Processing verb clusters

Bloem, J.

Publication date
2021

Document Version
Final published version

License
Other

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
In theory, linguists’ lives are becoming easier due to the increasing availability of digital texts that can serve as data for linguistic research. The internet is full of language data and its volume is increasing at a faster rate than any linguist could keep up with. Yet this data is only used to a limited extent, because it is not always clear how we can find, among this sea of data, the utterances and constructions that are of particular interest to us linguists. When we do try to do this, it results in such a large amount of data that any further processing will have to be automatic, using models that can generalize over large volumes of data.

This dissertation describes the word order variation that exists in Dutch two-verb clusters, such as *gezien hebben* ‘seen have’ or *hebben gezien* ‘have seen’ by applying quantitative research methods to larger text collections. The results show that the processing complexity of the utterance plays an important role in the choice between these two word orders. The dissertation consists of eight articles covering historical and synchronic aspects of verb cluster order variation, with a focus on new, computational research methods that can be applied to the problem. Chapters two, three and six mainly concern the use of digital tools for linguistic research, while the chapters four, five, seven, eight and nine apply such tools to research questions concerning verb cluster variation.
Processing verb clusters
Processing verb clusters

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen
op vrijdag 12 maart 2021, te 16.00 uur

door

Jelke Bloem

geboren te Groningen
### Promotiecommissie

**Promotores:**
- prof. dr. A.P. Versloot (Universiteit van Amsterdam)
- prof. dr. F.P. Weerman (Universiteit van Amsterdam)

**Overige leden:**
- prof. dr. H.J. Bennis (Universiteit van Amsterdam)
- prof. dr. P.C. Hengeveld (Universiteit van Amsterdam)
- prof. dr. J.E. Rispens (Universiteit van Amsterdam)
- prof. dr. J.E.J.M. Odijk (Universiteit Utrecht)
- prof. dr. A.P.J. van den Bosch (Universiteit van Amsterdam)
- dr. F. Van de Velde (KU Leuven)

Faculteit der Geesteswetenschappen

---

Het hier beschreven onderzoek werd mede mogelijk gemaakt door steun van de Fryske Akademy.
## Contents

Acknowledgements .............................................. xi
Author contributions ........................................... xv

1 Introduction .................................................. 1
  1.1 Verb clusters .............................................. 2
  1.2 New methods .............................................. 5
    1.2.1 Automatically annotated corpora ..................... 5
    1.2.2 Agent-based models ................................... 7
    1.2.3 Language modeling ................................... 7
  1.3 Topics in variation and change ......................... 8
    1.3.1 Processing ........................................... 8
    1.3.2 Variation affected by change ......................... 8
    1.3.3 Language contact .................................... 9
  1.4 Outline .................................................. 9

2 Applying automatically parsed corpora to the study of language variation 11
  2.1 Introduction .............................................. 12
  2.2 Verbal cluster variation .................................. 14
  2.3 Multifactorial modeling of language variation .......... 18
    2.3.1 Using automatically annotated data .................. 19
    2.3.2 Verbal cluster variation ............................. 20
  2.4 Method and data ......................................... 21
  2.5 Results ................................................ 24
    2.5.1 Europarl corpus .................................... 28
  2.6 Discussion .............................................. 28
    2.6.1 Linguistic interpretation ............................ 29
    2.6.2 Future work ........................................ 30
### 3 Evaluating Automatically Annotated Treebanks for Linguistic Research

3.1 Introduction ................................................. 34
3.1.1 Automatically Annotated Treebanks ....................... 35
3.1.2 Construction-specific Querying .......................... 36
3.2 Linguistic Studies using Automatically Annotated Treebanks 36
3.3 Current Approaches to Evaluation ........................... 38
3.4 Linguistically Informed Evaluation .......................... 40
3.4.1 Manual Evaluation of the Results ....................... 40
3.4.2 Manual Evaluation of the Text .......................... 41
3.4.3 Fall Back to Simpler Annotation ......................... 42
3.4.4 Search for Particular Instances ......................... 43
3.5 Conclusion .................................................. 45

### 4 Verbal cluster order and processing complexity

4.1 Introduction .................................................. 48
4.2 Related work .................................................. 51
4.2.1 Word order variation ..................................... 51
4.2.2 Processing complexity .................................... 52
4.2.3 Verb clusters ............................................. 54
4.2.4 Verb cluster order variation ............................. 58
4.3 The processing hypothesis ................................... 59
4.4 Method and data ............................................. 62
4.5 Measures of processing complexity ......................... 64
4.6 Results ....................................................... 78
4.7 Discussion .................................................... 84
4.8 Conclusion .................................................... 89

### 5 Testing the Processing Hypothesis of word order variation using a probabilistic language model

5.1 Introduction .................................................... 92
5.2 The case of Dutch two-verb clusters ......................... 94
5.3 Processing complexity ....................................... 95
5.3.1 Theoretical models ..................................... 95
5.3.2 Empirical measures ..................................... 97
5.4 Data ......................................................... 99
5.5 Language model .............................................. 99
5.6 Results ....................................................... 100
5.7 Discussion .................................................... 104

### 6 Lexical effects on Dutch verbal cluster order

6.1 Introduction .................................................... 108
6.1.1 Previous work ........................................... 110
6.2 Lexical and semantic associations .......................... 112
6.2.1 Associations with related constructions: Lexical preferences 113
6.2.2 Semantic differences between the orders 114
6.3 Measuring lexical associations 115
   6.3.1 Collostructional analysis 115
   6.3.2 Gain Ratio 117
   6.3.3 Residuals of logistic regression 118
6.4 Verb clusters 118
6.5 Method and data 121
   6.5.1 Method of analysis 121
   6.5.2 Corpus data 123
6.6 Results 125
   6.6.1 Initial distinctive collexeme analysis 125
   6.6.2 Adjectivity factor 127
   6.6.3 Particle verb factor 128
   6.6.4 Auxiliary verb factor 130
   6.6.5 Semantic analysis 130
   6.6.6 Gain Ratio 134
   6.6.7 Residuals of logistic regression 135
6.7 Discussion 137
6.8 Conclusions and future work 138

7 An agent-based model of a historical word order change 141
   7.1 Introduction 142
   7.2 Verbal clusters 143
   7.3 Methodology 144
   7.4 Results 148
      7.4.1 Dutch 150
   7.5 Discussion 152

8 Learned borrowing or contact-induced change: Verb cluster word order in Early-Modern Frisian 155
   8.1 Introduction 156
   8.2 Verb clusters in Frisian 158
      8.2.1 Modern Frisian 158
      8.2.2 Historical varieties of Frisian 160
      8.2.3 ‘Mixed Frisian’ texts and language contact 162
      8.2.4 Language contact, acquisition and change 163
      8.2.5 Verb clusters in ‘Mixed Frisian’ 165
      8.2.6 Hypotheses 168
   8.3 Verb clusters in Dutch 169
      8.3.1 Factors of variation 169
      8.3.2 Processing verb cluster orders 171
   8.4 Data and methodology 173
      8.4.1 Texts 173
      8.4.2 Annotation 175
      8.4.3 Automatic extraction 176
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>237</td>
</tr>
<tr>
<td>Summary</td>
<td>259</td>
</tr>
<tr>
<td>Samenvatting</td>
<td>267</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>275</td>
</tr>
</tbody>
</table>
Who would have thought I would finish this book in a Dutch department? Certainly not my Dutch teachers in high school, who I rarely seemed to get along with. Typically, my colleagues use this space to write about their lifelong passion for linguistics, but I went off to study computer things instead after that experience. During those studies, though, I realized that language is a very interesting type of data for computers to process, and that there is more to linguistics than memorizing arcane Dutch part-of-speech terminology that even people in the Dutch linguistics department don’t seem to use. There are lots of people who helped bring about this change of mind, supported me on the way to becoming a PhD student, and/or helped me during the project. I cannot possibly mention everyone, but I want to thank you all anyway and will mention a few specifics below.

