose form is the 1-2 order, or as a link between the verb and the 1-2 order in a network of constructions. The first option is unlikely because the main verb of a verb cluster is the complement of the verb cluster construction, which should not be able to affect word order according to most theories of language. It is more likely that any such preference would be stored as a preference of the verb cluster order construction for particular verbs.

In this paper, I will not discuss why particular words may prefer particular cluster orders (or the reverse). These preferences could have various origins, such as historical usage patterns, which would have to be explained for each verb individually. My interest is in simply finding evidence for or against the existence of such preferences.

Current theories on verb cluster formation do not allow for lexical preferences to affect

¹The term ‘lexical idiomaticity’ as defined by Speelman is more broad, though. It includes lexical restriction as well as association. When a construction has lexical restrictions, some combinations of words and constructions are ungrammatical rather than just preferred or dispreferred. I will use the term lexical preference as there are no apparent restrictions, lexical or otherwise, on two-verb cluster ordering with participial verbs.

word order. If such lexical preferences exist, we will have to account for them. An analysis that accounts for the role of syntax, semantics and lexical semantics simultaneously can be provided by construction-based approaches to language (e.g. Fillmore 1985; Goldberg 2002). In these approaches, syntax (form) is inextricably linked to semantics (meaning). A ‘lexicon’ of constructions is proposed, sometimes called a constructicon, which contains form-meaning pairs at various levels of abstraction, ranging from a highly abstract subject-predicate construction to fixed multi-word expressions to individual words. In construction-based approaches, both verb cluster orders would be distinct constructions, since they differ in form. As a consequence, they can have distinct properties that are stored in the constructicon, such as semantic properties, functional differences or links to other constructions.

A problem for construction-based approaches to verb clusters is the lack of clear semantic differences between the two orders. According to the approach just described, a form without a distinct function should not exist. Perhaps there is just a single verbal cluster construction, and factors external to the lexicon determine the word order. Alternatively, it might be the case that the verb cluster orders do have semantic properties, but they have not been noticed yet in previous research. One study by Pardoën (1991) does argue that there is an interpretation difference between the two orders — in this view, 1-2 orders are assigned a dynamic interpretation, while 2-1 orders are assigned a stative interpretation. For this reason, I have also tested for semantic preferences of the word orders in this study, including *STATIC* and *DYNAMIC*.

1.2. *Semantic preference*

As I mentioned earlier, lexical associations might be explained by the fact that the lexical verb belongs to a larger class of verbs, which the construction prefers. And according to construction-based approaches to language, semantic differences should exist. It therefore makes sense to test for semantic preferences among the main verbs used with the word orders. Such semantic preferences have also been observed for the English dative alternation for example, where the semantic class of the ditransitive verb (*TRANSFER* or *COMMUNICATION*) statistically explains part of the observed word order variation (Bresnan et al. 2007).

Because Dutch verbal cluster word orders do not appear to express a meaning difference, I did not have a hypothesis on what sort of semantic preferences might be found among the verbs associated with a particular order (perhaps with the exception of adjectival participles, which I will discuss later). Nevertheless, I have studied general semantic properties of the associated verbs in an attempt to find semantic associations.

If the word orders have distinct semantic preferences, and if lexical associations with verbal cluster word orders are observed, this would support the hypothesis that the word orders are constructions. To look for lexical associations that may indicate lexical preferences, and for semantic associations of the verb cluster orders, I will use a collostructional analysis method.

The rest of the paper is structured as follows. I discuss how lexical associations can be measured in section 2. Section 3 will discuss related work on Dutch verb clusters, and section 4 describes the corpus data I used for this study, as well as the way in which the analysis was performed. Section 5 presents the results of the study, followed by a discussion in section 6.

2. Measuring lexical associations

Various corpus studies have already been performed where associations between words and syntactic structures were used as evidence, particularly studies focusing on the meaning difference between constructions (e.g. Hilpert 2008; Coleman 2009; Levshina 2011). A method has been developed to measure such associations between words and constructions, called collostructional analysis (Stefanowitsch & Gries 2003). Statistical association measures, such as χ^2 , are used to calculate which words (or lexemes) co-occur with a construction most strongly in a specific syntactic position. A lexeme that is strongly associated with a construction is called a collexeme of that construction, and the combination of the collexeme and the construction is called a collostruction (hence the name collostructional analysis).

Collostructional analysis is an extension of collocational analysis. While collocational analysis measures co-occurrence of lexical items, collostructional analysis measures the association strength between constructions (usually words) that occur in other constructions (usually a grammatical structure). The method is based on the theoretical assumptions of the Construction Grammar approach to language, where constructions are seen as pairs of form and meaning and some part of the form or some part of the meaning is not entirely predictable from the components of the entire construction. The scope of what is considered a construction is quite broad, ranging from individual words to more abstract grammatical elements such as the past tense construction, which can be filled with any verb. Constructions can also be partially filled, such as a certain verb having a particular subcategorization frame, e.g. [SUBJ *give* O_i O_d]. In this case, the verb *give* is filled in, while the other elements are open slots that can be filled, e.g. O_d can be filled with a direct object that is semantically compatible with the construction. Any lexical item can occur in a more abstract construction as long as the lexical item is semantically compatible (though not necessarily identical) with the meaning of the construction it occurs in (Stefanowitsch & Gries 2003:p. 209-213).

The procedure described by Stefanowitsch & Gries (2003) is as follows. First, one particular construction that has one or more open slots to be filled by lexical items is chosen to be analyzed. Next, with the help of manual inspection and coding, all the lexemes occurring in the slot are extracted from a text corpus (Stefanowitsch & Gries 2003:p. 214-215).

In order to calculate the strength of association between the lexemes and the construction, Stefanowitsch & Gries chose the statistic Fisher's Exact Test (FET) since it is able to handle low-frequency data and it does not make any distributional assumptions. The input to FET is a 2-by-2 contingency table that contains the single and joint frequencies of the construction and the given lexeme, i.e. the frequencies of the lexeme in the construction, the lexeme in other constructions, the construction with other lexemes and finally all other constructions with all other lexemes. Table 1 is an example of this. The output is a p-value according to which the collostructions can be ranked: the smaller the p-value, the more strongly associated the construction and the collexeme are (Stefanowitsch & Gries 2003:p. 218-219).²

Finally, by way of linguistic analysis, the first ten to thirty ranks of the collexemes are ex-

²Subsequent studies have used measures of effect size such as the odds ratio for comparison and ranking, since p-values should not be compared directly — p-values of null-hypothesis significance tests are either significant or they are not, further interpretation is not meaningful.

amined, and Stefanowitsch & Gries classify them according to their semantic and sometimes also syntactical properties. For example, they state that all the collexemes occurring as objects of *cause* have a negative connotation. Furthermore, they group them according to their argument structure (Stefanowitsch & Gries 2003:p. 221). As a result of knowing what collexemes a construction has, they claim that one can objectively identify the meaning of the construction.

The work was later extended to a family of methods, including distinctive collexeme analysis (Gries & Stefanowitsch 2004), which provides a measure of the preference of a collexeme for one of two different constructions. The collexeme lists of two constructions are compared directly. Distinctive collexeme analysis is typically used for identifying meaning differences between grammatical constructions. Some of the collexemes that are associated with the same constructions generally share some semantic properties, as shown by Stefanowitsch & Gries (2003).

In the present work, I apply the collostructional analysis method to a phenomenon in which the form-meaning or form-function mapping is not so clear: verb clusters in Dutch.

3. Verb clusters

The Dutch language, like other Germanic languages, can use auxiliary verbs to express properties such as tense and aspect. This creates groups of verbs which, in some of the West-Germanic languages, have particular syntactic properties. When this is the case, these verb groups are called verb clusters. Languages with verb clusters have some degree of word order variation in multiple verb constructions with more than two verbs, and they all exhibit the *Infinitivus Pro Participio* (IPP) effect, where modal verbs that would normally be participles show up as infinitives (Wurmbrand 2006). Besides Dutch, verb clusters are also found in German, Frisian, Swiss-German, Afrikaans and various dialects and subgroups related to these languages. Verb clusters are a widely studied phenomenon in the syntactic literature on Dutch, particularly for their syntactic structure and for their optionality (Evers 1975; Haegeman & van Riemsdijk 1986; Zwart 1996; Barbiers 2008). The mechanism of verb raising is used to account for the existence of verbal clusters as a syntactic phenomenon, as illustrated in the introduction. A broad overview of verb raising across Germanic languages, as well as different theoretical accounts of the phenomenon, is provided in Wurmbrand (2006).

Dutch verb clusters were originally noted for their regional variation. In the literature, the 2-1 order construction in Example 3 (repeated from the introduction) is often called the ‘groene volgorde’ (green order), and the 1-2 order of Example 4 is called the ‘rode volgorde’ (red order):

- (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

This terminology dates back to a study on regional differences by Pauwels (1953). The orders were named after their color on a dialect map. This terminology has been used among Dutch

linguists since then, though because it is somewhat confusing and does not cover the cases with more than two verbs, alternatives have been used. In my terminology (1-2 and 2-1 order) I follow Stroop (1970), who calls the green order *descending* and the red order *ascending*. These terms refer to the idea that the finite auxiliary is the verb that is highest in the syntactic tree, while the main verb is the lowest. This lets us number the verbs: example 3 shows a 2-1 order. Larger clusters can thus be adequately described, e.g. a 4-3-2-1 cluster, where 4 is the main verb, although discussion in this paper is limited to two-verb clusters.

There are various types of two-verb clusters to consider, all of them exhibiting free variation but with different probabilities. Examples 3 and 4 show two-verb clusters with auxiliary heads, and a participial main verb. When I discuss clusters with auxiliary verbs (or ‘auxiliary clusters’), I am referring to the following types of auxiliary verbs that can occur as the head of a verb cluster:

- Auxiliaries of time (*zijn* ‘to be’ and *hebben* ‘to have’)
- The passive auxiliary (*worden* ‘to be’)
- The copular verbs *zijn* ‘to be’ and *worden* ‘to be’

These three verb lemmas are all of the auxiliary verbs that take participial main verbs, and this categorization follows the terminology used by De Sutter (2005). The two copular verbs are only included because they have the same form as other auxiliary verbs, which makes them difficult to distinguish in a corpus. Auxiliary clusters as defined here appear to exhibit the greatest amount of variation, and the present study includes only auxiliary clusters.

Other auxiliary verb types can occur as heads of verb clusters. The most frequent of these are clusters with modal verbs, such as *kunnen* ‘can’. These clusters are used in the 1-2 order overwhelmingly frequently, though 2-1 orders with modals do occur and are considered grammatical by most speakers (Pauwels 1953). Furthermore, clusters can be formed with the causal auxiliary *laten* ‘let’, and with some non-auxiliary grouping verbs such as *zien* ‘see’, or other copular verbs such as *blijven* ‘remain’. These constructions are also grammatical in both orders. I consider all these to be different constructions from the auxiliary clusters, since they have clearly different word order preferences.

Besides the theoretical syntax work on verbal clusters, there have been various studies looking into other aspects of language as possible explanations for the order variation. The rules and mechanisms that are used to account for the phenomenon of verb clusters in generative literature allow for a lot of optionality, and thus mainly outline the constructions in which order variation can occur. Coussé et al. (2008) provide a summary of recent work on verbal cluster variation, summarizing three dissertations on the topic (De Sutter 2005; Coussé 2008; Arfs 2007). A diverse set of factors that may influence the use of 2-1 and 1-2 orders has been discussed in these works, and Coussé et al. (2008) group them into four broad categories.

- Contextual factors include the regional background of the speaker and mode of communication. This factor does not affect intra-speaker variation, however.
- Rhythmic factors relate to the hypothesis that speakers may change the order of verbs to match the standard stress pattern of Dutch. This may not be so important in written

language, though. De Sutter (2009) did not find a strong effect of stressed syllables near the cluster in his corpus study of written texts.

- Discourse factors are mentioned with syntactic priming as an example. In a psycholinguistic study, Hartsuiker & Westenberg (2000) showed that verbal cluster orders can undergo structural priming, which also suggests that both verbal cluster orders may be considered to be distinct linguistic units or constructions.
- Semantic factors are also mentioned, even though verb cluster orders are said not to differ in meaning. Coussé et al. (2008) mostly seem to refer to the semantics of the auxiliary verb when discussing this factor, not the main verb. There is one exception to this, however. De Sutter (2005) showed that adjectival participial main verbs, such as *vermist* ‘missing’ (the man is missing / the missing man), occur more in the 2-1 order. This is probably because a real adjective could not come after the head verb in Dutch, a language that is verb-final at least in subordinate clauses.

There have been a few studies that took many factors into account at the same time, initially by De Sutter (2009), and expanded by Bloem et al. (2014). These two studies tested the association of 10 factors with various types of verbal clusters, and found clear effects for a variety of factors, such as the type of auxiliary verb and the sentence length. However, the models used in these studies do not explain all of the variation observed in the corpora. Bloem et al. (2014) reports a c-score of 0.76, where 1 is a perfect model (though that should be impossible to achieve). One possible cause of this is that they do not take the main verb into account as a factor in their model. There are simply too many possible main verbs to include as factors in a logistic regression model. In contrast, the method of collostructional analysis used in the present work makes it possible to quantify lexical associations, allowing for the analysis of the factor of main verb lexical preferences. However, collostructional analysis methods are not multifactorial, and therefore do not control for the other factors directly. Some of the factors that were found to affect verb cluster order by De Sutter (2009) and Bloem et al. (2014) are contextual, such as the length of the middle field of the clause the verb cluster occurred in. Such factors are not likely to affect lexical associations — there is at most an indirect link. For example, there could be a verb that occurs only in long clauses, and a lexical association with a word order computed for such a verb might also include the association between long clauses and that word order. However, other factors, such as the morphological complexity of the verb, are properties of the verb and may also affect the computed lexical associations. I will take these factors into account when I discuss the results. Next, I will describe the data and method used to perform this analysis.

4. Method and data

For my analysis, I have extracted data from a syntactically annotated corpus, the Wikipedia part of the Dutch-language Lassy Large corpus (van Noord et al. 2013). The Lassy Large corpora have their own search tools with which they can be queried, and in this way I automatically extracted the verb cluster constructions I am looking for. In this section, I will describe how I

	<i>horen</i>	\neg <i>horen</i>	totals
2-1	<i>c</i> =406	367,181	<i>vn</i> =367,587
\neg 2-1	974	198,876	199,850
totals	<i>cn</i> =1,380	566,057	<i>n</i> =567,437

Table 1: An example contingency table for the verb *horen* ‘to hear’, used in collostructional analysis

extracted the relevant data from the corpus and applied the collostructional analysis method to it.