First of all, I would like to thank Fred Weerman and Arjen Versloot, who not only supervised this PhD project, but also came up with the initial idea before I was involved. Even though I did not know what diachronic linguistics was during the interview, I am grateful that you had the confidence that I could make it work and you have taught me many things about the world of linguistics since. Arjen, your valuable insights and interest in doing things in a new way have motivated me to keep improving my research skills and to keep reading efficiently and critically. Thank you for your enthusiasm, and also for introducing me to many people in the world of Frisian linguistics. Fred, I am grateful that you gave me great freedom in this project to really make it my own. Your attention to detail and critical questions have taught me much about writing and have tremendously improved (as well as lengthened) the contents of this work. Your role in making the Dutch linguistics department a fun and social place to work must also be mentioned, for a freshly arrived student from Groningen this made everything much easier. I must also thank Olga Fischer here, for helping to keep us all on track during my progress interviews, as well as John Nerboune, for letting me know about the vacancy at the University of Amsterdam that resulted in this book.
I also wish to thank my co-authors for chapter 9, Marieke Olthof and Maud Westendorp, without whom this very interesting piece of the puzzle would not have existed. Chapter 6 uses a method that I learned about in a student project in Groningen together with Laura Deichmüller, and I thank her for letting me use her code contributions in the chapter 6 study.

Various people have graciously donated their time to improve the contents of this work by reviewing or proofreading sections or supporting me in the use of new methods and tools. Besides the numerous reviewers for the various conferences and journals these chapters have been submitted to, I want to specifically thank Gosse Bouma, Timothy Colleman and Jeannette Schaeffer for providing extensive and constructive reviews of chapters. For proofreading and feedback, I am grateful to Enoch Aboh, Jan Hulstijn, Caitlin Meyer, Marjolein Poortvliet, Margot Kraaijamp, Margreet van Koert, Maja Ćurčić, Camelia Bleotu and Odelia Ahdout. Furthermore, I owe thanks to Paul Boersma for helping me with some statistical questions, to Daniël de Kok and Gertjan van Noord for their help in getting me started with the Lassy Large corpus and associated tools, and to Jelle Zuidema and Rens Bod for early advice on the project, and allowing me to access the ILLC’s computing facilities.

Of course, I would like to express my appreciation to Hans Bennis, Kees Hengeveld, Judith Rispens, Jan Odijk, Antal van den Bosch and Freek Van de Velde for agreeing to be part of my doctoral committee and evaluating this work, quite possibly during their Christmas break.

I am grateful to all my colleagues at the department of Dutch Linguistics. Thanks to my senior (former) colleagues, with whom I had many hallway or door-way conversations that helped me along in scientific or educational endeavours: Suzanne Aalberse, Sible Andringa, Jan Don, Lotte Hogeweg, Folkert Kuiken, Tom Lentz, Sander Lestrade and Josje Verhagen. I want to specifically thank Hedde Zeijistra for welcoming me to the department on the first day, showing me around and answering my questions, making my start in the department much easier. I also thank the ACLC’s Grammar & Cognition group for their beneficial feedback on my practice presentations.

Importantly, my gratitude also goes out to Lilian, Marten and Marion, who have helped to answer many practical questions and solve so many problems, most of which I was not even aware of. I am thankful to Cees Lourens and Silvain Smit from ICTS, who were willing to support someone crazy enough to want to use Linux dual-boot systems, run servers and store large datasets at a Faculty of Humanities, for the work they put into my requests.

Many hours were spent with Margreet and Aafke, who were my officemates for over seven years, and I have appreciated our conversations, ranging from totally mundane to academic to important life advice. Just as many chats were had with my office neighbour Margot, as with our matching schedules, you were always around when a break was needed. Thanks! Caitlin, my running mate, supervisor ‘sibling’, fellow verb cluster aficionado and paranymph! While we started at the same time, you were a ‘local’ so I benefited tremendously from your insider knowledge of pretty much all the things I just listed. Your
verb cluster literature collection gave me a head start, but to no one’s surprise except your own, you did defend first! Thanks for always being there on this journey. Maja, thanks for your insights and friendship, and for helping me to find university housing and become Waddendijk neighbours!

While I have been around too long in the P.C. Hoofthuis to thank everyone who came and went during that time, I wish to thank, in no particular order, several other colleagues who have helped to make it such a fun and social environment to work in: Jing, Marjolein, Matthias, Sanne, Brechje, Camille, Janneke, Hernán, Natalia, Sybren, Dana, Patrick, Lois and Lorena. Sadly, I think I will be the final person to make this acknowledgement, which is to thank the former Bungehuis colleagues and PhDs, for many shared ACLC borrels, PhD lunches and events. Of course that includes those from Linguistics who arrived after the Bungehuis!

A special mention for my colleagues of the Fryske Akademy: Nika, Evelyn and Anne. While we did not end up meeting very often, when we did, we had a great time! I also wish to express my appreciation for my fellow verb cluster enthusiasts, Lotte and Liesbeth. We were writing dissertations on verb clusters at the same time, but in very different ways. This led us to organize a workshop on verb clusters together, which I really enjoyed. I also thank the participants of that workshop for their insightful discussions on the topic, which have contributed to this work. I also acknowledge the contributions of Freck and the QLVL crew, as well as Leen and Morgane in Leuven, as they welcomed me, extensively discussed my work and introduced me to the Belgian ways of lunches, coffee breaks and op café gaan.

Thanks also to my friends less involved with the work, who supported and advised me on my move to Amsterdam and/or helped to make my time in Amsterdam, at conferences and at the LOT schools more fun, including some already mentioned above, Mariya, Sibel, Rui, Anne, Tao, Angelina, Sveta, Bea, Elsan, Noushin, Eliza, Robert, Mira, the ILLC basketball crew including Milos (thanks for enlivening my birthday party with an interdisciplinary ACLC-ILLC Chomsky debate!), Jo, Ke and Raquel, the VU linguists — Fernanda, Yao, Siwei & co, and Het Bureau — Ruby, Mirella, Edoardo & co.

Finally, I am very grateful to my family, some of whom turned out to share my interest in linguistics — Sophie and Anne-Mieke, who I also thank for being my paranymph. Willeke, thanks for hosting me for that conference in Japan, and for always being a reliable source on Japanese. And of course my non-linguist parents Jaap and Dittie, for your support, for helping me move my stuff all over the place, and for making me do enough of my homework that I ended up here.
Author contributions

Chapter 1
This chapter was authored by Jelke Bloem.

Chapter 2

Bloem, Versloot and Weerman posed the research question on the nature of verb cluster order variation and whether it can be addressed by computational modeling. In order to answer this research question, Bloem conducted a corpus study, which was a replication study based on De Sutter (2009). Bloem collected the corpus data, wrote the necessary code, and extracted the features. The results were statistically analysed by Bloem. Bloem wrote the first version of the text. This was discussed during several supervision meetings. On the basis of valuable feedback from Versloot and Weerman, Bloem rewrote and revised this manuscript into its published form and then extended it with information scrapped due to the page limit for its chapter form.

Chapter 3
This chapter is adapted from:

Chapter 4


Bloem, Versloot and Weerman posed the research question on the generalization that underlies the factors affecting verb cluster word order variation that we saw in Chapter 2. In order to answer this research question, Bloem conducted a corpus study, which built upon the one in Chapter 2. Bloem collected the corpus data, wrote the necessary code, and extracted the features. The results were statistically analysed by Bloem. Bloem wrote the first version of the text. This was discussed during many supervision meetings. On the basis of valuable feedback from Versloot and Weerman, Bloem rewrote and revised this manuscript into its published form and then extended it with information scrapped due to the page limit for its chapter form.

Chapter 5


Chapter 6


Chapter 7

Bloem, Versloot and Weerman posed the research question on the diverging development of verb cluster word order in the West-Germanic languages. In order to answer this research question, Versloot developed a simulation and collected initialization values for it from corpus data, that Bloem further developed into an agent-based model. Bloem wrote the code for the agent-based model, ran the simulations, performed the analysis and created the visualizations. Bloem wrote the first version of the text. This was discussed during several supervision meetings. On the basis of valuable feedback from Versloot and Weerman, Bloem rewrote and revised this manuscript into its published form and then extended it with information scrapped due to the page limit for its chapter form.