I used a corpus that is fully syntactically annotated because I need to know about syntactic relations between verbs in order to extract the constructions that are relevant for the analysis. I need to know which verbs belong together in a cluster (rather than just being adjacent in the text but in different clauses), and which verb is syntactically higher. This kind of information is only included in corpora with full syntactic annotation, such as the treebanks created by natural language parsers.

4.1. Method of analysis

As mentioned before, I use distinctive collexeme analysis to obtain lexical associations, and perform this analysis on large amounts of corpus data.

In their first study using collostructional analysis, Stefanowitsch & Gries extracted their data using item-by-item inspection and manual coding. They argue that the syntactic (tree) structure is not always enough to extract a construction, and that the constructions should be looked at on a more abstract level of representation. However, I decided that it is better to automatically extract every syntactic pattern of a construction that occurs. This allows for the analysis of more data, at the risk of including some errors. These errors are less likely to be a significant factor in the analysis when there is more data. Furthermore, by using an automatically annotated corpus for Dutch, I have already introduced the chance of errors anyway — the annotation may contain errors, since automatic parsers are not perfect. Because verbs are still analyzed correctly most of the time, the use of a frequency cutoff for my data should mitigate their effect on the outcome of the analysis. Manual inspection of the data showed that erroneously annotated word forms only appear with low frequency, so by using a frequency cutoff and excluding low-frequency words, I mostly exclude such errors. The use of a frequency cutoff is also important to make sure that only words that are frequent enough for statistical analysis are included. Furthermore, to ensure that the extraction queries only returned the constructions I am looking for, I manually tested and inspected samples of the data that the queries returned.

In my distinctive collexeme analysis, I test for associations between each verb cluster order and particular main verbs. The method considers a construction to be a grammatical structure with open slots or variables for words. Example 4 can be abstracted to a structure [*heb* V], in which V is an open slot that takes any main verb that is semantically compatible with this construction (with *hebben* ‘to have’). Associations between instances of V and this particular

construction [*heb V*], as opposed to the construction [*V heb*], can then be calculated using a statistical measure of association, similar to collocation identification.

There are four frequency counts I need to extract from the corpus to be able to perform this calculation, as shown in Table 1. The value *c* is the frequency of co-occurrence of a potential collexeme with the 2-1 order construction. To extract this I looked up all instances of the construction and counted the frequencies of the verbs occurring in them. The value *vn* is the frequency of occurrence of the potential collexeme in all verb clusters. The value *cn* is the frequency of occurrence of the 2-1 word order in verb clusters in the corpus. Lastly, the contingency table requires the frequency of all constructions with all lexemes, *n*. In their analysis of the construction [*N waiting to happen*], Stefanowitsch & Gries (2003) use the total number of verb tags in their corpus, since their construction centers around the verb ‘wait’. Since I only look at verbs in clusters, I use the total number of verb clusters. The other values in the contingency table can be calculated automatically based on these four values.

As mentioned previously, a disadvantage of the distinctive collexeme analysis method is that other factors that may affect the lexical association cannot be controlled for in the computation, besides the factor ‘frequency’. This means that part of the lexical association may come from more general factors, rather than from a lexical preference. To compensate for this, I ran some additional tests to exclude certain factors. Furthermore, as in other collostructional analysis experiments, I performed a semantic analysis of the results to look for general semantic factors. For example, Stefanowitsch & Gries (2003) analyze their lists of collexemes by looking for semantic generalizations in them, such as ‘verbs denoting ways of speaking cleverly and deviously’. I have taken a more objective approach to this analysis by using semantic properties of verb senses from the Cornetto lexical-semantic database (Vossen et al. 2013).³

4.2. Corpus data

To obtain data on Dutch verb cluster constructions, I used the Wikipedia section of the Lassy Large corpus. This section consists of the entire contents of the Dutch-language Wikipedia website on the 4th of August, 2011, automatically annotated with the Alpino parser for Dutch. This parser is currently the state of the art for Dutch, performing with an average concept accuracy (in terms of correct named dependencies) of 86.52% (van Noord 2009). The accuracy of the parser on Wikipedia text specifically is 88.38%. Lassy is the largest syntactically annotated corpus for Dutch, and the Wiki part contains 145 million tokens. While its syntactic annotation is automatically assigned and may contain errors in its analysis of verb clusters, the size of the dataset is very important in collostructional analysis. Most of the collexemes occurring in any given construction occur rarely in that construction (Stefanowitsch & Gries 2003), just as most words are infrequent (Zipf’s law), so a large dataset is required in order to draw general conclusions about collexemes.

The corpus is stored in an XML-based format, containing the syntactic tree structure and various attributes of each node in the tree that were annotated. To indicate non-local dependencies (for phenomena such as WH-movement and subject extraction), commonly a problem for tree

³ Accessed via the web interface available at <http://cornetto.clarin.inl.nl/index.html>

representations of syntactic structure, there are links between different parts of the tree.

The Lassy corpus can be explored and searched using a set of tools⁴, one of which allows users to search the treebank using XPath⁵, a standard querying language for XML. This tool makes it possible to select nodes of the treebank based on their attributes (such as finite verbs) or based on what nodes it is linked to in the tree (for example with daughter nodes that are also a verb phrase, forming a cluster).

To obtain the necessary frequency counts for a collostructional analysis, I started out by querying for all of the sentences containing two adjacent verbs in the same subordinate clause, checking that one of them (the main verb) is below a VP in the syntactic tree. I took various exceptions into account, such as conjunctions of VPs, or cases in which there is a verb particle or infinitival marker between the two verbs. Therefore, any particle verb with a stranded particle is also included in the dataset, and counted as that particle verb, not as its root. I also had to make sure the two verbs were not part of a larger verb cluster, since I am only looking at two-verb clusters in this study.

Once I had obtained all the sentences containing something that matched my definition of a two-verb cluster with an auxiliary head, I retrieved their treebank files and stored them as subcorpora, containing only those sentences with verb clusters, to speed up querying. From this, I extracted all the frequency counts necessary to perform a collostructional analysis using the alpinocorpus-python API⁶, an interface for querying the corpus data and processing it using the Python programming language.

I extracted all of the verbs that occur as the main verb of a verb cluster, headed by one of the three auxiliary verbs that take participial main verbs: *hebben* ‘to have’, *worden* ‘to be’ and *zijn* ‘to be’, as listed in section 3. Most of the variation in Dutch two-verb cluster orders can be found in these constructions. I extracted the verbs by their lemma attribute, such as *afvragen* ‘to wonder’, rather than the verb root, such as *vraag-af* or surface forms, such as *afvroegen* ‘wondered’. From this list of verbs I counted the frequencies for the contingency table as described in section 4.1.

To measure the association strength between words and constructions based on these frequencies, I used two different measures. One is Fisher’s Exact Test, as proposed by Stefanowitsch & Gries (2003). I furthermore compute odds ratios as a measure of effect size. The associations were calculated using the *fisher_exact* function from the SciPy library for Python⁷, which provides odds ratios as well as p-values from the test.