Chapter 8


Bloem, Versloot and Weerman posed the research question on verb cluster order in historical Frisian texts and possible contact effects. In order to answer this research question, Bloem conducted a corpus study of Early-Modern Frisian texts. Bloem processed the Early-Modern Frisian corpus data, wrote the necessary code, and extracted the features. The results were statistically analysed by Bloem. Versloot manually evaluated exemplars extracted from the corpus, provided access to the corpus, conducted a corpus study of the Basle Wedding Speeches, and wrote a first version of Section 8.2.5 on this. Bloem wrote the first version of the other sections. This was discussed during several supervision meetings. On the basis of valuable feedback from Versloot and Weerman, Bloem rewrote and revised this manuscript into its published forms and then adapted the material into its chapter form.

Chapter 9

This chapter is adapted from: Olthof, Marieke, Westendorp, Maud, Bloem, Jelke, & Weerman, Fred (2017). Synchronic variation and diachronic change in Dutch two-verb clusters. Tijdschrift voor Nederlandse Taal- en Letterkunde, 133(1), 34–60.

Olthof, Westendorp, Bloem and Weerman posed the research question on diachronic aspects of current synchronic verb cluster order variation. Data from the Corpus of Spoken Dutch was collected by Olthof and Westendorp, who also
wrote the first version of the text, which was reviewed and discussed by Bloem and Weerman. The results were statistically analysed by Bloem, and the text was further developed into the published version by Bloem, as discussed by Olthof, Westendorp, Bloem and Weerman during several meetings and based on their written feedback.

Chapter 10

This chapter was authored by Jelke Bloem.
CHAPTER 1

Introduction

This dissertation discusses various aspects of the word order variation found within verbal clusters in West-Germanic languages, particularly Dutch and Frisian. The interesting thing about this variation is that its function is not clear. Typically, when we say something different we mean something different. With these verb clusters, however, there is no clear motivation for using one word order over the other. Example 1.1, taken from an article on chess in a corpus of Wikipedia texts, contains two-verb clusters in both possible orders:

(1.1) *Indien het aangevallen stuk niet gedekt1 wordt1 ... terwijl tegelijkertijd het schaak wordt2 opgeheven2 ..., volgt stukverlies* simultaneously the check is resolved follows capture

‘If the attacked piece is not covered when the check is resolved, capture follows.’ [wik_part0049/25627-3-1]1

The two verb clusters are highlighted in bold. Both clusters involve the auxiliary verb *worden* ‘to be’, and a participial verb. The auxiliary verb is numbered with 1 and the participial verb is numbered with 2, following their syntactic structure (the auxiliary verb 1 is the syntactic head of the participle 2, and therefore above 2 in the syntactic structure of this sentence). We can

---

1 Many of the example sentences in this dissertation are taken from the Dutch Lassy Large corpus (van Noord, 2009). These identifiers refer to the unique sentence identifiers in this corpus, and the reader can use them to find the sentence in the corpus, for example using PaQu (Odijk et al., 2017). This particular example can be retrieved based on its identifier using the following PaQu query: https://paqu.let.rug.nl:8068/xpath?db=lassywiki&xpath=%2FAlpino_ds%2Fsentence%5B@sentid%3D%2225627-3-1%22%5D&mt=std&xn=20.
thus say that the first cluster, *gedekt wordt*, is in the 2-1 order, while the second cluster, *wordt opgeheven*, is in the 1-2 order.\(^2\)

As both orders are possible, we can flip the orders and construct the following sentence in Example 1.2:

\[(1.2) \quad \text{Indien het aangevallen stuk niet wordt gedekt, terwijl tegelijkertijd het schaak opgeheven wordt, volgt stukverlies.}\]

‘If the attacked piece is not covered when the check is resolved, capture follows.’

Yet the meaning remains the same as that of Example 1.1, and from the perspective of a native speaker there is no other clear functional difference between the sentences (such as a focus difference).

Why, then, do we find two opposite verb cluster orders in this single Wikipedia sentence? Is it random, or are there patterns in the way these two orders are used, that might help us to explain why they are both in use? If so, can novel computational and quantitative methods help us find these patterns and account for them? And how does the history of the language play a role in the availability of two word order options? My dissertation focuses on these and other questions centering around verb cluster order variation, and new methods that can be used to study these questions.

The outline of this chapter is as follows. Section 1.1 describes what verb clusters are. Section 1.2 describes novel methods that can be used to find patterns in verb cluster usage and obtain evidence regarding hypotheses on when each order is used. Section 1.3 introduces some topics in variation and change that are relevant to verb cluster order variation. Section 1.4 presents my research questions and gives an overview of the outline of this dissertation.

### 1.1 Verb clusters

Verb clusters are constructions in which multiple verbs group together at the second of two fixed positions that are typically distinguished in descriptive accounts of Dutch sentence structure. In the Dutch grammar *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al. 1997), these fixed positions are called (verbal) poles (or *Klammern* in the German tradition). In a prototypical main clause, the subject comes before the first position, while the finite verb occupies the first position. This is also known as verb-second (V2). If there are any other verbs they are positioned on the second position, while dependent constituents of these verbs may occur before or after the second position. In a prototypical subordinate clause, the first position is reserved for a subordinator.

---

\(^2\)This numbering convention can extend to larger verb clusters and will be used throughout this dissertation, as it is commonly used in the verb cluster literature.
Introduction

(such as a relative pronoun) while all of the verbs are placed on the second position, creating a verb cluster. These verb clusters that may appear at the second position are notable in that they exhibit a lot of order variation: within languages, between closely related West-Germanic language varieties in which they occur, and even within speakers.

Apart from Dutch, verbs group together in all other Germanic languages in some way or another. In the syntactic literature, particular attention has been paid to Afrikaans, Frisian, German, Swiss German, West Flemish, and related dialects, where these verbal groups are considered ‘verb clusters’. This distinction between ‘verb clusters’ and other groups of verbs is made on the basis of several syntactic properties, of which the main one is that the linear order of the verbs in the actual utterance, and the order of selection of the complement verbs in the syntactic structure, do not always match in these languages (Wurmbrand, 2006). Verbal heads of verb phrases may select their complements to the left and to the right in the linear order. There are no rigid head-initial or head-final ordering patterns in verb clusters. Example 1.1 in the introduction is a demonstration of this: there are two verb clusters with the auxiliary verb worden, which can be considered the syntactic head, and it selects a main verb in both cases (gedekt and opgeheven). In the first cluster this main verb comes before the head (the 2-1 order, complement-head), and in the second cluster it comes after the head (the 1-2 order, head-complement).

Verb clusters consist of a main verb and one or more clustering verbs, which are typically auxiliary verbs, but can be any verb that selects a verbal complement. Example 1.1 contains two clusters of two verbs, the auxiliary verb worden ‘to be’, selecting a participial main verb as its dependent. Depending on the clustering verb used, the main verb may also be infinitival, such as in Example 1.3:

(1.3) ... zodat hij de Jedi kan uitroeien

‘... so that he can exterminate the Jedi.’ [wik_part0013/5954-13-2]

Kunnen ‘can’ is a modal verb, and modal verbs select infinitival main verbs. More types of clustering verbs are possible: an overview is provided by Augustinus (2015).

As evidenced by the fact that both possible linear orders are used in example 1.1, the syntactic order of selection and the linear order of the verbs are not always the same. This extends to larger verb clusters as well: clusters of three or more verbs are possible, and various word orders such as 1-2-3 (had kunnen zien, ‘had can see’) and 1-3-2 (had zien kunnen) have been attested throughout the Dutch-speaking area (Barbiers et al., 2006).

The examples I provided so far have been of verb clusters in subordinate clauses, where all verbs in the clause are located at the second position. Verb clusters can occur in main clauses as well, but for this to occur the main clause needs three or more verbs. The finite verb is positioned on the first position in
1.1. Verb clusters

the main clause (V2), requiring two additional verbs in the clause in order to form a verb cluster of the remaining two verbs on the second position (second pole). Therefore, subordinate clause verb clusters provide the clearest cases of free variation, with few syntactic and semantic restrictions. They have also been studied the most in previous work.

Verb clusters in the 1-2 and 2-1 orders have been studied extensively since the 1950s and before, and the role of many factors in their use has been discussed. The foundational study of Pauwels (1953) highlighted the role and the extent of regional variation, showing large differences in preferences for the orders throughout the Dutch-speaking area. The coloured dialect maps produced by this study were the origin of the terms ‘red order’ for the 1-2 order and ‘green order’ for the 2-1 order, terms that are still used in the field of Dutch linguistics. Other regional and stylistic factors have been discussed: the 2-1 order might be avoided because it is used in less prestigious language varieties (Sassen, 1963) or because it may have been considered an undesirable Germanism (van den Berg, 1949), as this order is used in High German. The mode of communication (spoken or written) also plays a role. Stroop (2009) shows that in spoken Dutch, the 2-1 order is preferred over the 1-2 order in participial two-verb clusters by contrasting findings from the Corpus of Spoken Dutch with findings by Arfs (2007a) in written Dutch. De Sutter (2005) proposes the amount of interaction in a discourse and the amount of editing as further stylistic factors that may affect verb cluster order.

It has been argued that speakers may change the order of verbs to match the standard stress pattern of Dutch. In a two-verb cluster, the auxiliary verb is unstressed and the main verb is stressed. If there is a stressed syllable directly before the cluster, speakers might choose to place the main verb at the end of the cluster (a 1-2 order) to avoid having two adjacent stressed syllables (De Sutter et al., 2007).

While the two orders are in free variation, semantic factors are mentioned to play a role. Different types of auxiliary verbs are associated with different word order preferences (De Sutter, 2005) and there is a study that argues that there is an interpretation difference between the two orders — in Pardoen’s (1991) view, 1-2 orders are assigned a dynamic interpretation, while 2-1 orders are assigned a stative interpretation.

Besides this literature discussing motivations for the choice between grammatical word orders, there is also a large body of literature that aims to account for the extent of the possible variation through syntactic accounts and derivations. Starting with Evers (1975), who proposes a verb raising rule to account for the syntactic characteristics of verb clusters, various other proposals have been worked out within transformational theories of grammar. This includes derivations with a head-initial instead of head-final basic order (Zwart, 1996), and with a mono-clausal instead of a biclausal origin (Bennis & Hoekstra, 1989). Verb clusters have also been accounted for in monostratal theories of grammar, either as generalized raising with some additional features to account for word order variation (Bouma & van Noord, 1998), or in a framework where
complement raising and subject raising are distinct (Augustinus, 2015).

Verb clusters continue to be a fruitful source of PhD dissertations. Recent dissertations have all highlighted different aspects of verb cluster order variation: as a case of intra-constituent variation involving many factors (De Sutter, 2005), as a consequence of rhythmic structure (Arfs, 2007b), as a case of language change either for the purpose of focus marking (Coussé, 2008) or as a consequence of semantic broadening and syntactic extension (Coupé, 2015), as an instance of complement raising (Augustinus, 2015), and as a case of geographical syntactic variation (Hendriks, 2018). The general picture that emerges is that there are many dimensions and sources of variation, and it is unlikely that a single factor will be found that can account for all the variation.

This highlights the importance of covering as much ground as possible: if so many aspects of language affect this variation, we should be thorough in studying anything that may influence language variation in terms of possible effects on verb cluster order variation. Therefore, the general theme of this dissertation is to explore additional potential sources of variation that have so far been understudied in the literature. Because verb clusters have already been studied quite thoroughly, I make use of new methods that have not been widely used in linguistics yet to study these potential sources of variation, while also allowing me to obtain new kinds of evidence on questions of verb cluster order variation.

1.2 New methods

Advances in language technology and computational methods have made new research methods possible, which can be used by linguists to obtain new kinds of evidence, construct new kinds of models and collect more data to study linguistic phenomena. In this dissertation, I aim to use such new methods whenever possible. This section briefly introduces a few new methods that are not yet widely used in linguistic research, but that I have used in this dissertation.

1.2.1 Automatically annotated corpora

The first new method is the use of large, automatically annotated corpora. These are text corpora that have been enriched with linguistic information (such as syntactic structure, parts-of-speech or coreference chains) automatically, by a natural language processing tool. This linguistic information can then be used to search for particular constructions, to calculate statistics over them, or to perform other quantitative research tasks. These corpora are increasingly becoming available, but have not been used for much linguistic research yet. For Dutch, there is the 700 million word Lassy Large treebank (van Noord et al., 2013), which I use in several chapters of this dissertation.
Van Noord & Bouma (2009) discuss automatic syntactic parsing for the creation of such corpora. 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’. 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 (van Noord, 2009). 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.

The advantage of automatically annotated corpora is that they are much larger than manually annotated corpora, as automatic annotation is far less costly than having humans provide annotations. The disadvantage is that natural language processing tools typically make more mistakes than well-trained human annotators. Nevertheless, the larger volume of data that these corpora contain can provide quantitative data on rare linguistic phenomena and constructions, or subtle probabilistic effects, that couldn’t otherwise be obtained. This makes it possible to study, for example, verb clusters with uncommon clustering verbs, very large verb clusters (Bloem, 2020), or cases where many factors affect the probability of one order or the other being used, as is the case for verb clusters.

Because automatic annotation makes mistakes, I first address the question of reliability. Can we replicate previous quantitative work on verb cluster order variation, which made use of manually annotated data, using only automatically annotated data? This question is addressed in chapter 2.

These corpora can be used to reveal new patterns in verb cluster order variation. Probabilistic patterns can be studied in more detail: smaller effect sizes require more data to prove. If a certain factor, such as the length of the sentence, makes a particular word order slightly more probable, only a large-scale corpus study can potentially reveal this. What factors can be found to affect verb cluster order variation in written Dutch, based on evidence from a state-of-the-art automatically annotated corpus? This question will be addressed in chapter 4.

Even if automatically annotated data provides the same results as manually annotated data, this does not help us judge the reliability of the results in cases where no previous work or manually annotated data is available. While the general parsing accuracy of automatically annotated corpora is typically known, this accuracy is only an average over all parsing steps, including simple things such as connecting determiners to their nouns. Are there ways in which linguists can estimate the reliability of an automatically annotated corpus for the study of a particular construction, such as verb clusters? This question is addressed in chapter 3.

A further advantage of large, automatically annotated corpora is that rare patterns can be studied: if a particular word in a verb cluster only occurs once in every 2000 verb clusters, a large corpus is necessary to obtain enough instances
of this type of verb cluster. This is relevant, for example, when studying the main verbs in a cluster: any Dutch verb can be the main verb of a verb cluster. Is there any evidence that main verbs, or the semantics thereof, affect verb cluster word order choice? This question is addressed in chapter 6.

1.2.2 Agent-based models

The second new methodology is the use of agent-based models. Most commonly, when language is modeled, we create a model of the language of a particular, prototypical speaker. Such a model cannot represent factors of interspeaker variation, such as regional variation or language contact, even though such factors often play a role in language variation. Agent-based models do allow the researcher to incorporate such factors by modeling multiple speakers (agents), as well as their (means of) interaction. These models can then be used to study language change in a simulated population. Within the domain of linguistics, agent-based models are typically used to model the emergence of new structures on the basis of their function in interaction, using artificial languages. An early overview of such work is provided by Steels (1997). However, these models could also be used to model known instances of language change in existing languages, based on historical corpus data. Possible causes of the change can be input into the model to generate new hypotheses about what factor might have caused the change in reality. In the case of Germanic verb clusters, we know that the 1-2 order became the dominant order in some languages (i.e. English), and the 2-1 order became the dominant order in other languages (i.e. German). This can be modeled by modeling verb clusters for a population of agents, including the Proto-West-Germanic word order preferences and known changes in verb cluster contexts around this time. What factors could have caused these related languages to develop such diverging verb cluster orders? This question is addressed in chapter 7.

1.2.3 Language modeling

The third new method is the use of probabilistic models of language. These models, commonly employed in natural language processing tasks, can learn the probabilities of combinations of words (or other linguistic units), and use those, for example, to predict what the next word in a discourse might be. Evidence suggests that prediction also plays a role in human language processing: we probabilistically predict upcoming language input at various levels of representation (Kuperberg & Jaeger, 2016). Surprisal-based theories of language processing hypothesize that elements that are more difficult to predict, are more difficult to process (Hale, 2001). Probabilistic models of language can be used in the context of these theories to measure the difficulty of processing particular constructions. In the context of verb clusters, these models can provide evidence for ease or difficulty of processing verb cluster orders or verb cluster contexts, assuming that surprisal-based theories of language processing are correct. Which
two-verb cluster order is easier to process? This question is addressed in chapter 5.