The p-value provided by Fisher’s Exact Test merely allows one to say whether an association is significant or not, based on a threshold (such as $\alpha = 0.05$). It does not allow ranking the verbs, while ranking is essential for a collostructional analysis. Furthermore, there is not an established threshold between association and non-association to use when performing a collostructional analysis. Therefore, I use the odds ratio as a measure of effect size. Using this ratio we can say, for example, that a verb is two times more strongly associated with the 1-2 order than with the 2-1 order, though we need the Fisher’s Exact Test to say whether this is statistically significant. Using these odds ratios, I can rank and compare the collexemes.

⁴Available at <http://rug-compling.github.io/dact/>

⁵Description available at <http://www.w3.org/TR/xpath/>

⁶API available at <https://github.com/rug-compling/alpinocorpus-python>

⁷SciPy library available at <http://www.scipy.org/>

Rank	1-2 order	Association	2-1 order	Association
	Collexeme		Collexeme	
1	toebehoren (76:0) 'to belong to'	inf ***	bekend (2064:75) 'to confess'	50.36710 ***
2	malen (56:0) 'to grind'	inf ***	kampen (157:6) 'to deal (with)'	47.69147 ***
3	toeleggen (212:6) 'to focus'	19.382863 ***	vertrouwen (49:2) 'to trust'	44.643377 ***
4	opzeggen (70:2) 'to cancel'	19.206565 ***	noodzaken (121:5) 'to necessitate'	44.103544 ***
5	plaatsvinden (1004:30) 'belong to'	18.323638 ***	benodigd (118:5) 'to require'	43.009776 ***
6	kwijtraken (171:8) 'to take place'	11.726689 ***	bevriend (128:7) 'to befriend'	33.325336 ***
7	binnenkomen (63:3) 'to come in'	11.524063 ***	verplichten (391:23) 'to oblige'	30.999251 ***
8	opsommen (55:3) 'to sum up'	10.06087 ***	gebruikmaken (90:7) 'to make use'	23.429862 ***

Table 2: Top eight collexemes of the 1-2 and 2-1 verb cluster orders

5. Results

In this section, I first describe the results of the standard distinctive collexeme analysis on the data just described, and then the additional tests that were performed to control for various general factors that are known to relate to verb cluster order variation.

5.1. Initial analysis

Following the method of distinctive collexeme analysis outlined in section 4.1, I have ranked the verbs that occur in each word order by their association strength with the word order. This means that verbs that occur in the 1-2 or 2-1 order more often than would be expected by chance are ranked higher. In this analysis, I have only used verbs that occur more than 50 times in the data set, to avoid spurious results for infrequent verbs. This frequency cutoff leaves me with 1,231 verbs in total. Table 2 shows the eight most strongly associated verbs for each word order, along with their association strength. The association is measured in terms of odds ratios. One asterisk indicates statistical significance at the $\alpha = 0.05$ level according to Fisher's Exact Test, with two and three asterisks indicating significance at $\alpha = 0.01$ and $\alpha = 0.001$, respectively. All of the listed verbs are significant, due to the large sample size and the exclusion of low-frequency verbs.

We can observe that out of the 1,231 verbs, there are two verbs that occur exclusively in one of the two word orders. The table shows these verbs at the top of the list for the 1-2 order, the order that is also more frequent overall (64.8% of the clusters are in the 1-2 order). The odds ratios for these top collexemes indicate strong word order preferences for the other top verbs as

well — for example, the word that is ranked 8th for the 2-1 order, *gebruikmaken* ‘to make use’, is still 23.4 times more likely to occur in the 2-1 order than what would be expected by chance, if the word order was random. These clear effect sizes continue well down the ranking: the verb *spelen* ‘to play’ is ranked as the 300th collexeme of the 1-2 order, yet is still 1.61 times more likely to occur in this order.

Taking the list of all 1,231 verbs into account, I note that a large number of the verbs under consideration have statistically significant associations with either of the two word orders. Even though there are various problems with null-hypothesis significance tests of frequency-based measures in large corpora (Gries 2005), I will report a summary of these tests as an indication of the prevalence of lexical associations. However, I consider the computed association values listed in the table to be a more important result, because they express effect sizes rather than null-hypothesis probabilities. With the common significance threshold of $\alpha = 0.05$, 864 verbs (70.2%) are significantly associated with an order. With $\alpha = 0.01$, the number is 756 (61.4%), and with $\alpha = 0.001$, 631 verbs (51.3%) have significant associations.

It can be argued that the significance threshold of a collostructional analysis should involve a Bonferroni correction for performing multiple tests on the same data. I note that Gries (2005) observed larger than expected amounts of false positives when comparing word frequencies with χ^2 -tests in different subcorpora, though I am not aware of any such experiment on collostructional data. Because I am testing for lexical associations between unique pairs of words and orders, I view each lexical item as a separate thing to be tested, thereby not requiring a correction. Nevertheless, I will report the results with corrected thresholds for completeness: with $\alpha = 0.05$, the corrected value is $0.05/1231 = 0.0000406$. Under this condition, 531 verbs (43.22%) have significant associations.

5.2. Adjectivity factor

For verb clusters, there is one general semantic effect that we might expect to observe, as discussed by De Sutter (2005): that of the adjectival participial verbs mentioned in section 3. Because some words may serve either as adjectives or participial verbs depending on their semantic context, this property of adjectivity is not completely unambiguous, and it is also not annotated in the corpus. De Sutter lists 11 factors that may indicate the adjectivity of a participle though most of them cannot be taken from an annotated corpus. He discusses three ways to measure it in a corpus:

- Testing for usage with the auxiliary verb *zijn* ‘to be’, which is frequently used as a copula verb
- The ratio of adjectival use in the corpus: if the adjectival form is used more often, the meaning is more likely to be adjectival
- The ratio of participial use in the corpus (in all contexts, not just verb clusters)

I did not see the relevance of the third measure, so I have tested my lists of collexemes against the first two measures.

The first measure, usage with *zijn*, involves calculating for each main verb what I call the *zijn*-ratio: the percentage of clusters with this main verb and the auxiliary verb *zijn*. Out of the top 20 collexemes for the 2-1 order from the first analysis listed in Table 2 (only the top 8 are shown), 15 have a ratio above 0.5. For the 1-2 order, there are only 5. These counts indicate that there is some correlation with word order, though no absolute tendency. When I calculate the correlation with the 2-1 order associations (the values listed in Table 2) over all of the verbs I get a value of $c = 0.200$, a weak correlation. I repeated the procedure with *worden*, which may also take adjectival verbs, and found that $c = 0.140$, no correlation.

The second measure is difficult to use reliably in an automatic way, because there may be mistakes in the automatic annotation of the corpus. It involves measuring how often a participial verb form was used as an adjective as opposed to a verb in the entire corpus, but an adjectival use may have been mistagged as a verb, or the other way around. They are difficult to distinguish. Nevertheless, I calculated this adjective ratio for each verb. Out of the top 20 collexemes for the 2-1 order with a frequency over 1000, 11 have a ratio above 0.5. For the 2-1 order, there are 4. The correlation is again a weak one, with $c = 0.253$. The top adjectival verbs in 2-1 order also match the ones reported by De Sutter (2005).