1.3 Topics in variation and change

As mentioned, verb cluster order variation is an alternation between linguistic forms that can potentially be affected by all sorts of sources of variation in language. My dissertation covers several possible sources of variation that have not been fully explored in relation to verb clusters.

1.3.1 Processing

Advances in cognitive science and the increasing availability of psycholinguistic evidence has highlighted the role of performance aspects of human cognition in shaping our grammar and language. MacDonald (2013) presents a range of phenomena in which cognitive constraints and processing difficulty are shown to affect language production, including word order. Minimizing processing effort has been proposed as an explanation for optionality phenomena and cases of grammatical variation, such as for English particle movement (Gries, 2001). De Sutter (2005) brings up processing as a possible explanation for verb cluster order variation: some of the factors that are found to correlate with verb cluster order variation, can be explained in terms of processing mechanisms. However, this psycholinguistic interpretation is not explored further with follow-up experiments, even though it could potentially account for a large part of the interspeaker variation in verb cluster orders. Do either of the orders occur in contexts that are generally more difficult to process, and which order is more difficult to process? I address these questions in chapters 4 and 5.

1.3.2 Variation affected by change

While the history of Dutch verb cluster order variation has been mapped by previous corpus studies (Coussé, 2008; Coupé, 2015), there are still open questions on either end of the time scale. On the mediaeval end, it is unclear why verb clusters might have developed from Proto-West-Germanic with diverging word orders. In English, verbal word order converged on the 1-2 order, while in German, the 2-1 order took over in nearly all contexts. In Proto-West-Germanic, there appears to have been variation (Coussé, 2008). Why did the orders diverge? This question is addressed in chapter 7.

On the modern end of the timescale, it is unclear whether the processes of language change described by Coussé (2008) and Coupé (2015) are still ongoing. If this is the case, the present day variation must at least partly be explained as a consequence of language change, with two competing forms available in the grammar. While it is not feasible to track modern language change in real-time, the apparent-time paradigm makes it possible to gain some insight
into language change by comparing younger and older speakers. Furthermore, a study of spoken language can reveal the most current trends, as written language is typically more conservative. Do younger speakers use more 1-2 orders than older speakers in their spoken Dutch? This question is addressed in chapter 9.

1.3.3 Language contact

The matter of (historical) language contact has received hardly any attention in verb cluster studies, even though language contact can clearly cause and affect variation. An example of this can be seen in West-Frisian, where speakers appear to have started using both possible word orders in two-verb clusters fairly recently, even though use of the 1-2 order is nonnormative. As all speakers of West-Frisian are bilingual and also speak Dutch, this change is assumed to be influenced by language contact with Dutch (de Haan, 1996). However, this is not the whole story, as the Frisian language has an extensive history of language contact, and this 1-2 order also appears in historical texts (Larooij, 1991; Hoekstra, 2012). This raises questions regarding the extent and effect of this historical language contact. Is the modern use of this word order actually a new development taken from Dutch, a consequence of hundreds of years of language contact, or a continuation of an older language-internal development? And what kinds of language contact between Dutch and Frisian may have affected this change? These questions are addressed in chapter 8.

1.4 Outline

The rest of the dissertation investigates the topics touched upon in this introduction in more detail. The core of my dissertation consists of eight chapters that address the topics and methods introduced above. These chapters were originally written as stand-alone research papers that have been published in computational linguistics conference proceedings or linguistics journals. The chapters can therefore be read independently of each other. While each chapter has a different focus, they partially overlap in the topics they introduce and cover. Specifically, the dissertation consists of the following chapters:

In chapter 2, I argue for the use of automatically parsed corpora to study language variation by replicating key findings of De Sutter’s (2005) study using data from a larger, automatically annotated corpus. Despite some limitations imposed by the automatic approach, the conclusions of De Sutter are confirmed, and the automatic approach allows for relatively easy extension of the study to additional verb cluster constructions and an additional domain of text.

Chapter 3 addresses the problem of errors made by automatic annotation tools in studies such as the one in chapter 2, and how their impact on the results of a linguistic study can be minimized. This is done by proposing various tricks for calculating construction-specific annotation accuracy scores that may help a linguist studying a particular construction.
Chapter 4 delves further into the choice between possible verb cluster orders in two-verb clusters, investigating the factor of processing complexity as a way to account for the involvement of a wide range of factors, from sentence length to priming effects, in this choice. By testing for associations with factors indicating a higher or lower processing complexity of the verb cluster and its context, we find strong indications that the 1-2 order is easier to process than the 2-1 order.

Chapter 5 operationalizes the factor of processing complexity in a different way: as an effect of surprisal. Assuming that more surprising elements are more difficult to process, we apply a probabilistic language model to compute surprisal over a large set of verb clusters from a corpus. The results show that uniform information density, or the difference in surprisal between the verbs, differs between the two possible verb cluster orders, providing further evidence for the idea that the 1-2 order is easier to process.

Chapter 6 addresses a semantic aspect of verb cluster word order that can only be quantitatively studied using very large corpora: the effect of the main verb and its semantic properties. There are many possible main verbs that can be used in a verb cluster, so it is difficult to find enough evidence to draw generalizations unless the corpus is very large. Using the method of distinctive collexeme analysis, we find main verbs with statistically significant preferences for either of the two orders, and draw some tentative semantic generalizations among these verbs.

Chapter 7 introduces agent-based modeling as a tool for historical linguistics by modeling the historical development of verbal cluster word order in Germanic languages. The model is used to generate two hypotheses regarding the development of the 1-2 order in English and the 2-1 order in German, on the basis of historical data from Old English, Old High German and Old Frisian.

Chapter 8 addresses the extent to which language contact may have shaped Frisian verb cluster order, particularly in Early-Modern Frisian. Prescriptively, it is said that the 2-1 order should be used in Frisian, but the 1-2 order shows up in modern usage as well as in historical texts. This chapter discusses whether this can be attributed to language contact, and to what kind: contact through bilingualism, or through learned borrowing.

Chapter 9 investigates the relation between diachronic change and synchronic variation in the word order of Dutch two-verb clusters. Variation and change often go hand-in-hand, and the results of this chapter provide evidence that two-verb cluster order preferences are still shifting in spoken language, showing that diachronic change is another factor in the synchronic variation of verb cluster orders that we observe.

Chapter 10 summarizes the findings of this dissertation, discusses their implications and provides suggestions for future research.
Applying automatically parsed corpora to the study of language variation

Chapter Highlights

Problem Statement

• Quantitative linguistic studies have shown that word order variation is affected by many different factors. Accurately estimating the effect of each of these factors requires large amounts of text.

Research Questions

• Can automatically extracted data from large, automatically parsed corpora be used to replicate the findings of a quantitative study of word order variation that was conducted on a manually annotated corpora with manually selected data?

• Previous studies have found that various factors play a role in Dutch two-verb cluster word order. Do these findings hold when these factors are studied jointly in a multifactorial model, based on corpus data?

2.1 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 (2.1) or as a prepositional dative (2.2):

(2.1) He gave his friend the ticket.

(2.2) 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 multifactorial 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 multifactorial 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 demonstrate the benefits of using automatically annotated corpora for the study of language variation by replicating a previous, manual multifactorial 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. This extension of the study allows us to generalize...
the claims of the previous study to standard Dutch two-verb cluster variation more broadly. 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 the Dutch syntactic literature (Evers, 1975; den Besten & Edmondson, 1983; Haegeman & 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, 2007b). In addition, a methodologically sound quantitative study, which did not make use of automatically annotated data, already exists for comparison (De Sutter, 2009).

Dutch verb clusters allow for a lot of optionality. The verbal clusters found at the ends of Dutch clauses allow for almost free word 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:

\begin{align}
\text{(2.3) } & \text{Ik \textit{denk} \textit{dat} \textit{ik \textit{het} \textit{begrepen} \textit{heb}} \\
& \text{I think that I it understood have} \\
& \text{‘I think that I have understood it.’}
\end{align}

\begin{align}
\text{(2.4) } & \text{Ik \textit{denk} \textit{dat} \textit{ik \textit{het} \textit{begrepen}} } \\
& \text{I think that I it have understood}
\end{align}

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 multifactorial 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 multifactorial 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 found in de Sutter’s smaller model are still present. We also compare our expanded model 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 verb cluster types.