This adjectivity factor should be controlled for if we want to observe lexical preferences. Even though it is a property of the verb, is arguably a more general feature or semantic class that can apply to many verbs, not a lexical preference specific to one verb. In section 5.4 I will discuss an analysis where adjectivity is controlled for.

5.3. Particle verb factor

One thing to note is that the top collexemes for the 1-2 order consist mostly of particle verbs. Like English, Dutch has a class of verbs that incorporates a particle, such as ‘wash **up**’, or *afwassen* in Dutch. These verbs have the syntactic property of being separable: the particle may appear in a different place in the sentence in some contexts. In Dutch, the particle and verb are written together when they are not separated. It has previously been observed that particle verbs are more strongly associated with the 1-2 order (Bloem et al. 2014), so it might be the case that the order preferences I observe are not associated with the individual verbs, but with the syntactic property of being separable. In Table 2, 7 out of 8 verbs in the 1-2 order list are particle verbs. In the 1-2 order list, one verb is separable (*gebruikmaken*, ‘to make use (of)’), though the separable unit is a noun, not a particle.

I also noticed that the relative word order frequencies of the verbs (how often a verb occurs in the 1-2 versus the 2-1 order) are not normally distributed. Figure 2 shows a frequency distribution of the relative word orders. We can observe two peaks in the distribution, one around 60-65% 1-2 orders and one at 80-85% 1-2 orders, meaning that there are relatively many verbs with these proportion of 1-2 orders. One would expect that there is just one peak, around the general average proportion of 1-2 orders, corresponding to a normal distribution. In that case, most verbs would follow the average input pattern, but some would have more specific associations with one order due to lexical preferences. The second peak indicates that there is a second factor involved here, which might be the particle verbs as a distinct class with a distinct word order pattern.

To test this possibility, I ran another distinctive collexeme analysis in which particle verbs

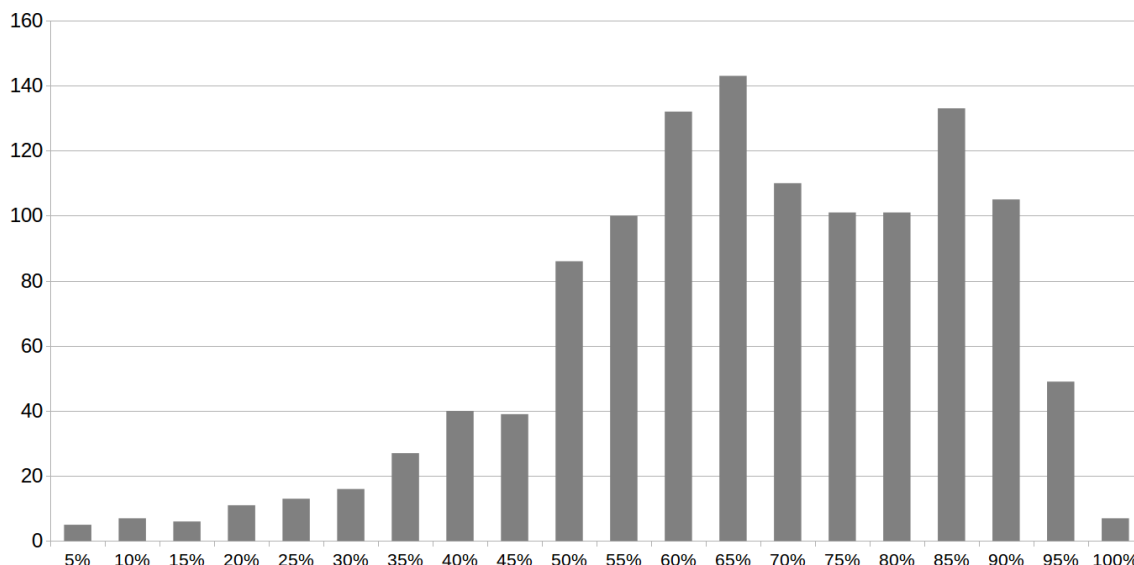


Figure 2: Frequency distribution of verb orders. The X-axis shows frequency bins of 1-2 order percentages, and the Y-axis shows how many verbs have this percentage of 1-2 orders. Values on the X-axis are the upper bound of the bin, i.e. 5% means verbs that have 0-5% 1-2 orders.

were excluded as much as possible. I filtered them out by excluding any verbs that start with any of a list of particles. This is not a foolproof method, as some verbs may start with something that looks like a particle, but at least it makes sure that anything that might look like a particle does not factor into the calculation. In this filtered analysis, there are 758 verbs. Out of those, 421 (59%) are significantly associated with either of the two word orders at $\alpha = 0.05$. With particle verbs, the percentage was 70.2%. This indicates that particle verbs can be linked to some of the observed word order preferences, but not nearly all of it. There are a large number of non-particle verbs that have an association with a particular order. Furthermore, the second peak in the word order distribution disappeared. In this analysis it is more like a normal distribution, with its peak at 65% 1-2 orders. This observation suggests that there are no longer two classes of verbs with a different mean word order preference.

5.4. Auxiliary verb factor

A further confound lies in the fact that different main verbs are used with different auxiliary verbs, which may affect word order associations. Previous work has found that the auxiliary verbs have distinct word order associations. For example, *hebben* ‘to have’ has a stronger 1-2 order association (Bloem et al. 2014). Verbs that only occur with *hebben* thus might have an association with the 1-2 order merely due to their association with *hebben*. In the previous analysis, that association would be attributed to the main verb.

To control for this auxiliary verb factor, I repeated the distinctive collexeme analysis, again excluding particle verbs, and using only frequencies from verb clusters with *hebben* as the

1-2 order			2-1 order	
Rank	Collexeme	Association	Collexeme	Association
1	hangen (55:4) 'to hang'	5.029647 ***	kampen (157:6) 'to struggle'	71.752328 ***
2	grijpen (60:5) 'to grab'	4.3885610 ***	zeggen (65:4) 'to say'	44.426110 ***
3	staan (1271:107) 'to stand'	4.1433800 ***	danken (400:44) 'to thank'	25.1134143 ***
4	verrichten (161:15) 'to execute/perform'	3.9102099 ***	lijden (142:32) 'to suffer'	12.156137 ***
5	verlenen (110:11) 'to grant'	3.6499705 ***	munten (55:24) 'to target'	6.2607829 ***
6	goedkeuren (47:5) 'to approve'	3.4391766 **	bedoeld (54:26) 'to intend'	5.6737062 ***
7	tonen (100:12) 'to show'	3.0424888 ***	bieden (137:69) 'to offer'	5.4343953 ***
8	hebben (1800:206) 'to have'	2.9718374 ***	begrijpen (33:23) 'to understand'	3.9170367 ***
9	stellen (277:35) 'to set, to state'	2.8690793 ***	bedriegen (49:35) 'to deceive/cheat'	3.8232165 ***
10	brengen (596:77) 'to bring'	2.7695827 ***	verboden (50:36) 'to forbid/ban'	3.7929261 ***
11	zitten (398:52) 'to sit'	2.7609043 ***	leven (274:203) 'to live'	3.7016798 ***
12	behoren (111:15) 'to belong'	2.7003267 ***	verraden (57:48) 'to betray'	3.2429590 ***
13	laten (150:21) 'to allow'	2.6023152 ***	vergeten (27:24) 'to err'	3.0706499 ***
14	roepen (48:7) 'to call/shout'	2.5084791 *	besteden (28:25) 'to spend'	3.0570449 ***
15	leveren (360:53) 'to deliver'	2.4532010 ***	dwingen (29:29) 'to force'	2.7293867 ***
16	verklaren (175:26) 'to declare/state'	2.4495074 ***	verkrachten (27:27) 'to rape'	2.729311 ***
17	verwekken (53:8) 'to beget'	2.4230324 *	regelen (30:32) 'to arrange'	2.5587384 ***
18	steken (97:15) 'to stab'	2.3608368 ***	slapen (34:38) 'to sleep'	2.4420665 ***
19	geven (1094:163) 'to give'	2.3488134 ***	liegen (52:59) 'to lie'	2.4060080 ***
20	nemen (750:116) 'to take'	2.2963172 ***	eten (74:86) 'to eat'	2.3494722 ***