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 verb cluster in terms of their syntactic structure in the annotation format used by the corpus.

In section 2.2 we briefly discuss Dutch verbal clusters and the variation found in them. We then discuss previous work on modeling of Dutch verbal
2.2 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 2.3 is called the 2-1 order and construction 2.4 is called the 1-2 order. 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, letting 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 2.3 and 2.4 show two-verb clusters with auxiliary verbs as heads, selecting a participial main verb. Following De Sutter (2009) we consider this type of cluster 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 (*wollen* "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. For example, in two-verb clusters, modal auxiliaries select infinitival main verbs, rather than participial ones. In Dutch we can observe that the 1-2 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 2-1 order only 0.5% of the time. This is evidence that they are used, though rarely.

**Clusters with other head verbs** There are lexical verbs such as *staan* ‘to stand’ and *helpen* ‘to help’ that can also form a verbal cluster in certain contexts. This list includes causal verbs, aspectual verbs, and lexical verbs that
are harder to classify. These verbs seem too diverse to be grouped together. For brevity and due to their relative rarity (together they make up about 5.5% of the clusters) we did not include these constructions in the present study, though it would certainly be possible to explore them in a large corpus study and perhaps contribute to their classification. For a literature overview on these verbs, and a corpus study investigating what verbs can be clustering verbs, see Augustinus (2015).

Te-infinitival clusters  In the cluster types we discussed so far the auxiliary verb was finite, but there is another verb cluster construction in which both the auxiliary and the main verb are infinite, and the main verb is marked by the infinitival marker te "to". 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.

Verb particles in clusters  We just mentioned that verbal clusters cannot be interrupted, but there are exceptions to this. The most notable type of cluster interruption, also known as ‘cluster creeper’ in the literature, is also common in standard Dutch. Particle verbs may occur in a cluster in their split form, and there seem to be few restrictions to where the particle can be placed in the verbal cluster.

In the descending order, the particle can be placed at the start of the cluster. This is shown for the verb terechtkomen in 2.5, where terecht is the separable particle. Alternatively, the particle can be placed together with the main verb, as in 2.6:

(2.5)  
Hoe het daar terecht is gekomen is onzeker .
How it there up is ended is uncertain .
‘How it ended up there is unclear.’ [wik_part0048/24676-2-2]

(2.6)  
Hoe het daar is terechtgekomen is onzeker .
How it there is ended-up is uncertain .
‘How it ended up there is unclear.’

In the 2-1 order, the particle is always at the start of the cluster, since the main verb comes first and the particle can only come before the main verb. There are some instances in written texts where the main verb is preceded directly by a separated particle but we consider this to be an alternate spelling, not a different construction. Normally, the separable verb is written as a whole when there are no intervening elements.

There are other forms of cluster interruptions besides interruptions by particles, but they are rarely accepted by standard Dutch speakers. In Flemish dialects, they are more commonly attested and accepted (Hendriks, 2013)
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:

(2.7) \textit{De fuut kan in alle wateren van enig formaat aangetroffen.}

\par The grebe can in all waters of some size found be.

‘The grebe can be found in all bodies of water of substantial size.’

(2.8) \textit{Wegwerpbatterijen kunnen niet worden opgeladen.}

\par 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. From this, we conclude that main clause clusters can 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. Generative 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 & 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, 2007b). 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).
Contextual factor  There are two different factors that Coussé et al. (2008) consider contextual, the regional background of the speaker and the mode of communication. Firstly, the Dutch dialect continuum contains varying distributions of 1-2 and 2-1 orders, and the dialect of a speaker’s native region may affect their standard language. As a general trend, there is more of a preference for the 1-2 order in the areas where the dialect is close to standard Dutch. While the regional factor is important, it does not explain intraspeaker variation. Secondly, it has been shown that speakers prefer the 2-1 order in spoken language and other highly interactive modes of communication. There might be a prescriptive element as well — edited writing is less likely to use the 2-1 order.

Rhythmic factor  It has been argued that speakers may change the order of verbs to match the standard stress pattern of Dutch. In a two-verb cluster, the auxiliary verb is unstressed and the main verb is stressed. If there is a stressed syllable directly before the cluster, speakers might choose to place the main verb at the end of the cluster (a 1-2 order) to avoid having two adjacent stressed syllables. However, De Sutter (2009) did not find a strong effect of stressed syllables near the cluster in written texts.

Semantic factor  Semantics, particular lexical semantics, may also play a role in the choice of order. One finding in this area is that adjectival participles have a preference for the 2-1 order. Apart from that, not much work has been done in this area, perhaps due to the small corpora that have been used so far. To study lexical semantics, many instances of clusters with particular words are required. It has also been argued that 1-2 and 2-1 orders have their own interpretations. Pardoen (1991) claims that the 2-1 order goes with a stative interpretation, and the 1-2 order with a dynamic interpretation.

Discourse factor  It has been found that syntactic persistence may influence verbal cluster order. When a speaker has used an 1-2 order, this construction is primed, and they are more likely to use the 1-2 order again for the next verbal cluster.

From this overview, 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 multifactorial modeling technique used by De Sutter (2005) fits this criterion.

We will now discuss some related studies where multifactorial 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 multifactorial study on verbal cluster order that we are replicating.
2.3 Multifactorial 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. Statistical methods such as regression analysis or ANOVA, which both allow more than one independent variable, are used for this. Starting with Gries (2001), this methodology has been applied to the study of language variation from corpus data. A well-known example is the dative alternation study of Bresnan et al. (2007). In these studies, multifactorial 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 accounts for 84% of the variation using a multifactorial 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.

For this reason, the study used natural language data from the British National Corpus. Variables were manually annotated and analyzed with a generalized linear model (GLM), a generalization of linear regression. Furthermore, Classification and Regression Trees (a type of decision tree classifier) were used to verify the results. While this model is additive and therefore can only show the importance of each variable in addition to the previous one (in terms of information gain from the split the variable defines), it makes no distributional assumptions. To explain his results, which show many different variables affecting the choice of construction, Gries proposes a processing hypothesis — in linguistic contexts that are more difficult to process (such as long sentences), speakers are more likely to choose the construction that requires the least processing effort.

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.

2.3.1 Using automatically annotated data

Corpora that have been syntactically 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 the syntactic formalism used by the parser — 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 many 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 & 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. Firstly, the finding of previously unknown (uses of) constructions. Large corpora are more likely to contain rare linguistic phenomena. Secondly, more accurate frequency counts can be obtained, not just of words but also of constructions. Lastly, correlations between words and constructions can be found more easily (i.e. two different reflexives and their verbs).

For the Dutch language, the potential uses of the 500 million word automatically annotated LASSY Large corpus have been discussed by van Noord (2009). This work 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
2.3. Multifactorial modeling of language variation

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 case of the dative alternation, discussed in the introduction (examples 2.1, 2.2). Lehmann & 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:

(2.9)  \textit{Anna probeerde} (\textit{om}) een bom onschadelijk te maken
\textit{Anna tried} (\textit{CMP}) a \textit{bom harmless} to make

‘Anna tried to defuse a bomb.’

Bouma (2017b) provides further examples, and creates a model of this variation using data from part of the LASSY Large corpus. This model is multifactorial, 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 \textit{proberen} in example 2.9) 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 \textit{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.

2.3.2 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 English in De Sutter (2009). In this study, various variables from previous work on verbal clusters were modeled using a multifactorial model. Like in Gries (2001), and unlike in Bouma (2017b), 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
Applying automatically parsed corpora to the study of language variation

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 clusters 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 2.5.

2.4 Method and data

We create a multifactorial 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 common 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). Identifiers displayed after the translation of example sentences in this chapter are the identifiers of
the sentence in the Lassy Large corpus. 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 written 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 cannot 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, for example, in the sentences with verb clusters included in the Syntactic Atlas of Dutch dialects (DynaSAND, Barbiers et al. 2006). 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 do 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 2.5.1.