Table 3: Top twenty collexemes of the 1-2 and 2-1 verb cluster orders with *hebben* 'to have', excluding particle verbs

auxiliary verb. This also controls for the adjectivity factor I discussed in section 5.2, because adjectival verbs are only used with *zijn* and *worden*, not with *hebben*. This leaves only the possibility that verbs that are used with *hebben* but also with another auxiliary verb had their lexical preferences influenced by the word order preference of that other auxiliary verb. However, this would still be a lexical preference, stored as a property of the main verb, and therefore does not need to be controlled for. These restrictions leave fewer verbs for the analysis, though. There are 218 verbs that occur more than 50 times in clusters matching these criteria (compared to 758 in the previous analysis), and their frequencies are lower than in the previous analysis.

Table 3 shows the 20 most strongly associated verbs for each word order, along with their association strength, in the same format as the previous results table. Again, there are many verbs for which the association values deviate from 1 — in both top 20s, all of the collexemes are over twice as likely to occur in the order they are associated with.

Out of the 218 verbs in this analysis, 112 (51%) are significantly associated with either of the two word orders at $\alpha = 0.05$. In the analysis where only particle verbs were filtered, the percentage was 59%. There is still a clear tendency for verbs to have significant lexical associations, even within the class of verb clusters with *hebben*. If a Bonferroni correction is applied, the corrected value is $0.05/218 = 0.000229$. Under this condition, 56 verbs (25.69%) have significant associations.

5.5. Additional semantic analysis

Even though I have now controlled for various factors in my computation of the association values, there are more general properties of the main verbs that might account for part of the associations, besides lexical preference.

In a collostructional analysis, the top collexemes for each construction are usually analyzed semantically. The verb cluster orders are generally considered to express no semantic difference and, therefore, I have no specific hypothesis on any possible semantic association of each word order to test. The claim of Pardoën (1991), who suggested a *STATIC* versus *DYNAMIC* interpretation distinction, should be tested, however. Because of this, I decided to take an objective approach by using semantic properties of Dutch verb senses from the Cornetto lexical-semantic database (Vossen et al. 2013), which includes *DYNAMIC* as a feature. Unfortunately, the Lassy Large corpus is not annotated for word senses so I had to take the properties of the most common senses of each verb, as judged by a native speaker. If there were multiple commonly used senses with conflicting properties, the verb was annotated for both properties.

I performed this annotation for the top 20 collexemes of each word order from the analysis in the previous section (shown in Table 3). The results are summarized in Table 4. It lists all of the semantic features from the Cornetto database, as well as the polarity property. I list the description of each feature below, taken from the user documentation (Maks et al. 2013):

- *VALENCY*: the number of arguments of the verb
- *TRANSITIVITY*: whether a verb can take direct objects, and how many
- *CONTROL*: the subject of the verb is capable of acting with volition

- **DYNAMIC**: the verb expresses a non-static, changing situation
- **ATTRIBUTIVE**: the verb expresses a relation of ownership
- **SPATIAL**: the verb expresses a location or movement of (one of) the participants
- **COGNITION**: the verb demands emotional, perceptual or mental activity
- **POLARITY**: what kind of attitude the word's most common senses express

The polarity property was annotated on the basis of an automatic sentiment analysis system, so there may be some errors in the annotation of this property in the database.⁸

Property	Values	1-2 order	2-1 order
VALENCY	Mono	0	4
	Di	10	12
	Tri	11	6
TRANSITIVITY	Intransitive	5	5
	Transitive	16	17
	Ditransitive	4	0
CONTROL	+CONTROL	16	15
	-CONTROL	6	6
DYNAMIC	+DYNAMIC	16	17
	-DYNAMIC	6	4
ATTRIBUTIVE	+ATTRIBUTIVE	1	0
	-ATTRIBUTIVE	19	20
SPATIAL	+SPATIAL	1	0
	-SPATIAL	19	20
COGNITION	+COGNITION	2	6
	-COGNITION	19	14
POLARITY	NEGATIVE	3	8
	POSITIVE	3	1

Table 4: Frequency count of semantic properties of the top 20 words in each word order

For each semantic feature, I counted how many of the top 20 collexemes for each order have this feature. The counts do not always add up to 20 because a verb may have multiple frequent senses with conflicting properties, in which case they were counted as having both features. We can observe that the counts for the 1-2 and 2-1 order are quite similar for most features, which indicates that there are no meaning associations of those features with a particular word order. Based on Pardoën's (1991) claim that 1-2 orders are assigned a dynamic interpretation while 2-1 orders get a stative interpretation, we might expect to see a difference in the number of verbs

⁸I noticed a few possible errors, for example, the first sense of *slapen* 'to sleep' is annotated as having negative polarity

with the DYNAMIC feature between the two orders. This is not the case however — there is one more DYNAMIC verb in the 2-1 order, a negligible difference.

One notable difference can be observed for the POLARITY property. By just looking at the verbs on the right-hand side of Table 3 one may already get the impression that many of the verbs associated with the 2-1 order have a somehow negative sense. The semantic analysis in Table 4 confirms this, listing 8 negative verbs for the top 2-1 order verbs, and just three for the top 1-2 order verbs. There are also slightly more positive 1-2 verbs. Another difference is in the COGNITION verbs, of which there are more for the 2-1 order. There was only one verb with the ATTRIBUTIVE feature and one with the SPATIAL feature in either of the top 20 collexeme lists, which were both in the 1-2 order list.

A further difference can be observed in the VALENCY and TRANSITIVITY counts: among the 2-1 order verbs we find four verbs that take only a subject argument while the 1-2 order list has no such verbs. On the other hand, the 1-2 order list features four ditransitive verbs while the 2-1 order list has none.

In summary, it appears that there are a few semantic features that are more common among verbs that are strongly associated with one order, compared to the other. The negative polarity property is the clearest of these differences, however, even there it is only a tendency — there are also many neutral verbs and even a positive verb among the top 1-2 order collexemes, and there are negative verbs among the top 2-1 collexemes as well.

Lastly, it is interesting that there are some similar verbs in the overall list of collexemes that are positioned adjacent to each other. An example of this can be seen in Table 3, where the bottom two verbs in the 1-2 order list are the verbs for ‘give’ and ‘take’. In the 2-1 order list the bottom verb is ‘eat’ with rank 20, and in the 21st position I found *drinken* ‘to drink’. It appears that at least some semantically similar verbs have very similar cluster order associations.

6. Discussion

I have calculated lexical associations between main verbs and their word order in clusters, and found that significant associations are quite common. A majority of the verbs are significantly associated with either the 1-2 or 2-1 word order in the corpus that I used, even after controlling for various confounding factors. I found that the two word orders exhibit some distinct semantic preferences, but not to such an extent that all of the associations can be explained in this way.

This result indicates that verb cluster word order associations are not only a result of general contextual factors and properties of the verb, but also of a lexical preference for one of the word orders, specific to a particular main verb. Such a preference has to be stored in the mental lexicon or constructicon in some way, whether it is as a connection or as a property of the verb. These lexical preferences contribute to the choice of a word order, along with other factors.

The results also suggests that the word orders are distinct linguistic units or constructions, rather than different realizations of an underlying unit. The word orders appear to have some semantic associations, with the 2-1 order showing an association with negative verbs and cognition verbs. Such associations can be explained if each word order is a construction with distinct semantic preferences. Furthermore, in a construction grammar view of language, the existence of semantic and lexical preferences would mean that the word orders are distinct constructions. If

a word order is chosen not purely on the basis of general features, but also due to an association with a specific verb, this must be stored as a preference of a word order construction for that verb, or a preference of the verb for a word order construction. Even though the orders do not exhibit a clear meaning difference, the semantic associations I found suggest that there may be differences in semantic preferences. A further possibility that this study did not explore is that there may be a functional difference between them, related to information structure or facilitation of linguistic processing.

However, I can not exclude the possibility that some of the variation my analysis attributed to specific verbs was actually caused by something else. I excluded various factors from the analysis, but there can always be other factors that I did not know about, and it is not possible to exclude every other possible factor. I have discussed most of the known factors from the literature. The only non-contextual factor from Bloem et al. (2014) I did not analyze or control for was the inference factor — a cluster occurring as part of a fixed expression. Controlling for this is difficult to do on the basis of verb types. It is not a direct property of the verb, although there are surely verbs that occur in fixed expressions more often than others. Still, I doubt this can account for all of the remaining lexical associations I observed. The fact that the distribution of the proportions had a single peak, after removing the particle verb factor, also indicates that there are no other major categories of verbs in the data with different order preferences. Smaller effects cannot be excluded in this way, however — I saw that the 2-1 order weakly correlates with adjectival semantics, even though there was no peak in the distribution to indicate this.

7. *Conclusions and future work*

With this study, I have shown that significant associations can be observed between lexical main verbs and the word order of Dutch two-verb clusters. After controlling for known factors affecting the word order variation, lexical and semantic associations with particular word orders were still observed, indicating that the two verb cluster word orders have distinct lexical and semantic preferences. These preferences may influence the speakers' choice between the two grammatical word orders, together with more general factors such as the contextual factors mentioned by Coussé et al. (2008).

While it is usually assumed that there is no meaning difference between the two word orders, my collostructional analysis did show some semantic associations between lexical main verbs and the word orders. This can be viewed as evidence for the hypothesis that the two word orders are distinct constructions. This was already suggested previously by the finding that the finding that the orders can be primed, i.e. a 1-2 order is more likely to be followed by another 1-2 order than by a 2-1 order (Hartsuiker & Westenberg 2000). Such priming is only possible if one order can be activated independently of the other order. It also seems difficult to conceive of a theory that allows for lexical preferences of a word order for particular lexical verbs to be stored in the lexicon, when the two word orders are not distinct linguistic units (i.e. constructions).

The fact that semantic associations with verb cluster word orders were observed also indicates that it can be interesting to apply collostructional analysis to a phenomenon with an unclear form-function mapping, even though the method was intended for studying semantic differences.

Perhaps more semantic associations of verb cluster constructions can be found if different or

additional semantic categories of verbs are used in the analysis. Finding such associations could allow for more of the observed word order variation to be explained in a more general way than as lexical preferences. Discovering these semantic generalizations would have to be done by hand by linguistic experts, as it was done by Stefanowitsch & Gries (2003) in their study, though they had far fewer items to analyze. Furthermore, no obvious semantic generalizations stood out to us, nor to some conference audiences that were presented with the top 10 collexemes for both orders. The meaning generalizations that Stefanowitsch & Gries (2003) attempted to identify in their collostructional analyzes were hypothesized in advance, while I have no idea what further meaning differences there might be between the two verb cluster orders, and none are hypothesized in the literature.

Alternatively, the semantic analysis could be done computationally using some measure of semantic similarity and comparing whether similar verbs are similarly ranked in the list of collexemes. I noted that there are some pairs of words with a high semantic similarity, as well as very similar word order preferences. A measure of semantic similarity would tell us whether this is the case for other semantically similar verbs as well. A crude measure of similarity that could be used is path similarity in a lexical-semantic database. If two verbs have similar hypernyms in such a database, their meaning is probably similar. A more accurate analysis could be performed using a distributional semantic network, which models the meaning of words on the basis of their collocations, an idea that is quite compatible with collostructional analysis. However, I am not aware of the existence of any distributional semantic network for the Dutch language.

When I did exclude factors, I did so by filtering out a large part of the total dataset of verb clusters. For example, I used only clusters with *hebben*, but this is a minority of the total number of verb clusters I extracted from the corpus. It might be possible to do a collostructional analysis with two open slots, one for the main verb and one for the auxiliary, but this would still divide the verb frequencies for each order into frequencies for each order for each auxiliary, leading to smaller samples of each category and less reliable results.

Lastly, it would also be interesting to investigate the nature of lexical preferences in Dutch verb clusters. Lexical preferences could be stored as properties of the words, or as connections between constructions. In future work, it might be possible to test whether a verb's association with the 1-2 order is also a link in the mental lexicon by experimentally testing whether that verb also primes the 1-2 order, or whether the 1-2 order primes it.

Acknowledgements

I would like to thank Arjen Versloot and Fred Weerman for their supervision of this work, and an anonymous reviewer for their comments. Some parts of this work were based on an earlier unpublished collostructional analysis study on reflexives. That study was joint work with Laura Deichmüller, whose contributions I am grateful for.