The verbal clusters were extracted from the corpus using XPath 2.0 queries via the DACT command line tools (de Kok, 2010). 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 using the DACT Python bindings, and further processing (i.e. counting constituent lengths) with custom Python scripts.

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 resulting 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 Table 2.1. To operationalize the Type of the auxiliary verb, De Sutter divided the auxiliary verbs up into five grammatical classes: zijn "to be" as a copular 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 show very similar order patterns in preliminary checks.

The variable **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 can 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: **highly** informational (nouns, verbs, numerals), **intermediately** 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 form in the entire Lassy Large corpus. Lastly, we added two variables to distinguish the new cluster types discussed in section 2.2 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 (De 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 in the present study, 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).
2.5 Results

Table 2.1 shows the comparison of our model to that 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 $\beta$ 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 cannot have an effect compared to itself, and 1 is the neutral value.

As a first observation for the results in table 2.1, 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. 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
<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Odds ratio De Sutter (2009)</th>
<th>Odds ratio This study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of auxiliary</strong></td>
<td>Copular zijn</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Auxiliary of time</td>
<td>18.30 ***</td>
<td>1.90 ***</td>
</tr>
<tr>
<td></td>
<td>Passive zijn</td>
<td>7.82 ***</td>
<td>2.19 ***</td>
</tr>
<tr>
<td></td>
<td><em>zijn</em></td>
<td>11.73 ***</td>
<td>1.19 ***</td>
</tr>
<tr>
<td></td>
<td><em>worden</em></td>
<td></td>
<td>2.19 ***</td>
</tr>
<tr>
<td></td>
<td><em>hebben</em></td>
<td></td>
<td>132.42 ***</td>
</tr>
<tr>
<td></td>
<td>Modal verb</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Morphological structure of the main verb</strong></td>
<td>Non-separable</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Separable</td>
<td>3.87 ***</td>
<td>4.92 ***</td>
</tr>
<tr>
<td><strong>Length of middle field</strong></td>
<td>0-2 words</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3-5 words</td>
<td>2.03 ***</td>
<td>2.42 ***</td>
</tr>
<tr>
<td></td>
<td>6-8 words</td>
<td>2.29 ***</td>
<td>3.23 ***</td>
</tr>
<tr>
<td></td>
<td>9-11 words</td>
<td>2.29 ***</td>
<td>3.34 ***</td>
</tr>
<tr>
<td></td>
<td>12-14 words</td>
<td>2.57 **</td>
<td>3.33 ***</td>
</tr>
<tr>
<td></td>
<td>&gt;14 words</td>
<td>1.98</td>
<td>3.15 ***</td>
</tr>
<tr>
<td><strong>Information value</strong></td>
<td>Low</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>1.41</td>
<td>1.21 ***</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>1.94 ***</td>
<td>1.11 ***</td>
</tr>
<tr>
<td><strong>Inherence</strong></td>
<td>No fixed expression</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Fixed expression</td>
<td>2.26 ***</td>
<td>2.10 ***</td>
</tr>
<tr>
<td><strong>Grammatical relation of extraposition to head</strong></td>
<td>Adjunct/no extraposition</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>CMP of main verb</td>
<td>0.47 ***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CMP of preverbal head</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No extraposition</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>CMP/ADJ of main verb</td>
<td></td>
<td>51.44</td>
</tr>
<tr>
<td></td>
<td>CMP/ADJ of preverbal head</td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td><strong>Main verb frequency</strong></td>
<td>$\beta = 2.44^{E-06}$</td>
<td>$\beta = 3.73^{E-08}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$ASE = 7.74^{E-07**}$</td>
<td>$ASE = 1.05^{E-08***}$</td>
<td></td>
</tr>
<tr>
<td><strong>Finiteness of the head</strong></td>
<td>Finite head</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Infinite head</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td><strong>Context clause type</strong></td>
<td>Subordinate clause</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Main clause</td>
<td>0.34</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Comparison of the size of the effect of the variables on verbal cluster variation for the two studies. CMP = complement, ADJ = adjunct.
**2.5. Results**

**Extrapolation 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 (2005, p. 205-230). 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, and we included more data. These values do indicate that the models are good enough for prediction tasks. They are also similar to the c-score Bouna (2017b) 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 2.1 are the main measure of that.

De Sutter also reports that he was unable to include interaction effects to avoid multicollinearity. We make the same observation for the variables used by de Sutter, despite our larger sample size. Without interaction terms, all the Variable Inflation Factors (VIFs) (indicating multicollinearity) are very low ($< 2$). However, when we try to add the interaction term that improves the model the most (largest deviance reduction using stepwise additive modeling), which is an interaction between **Length of middle field** and **Information value**, the VIF of **Information value** shoots up to 18 million, indicating that this interaction basically explains the same variation as the variable **Information value** does by itself. The next best interactions show VIFs
of 23.55 and 41.20, while a VIF over 10 is indicative of serious multicollinearity. While it is not conclusive proof, this observation also indicates that the chosen variables are quite independent from each other, and do not explain the same parts of the variation.

Interactions with the different construction types (main clause clusters and infinitival clusters) are a different story. We find strong interactions particularly between the different auxiliary verbs and main clause/subordinate clause verbal clusters. This is a logical consequence of the fact that these two types of clusters are different constructions, in which different auxiliary verbs are used.

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, thus only including subordinate clause clusters with participial main verbs. 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.

We also created separate models including only main clause clusters or only infinitival clusters to check whether the same variables hold for these constructions.

For the main clause model, we simply excluded all data with a value of ‘subordinate clause’ for the *Context clause type* variable, allowing us to see the effect of each variable for main clause clusters only. For reasons of space, we cannot list the entire results table, but will summarize the differences. As indicated by the interactions we just discussed, one of the differences between the main clause model and the full model can be found for the *Type of auxiliary*. The auxiliary *worden* appears to be more strongly associated with the 1-2 order compared to the reference *zijn*. *Hebben* shows a similar effect size as in the full model with subordinate clause clusters included.

Higher values for *Length of middle field* appear to be less strongly associated with the red order. However, this might just be because the reference value is 0-2 words, which is almost impossible for a main clause. The longer categories (12-14 words and > 14 words) do trend slightly towards the 1-2 order. The unexpected effects found for *Grammatical relation of extraposition to head* are also present in the main clause only model. In fact, they are somewhat larger than in the full model (357.74 instead of 49.65 for comp/adj. of main verb, 0.39 instead of 0.46 for comp/adj. of preverbal head), note that odds ratio values are exponentially scaled from 1. For clusters with preverbal constituents with intermediate *Information value*, we also see a clear effect of 1.51 here, over 1.21 in the full model.

While we cannot point to specific explanations for each of these differences,
we can say that all of these variables, as originally identified by de Sutter for subordinate clause verbal clusters, also apply to variation in main clause verbal clusters. The model for te-infinitival clusters only similarly did not differ much from the full model.

### 2.5.1 Europarl corpus

In section 2.4 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 and/or use of auxiliary verbs 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). For this particular corpus, translation effects can be another explanation: much of the contents of the corpus were probably translated from English, where verb groups are always in the 1-2 order. 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.

### 2.6 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 automatically annotated corpora are an excellent source for obtaining more data for language variation studies. Although some variables (e.g. 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 multifactorial 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 enables the retrieval of 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.

2.6.1 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 & Weerman, 2016). Meyer and Weerman 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 & Weerman, 2016). 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. Our 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 & Weerman (2016) 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.

2.6.2 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 (2017b)
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.

Lastly, it is clear that further work needs to be done in explaining why these variables affect verbal cluster order. It appears that the processing hypothesis is not the entire story. An explanation of the processing difficulty of each of the variables that can convincingly point to either one of the two orders as the easier or more basic one, would be very interesting. Otherwise, more evidence may have to be found, either in the form of additional processing-related variables, or in the form of a different kind of evidence entirely, such as a language production task that controls for processing difficulty.
CHAPTER 3

Evaluating Automatically Annotated Treebanks for Linguistic Research*

Chapter Highlights

Problem Statement

- Automatically annotated corpora have many potential benefits as a data source for linguistic research, but it is not clear how to evaluate the quality of the annotation for this purpose.