Jelke Bloem

Amsterdam Center for Language and Communication, University of Amsterdam

j.bloem@uva.nl

<http://www.uva.nl/profiel/j.bloem>

References

- Arfs, M. (2007). *Rood of groen? De interne woordvolgorde in tweeledige werkwoordelijke eindgroepen met een voltooid deelwoord en een hulpwerkwoord in bijzinnen*. Ph.D. thesis, Göteborg University.
- Bader, M., T. Schmid & J. Häussler (2009). Optionality in verb-cluster formation. Winkler, S. & S. Featherston (eds.), *The fruits of empirical linguistics*, De Gruyter Mouton, vol. 2, pp. 37–58, URL <http://ling.auf.net/lingbuzz/000751/current.pdf>.
- Barbiers, S. (2008). Werkwoordclusters en de grammatica van de rechterperiferie. *Nederlandse Taalkunde* 13:2, pp. 160–187, URL <http://depot.knaw.nl/9698/1/Barbiers-NTK.pdf>.
- Bloem, J., A. Versloot & F. Weerman (2014). Applying automatically parsed corpora to the study of language variation. Hajic, J. & J. Tsujii (eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin City University and Association for Computational Linguistics, Dublin, pp. 1974–1984, URL <http://www.aclweb.org/anthology/C14-1186>.
- Bresnan, J., A. Cueni, T. Nikitina & R. H. Baayen (2007). Predicting the dative alternation. Bouma, G., I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, KNAW, Amsterdam, pp. 69–94, URL <http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BresnanEtAL.pdf>.
- Colleman, T. (2009). Verb disposition in argument structure alternations: a corpus study of the dative alternation in Dutch. *Language Sciences* 31:5, pp. 593–611, URL <http://dx.doi.org/10.1016/j.langsci.2008.01.001>.
- Coussé, E. (2008). *Motivaties voor volgordevariatie. Een diachrone studie van werkwoordvolgorde in het Nederlands*. Ph.D. thesis, Ghent University, URL <http://dx.doi.org/1854/12253>.
- Coussé, E., M. Arfs & G. De Sutter (2008). Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep: een taalgebruiksgebaseerd perspectief op de synchronie en diachronie van de zgn. rode en groene woordvolgorde. Rawoens, G. (ed.), *Taal aan den lijve : het gebruik van corpora in taalkundig onderzoek en taalonderwijs*, Academia Press, pp. 29–47.
- De Sutter, G. (2005). *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. Ph.D. thesis, University of Leuven, URL <http://hdl.handle.net/1979/98>.
- De Sutter, G. (2009). Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. Dufter, A., J. Fleischer & G. Seiler (eds.), *Describing and Modeling Variation in Grammar*, Walter De Gruyter, pp. 225–255, URL <http://hdl.handle.net/1854/LU-537795>.
- Evers, A. (1975). *The transformational cycle in Dutch and German*, vol. 64. Indiana University Linguistics Club, Bloomington.
- Fillmore, C. J. (1985). Syntactic intrusions and the notion of grammatical construction. Niepokuj, M., M. VanClay, V. Nikoiforidou & D. Feder (eds.), *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, vol. 11, pp. 73–86, URL <http://dx.doi.org/10.3765/bls.v11i0.1913>.
- Goldberg, A. E. (2002). Construction grammar. *Encyclopedia of Cognitive Science*, Macmillan Reference Limited Nature Publishing Group.
- Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgariff. *Corpus Linguistics and Linguistic Theory* 1:2, pp. 277–294, URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.5694&rep=rep1&type=pdf>.
- Gries, S. T. & A. Stefanowitsch (2004). Extending collocation analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics* 9:1, pp. 97–129, URL http://www.linguistics.ucsb.edu/faculty/stgries/research/2004_STG-AS_ExtendingCollocations_IJCL.pdf.
- Haegeman, L. & H. van Riemsdijk (1986). Verb projection raising, scope, and the typology of rules affecting verbs. *Linguistic inquiry* 17:3, pp. 417–466, URL http://dbnl.nl/tekst/haeg001verb01_01/haeg001verb01_01.pdf.
- Hartsuiker, R. J. & C. Westenberg (2000). Word order priming in written and spoken sentence production. *Cognition* 75:2, pp. B27–B39, URL [http://dx.doi.org/10.1016/S0010-0277\(99\)00080-3](http://dx.doi.org/10.1016/S0010-0277(99)00080-3).
- Hilpert, M. (2008). *Germanic future constructions: A usage-based approach to language change*, vol. 7 of *Constructional approaches to language*. John Benjamins Publishing.

- Hindle, D. & M. Rooth (1993). Structural ambiguity and lexical relations. *Computational linguistics* 19:1, pp. 103–120, URL <http://www.aclweb.org/anthology/J93-1005>.
- Lehmann, H. M. & G. Schneider (2012). Syntactic variation and lexical preference in the dative-shift alternation. *Language and Computers* 75:1, pp. 65–75.
- Levshina, N. (2011). *Doe wat je niet laten kan: A usage-based analysis of Dutch causative constructions*. Ph.D. thesis, University of Leuven, URL <https://lirias.kuleuven.be/handle/123456789/317068>.
- Maks, I., H. van der Vliet, A. Görög & P. Vossen (2013). User documentation of Cornetto LMF. Tech. rep., VU University Amsterdam, URL <http://cornetto.clarin.inl.nl/help/D9-Cornetto-Documentation.pdf>.
- van Noord, G. (2009). Huge parsed corpora in LASSY. Van Eynde, F., A. Frank, K. De Smedt & G. van Noord (eds.), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, LOT, vol. 12, pp. 115–126, URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.323.8323&rep=rep1&type=pdf>.
- van Noord, G., G. Bouma, F. Van Eynde, D. de Kok, J. Linde, I. Schuurman, E. T. K. Sang & V. Vandeghinste (2013). Large scale syntactic annotation of written Dutch: Lassy. Spyns, P. & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, Springer Berlin, pp. 147–164, URL http://dx.doi.org/10.1007/978-3-642-30910-6_9.
- Pardoen, J. (1991). De interpretatie van zinnen met de rode en de groene volgorde. *Forum der letteren* 32:1, pp. 1–22, URL http://www.dbnl.org/tekst/_for004199101_01/_for004199101_01_0001.php.
- Pauwels, A. (1953). *De plaats van hulpwerkwoord, verleden deelwoord en infinitief in de Nederlandse bijzin*. M. & L. Symons, Leuven.
- Rebuschat, P. & J. Williams (2012). Introduction: Statistical learning and language acquisition. Rebuschat, P. & J. Williams (eds.), *Statistical learning and language acquisition*, Walter de Gruyter. Berlin, vol. 1, pp. 1–12, URL http://eprints.lancs.ac.uk/65357/1/Rebuschat_Williams_2012_Intro_libre.pdf.
- Speelman, D., J. Tummers & D. Geeraerts (2009). Lexical patterning in a construction grammar: The effect of lexical co-occurrence patterns on the inflectional variation in dutch attributive adjectives. *Constructions and Frames* 1:1, pp. 87–118, URL <http://dx.doi.org/10.1075/cf.1.1.05spe>.
- Stefanowitsch, A. & S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8:2, pp. 209–243, URL <http://dx.doi.org/10.1075/ijcl.8.2.03ste>.
- Stroop, J. (1970). Systeem in gesproken werkwoordsgroepen. *Taal en tongval* 22, pp. 128–147, URL http://dbnl.nl/tekst/stro008syst01_01/stro008syst01_01.pdf.
- Vossen, P., I. Maks, R. Segers, H. Van Der Vliet, M.-F. Moens, K. Hofmann, E. T. K. Sang & M. de Rijke (2013). Cornetto: a combinatorial lexical semantic database for Dutch. Spyns, P. & J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, Springer, pp. 165–184, URL http://dx.doi.org/10.1007/978-3-642-30910-6_10.
- Wurmbrand, S. (2006). Verb clusters, verb raising, and restructuring. Everaert, M. & H. van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Blackwell Publishing, pp. 229–343, URL <http://dx.doi.org/10.1002/9780470996591.ch75>.
- Zwart, J.-W. (1996). Verb clusters in continental West Germanic dialects. Black, J. R. & V. Motapanyane (eds.), *Microparametric syntax and dialect variation*, John Benjamins Publishing, vol. 139 of *Current Issues in Linguistic Theory*, pp. 229–258, URL <http://dx.doi.org/10.1075/cilt.139.12zwa>.