- While treebanks are usually evaluated with a general measure over all the data, linguistic studies often focus on a particular construction or a group of structures, and we would like to estimate annotation quality over all instances of a particular construction.

3.1 Introduction

This chapter addresses an issue that is important for the application of large, automatically annotated corpora to linguistic research. A disadvantage of corpus-based methods in linguistics is that many phenomena of interest to theoretical linguists are used infrequently in naturalistic speech, and therefore are less likely to occur in smaller corpora. Automatically annotated linguistic resources contain the ‘big data’ that is necessary to study these phenomena empirically, but they inevitably contain errors as well, due to the imperfect natural language processing tools that were used to annotate them. As linguists are increasingly using large corpora as a source of empirical evidence, these issues have been acknowledged, but not explored systematically. In this work, I discuss four different approaches to evaluating data from automatically parsed corpora when a particular linguistic phenomenon is being studied. Current methods for evaluating the quality of annotation are too general for the purposes of studying a particular construction, and only measure the overall accuracy of the annotation of a corpus. While I use treebanks to illustrate the approaches because they seem to be the most common type of automatically annotated language resource, the approaches are also applicable to language resources created using other forms of automatic annotation, such as part-of-speech tagged historical texts or semantically parsed corpora.
3.1.1 Automatically Annotated Treebanks

Treebanks are text corpora that have been enriched with syntax trees or syntactic graphs (e.g. when dependency grammars are used), allowing linguists to search those texts for particular syntactic constructions and morphological features. Such queries will result in a list of only those constructions, which is much easier to study than an entire text. Due to advances in natural language processing, it has become possible to syntactically parse large amounts of text automatically, with fairly good accuracy. This has resulted in the creation of treebanks that are much larger than traditional, manually annotated corpora. From a linguistic perspective, the main advantage of these large-scale treebanks over manually annotated corpora is that they can be used to investigate rare constructions, co-occurrence patterns of uncommon words or small probabilistic effects. They also provide larger sample sizes or lists of examples of naturalistic data for more common linguistic phenomena.

Probabilistic effects in language have been discussed particularly in the study of alternations, i.e. multiple near-synonymous constructions that form two grammatical options for expressing the same meaning. Corpus studies of such phenomena have revealed that a number of factors from various domains of language (i.e. phonetics, semantics, pragmatics) may affect the choice between alternative constructions in such an alternation to varying degrees. The size of these effects can be interpreted as probabilities. The linguistic implications of the observation of such effects have been discussed by Bresnan (2007), who studied the English dative alternation. This multifactorial study experimentally tested whether probabilistic effects found in a previous corpus study corresponded to speaker preferences in a rating experiment between the two constructions of the dative alternation, i.e. ‘I gave her the book’ and ‘I gave the book to her’. An earlier example of a multifactorial study of a linguistic optionality can be found in Gries (2001), who shows that many factors affect particle placement in English. In this study, Gries also computes effect sizes to quantify how influential the factors are, though the term ‘probability’ is not explicitly mentioned.

The main disadvantage of automatically annotated corpora is the error rate. While manually annotated or manually checked treebanks contain some errors, automatic annotation comes at the cost of annotation accuracy. The errors made by the parser will include systematic errors, where the parser has more difficulty with certain types of constructions than others. The parser may even be unable to annotate a particular construction correctly, thereby failing to provide the necessary means to search for it in a large corpus. Therefore, when using automatically annotated treebanks for linguistic study, some sort of evaluation is necessary to make sure that the construction of interest was annotated correctly, or at least well enough for the purposes of the study. While the accuracy rate of the parser that was used to annotate the treebank is usually known, such a general measure of evaluation is not meaningful for most linguistic studies.
3.2. Linguistic Studies using Automatically Annotated Treebanks

3.1.2 Construction-specific Querying

When studying a particular linguistic phenomenon or construction in a corpus, it may be more relevant to view the task as a form of information retrieval — all of the sentences instantiating the construction have to be retrieved from a larger data source (the corpus). The main difference with most information retrieval tasks is that the success of the search process depends on the quality of the annotation, rather than the quality of the search algorithm or the query. Nevertheless, I will use two basic measures from information retrieval, precision and recall, to illustrate the various approaches to evaluating the annotation quality of a corpus for a particular construction. In the context of this chapter, precision is defined as the fraction of results from a corpus query that are instances of the construction that is being searched for, while recall is defined as the fraction of instances of the construction in the corpus that are retrieved. I will assume that corpus queries are perfectly written following the annotation format of the corpus being queried to specify exactly what the researcher wants to retrieve. Under this assumption, any imperfections in precision and recall occur only due to incorrect annotation. In reality, other issues may affect precision and recall as well, such as inaccurate formulation of the query or a lack of distinction between certain phenomena in the annotation format of the corpus. Since those would be problems of information retrieval rather than automatic annotation quality, I will not focus on them in this discussion.

3.2 Linguistic Studies using Automatically Annotated Treebanks

Automatically annotated treebanks are a useful source of information in any study where large sample sizes are beneficial. Such treebanks have been made available for various languages. For Dutch, there is the 700 million word Lassy Large treebank (van Noord et al., 2013). For German, the 200 million word TüPP-D/Z (Müller, 2004) is available with automatic annotation. For English, the Google Books n-gram corpus (Lin et al., 2012) has been annotated syntactically, as well as the 4 billion word Gigaword v5 corpus (Napoles et al., 2012), to name but a few. Treebanks for a specific domain or language can be created as long as an automatic parser is available. This is the case for many major languages. Efforts have been made to make this technology more accessible to linguists who do not necessarily have a technical background, using techniques like example-based querying for treebanks (Augustinus et al., 2012; Odijk et al., 2018), or systems where researchers can upload their own corpora to be automatically annotated, such as PaQu (Odijk, 2015).

These treebanks have already been used to study various linguistic phenomena. For Dutch, several applications of the Lassy Large treebank are discussed by van Noord & Bouma (2009). A study of extraposition of comparative objects by van der Beek et al. (2002) was used to illustrate the grammar used by the
Alpino parser, but it attracted criticism for allowing too much extraposition. As evidenced by a note (van der Beek et al., 2002, 364, note 8), a reviewer claimed that such extraposition was not possible from the front of the sentence (the topic), however, a search of the large corpus revealed that such sentences were in fact being used in particular contexts. It was judged to be a probabilistic phenomenon, more or less acceptable depending on various factors. This shows that automatically annotated treebanks can also be used to refute claims based on linguistic theory. Bastiaanse & Bouma (2007) used syntactic structures from the treebank to argue that patients with Broca’s aphasia have difficulty with constructions of higher linguistic complexity, rather than due to a frequency phenomenon. Bouma & Spenader (2008) studied the distribution of the Dutch reflexives \textit{zich} and \textit{zichzelf} with regards to the verbs with which they co-occur, where different verbs can select one or both of the options. These are examples of studying rare constructions or co-occurrence patterns, a task that automatically annotated treebanks are particularly suitable for.

Another such task is the study of probabilistic effects. Bloem et al. (2014) used a part of the Lassy Large treebank to study Dutch verb cluster constructions, a word order variation in which a variety of factors play a probabilistic role. Like English, the Dutch language may use auxiliary verbs to express features such as tense and aspect. In verb-final subordinate clauses, these verbs are grouped together at the end of the clause, and in main clauses, the first (finite) verb goes to the second position while the others are grouped together at the end of the clause. In subordinate clause two-verb clusters, both logical orders are grammatical:

(3.1) \textit{Ik zei dat ik het gehoord heb}  
\hspace{1cm} I said that I have heard
\hspace{1cm} ‘I said that I have heard it.’

(3.2) \textit{Ik zei dat ik het heb gehoord}  
\hspace{1cm} I said that I have heard
\hspace{1cm} ‘I said that I have heard it.’

Speakers may choose between the orders depending on a variety of factors relating to discourse, semantics, mode of communication or processing complexity (De Sutter, 2009; Bloem et al., 2014). Larger clusters of verbs are also possible, but not all of the logical orders are grammatical when three or more verbs are involved, although there is still variation.

Studying this phenomenon involves searching a treebank for groups of verbs in a particular syntactic configuration: an auxiliary or modal verb heading a participial or infinitival main verb. This study also replicates earlier work on a manually annotated corpus (De Sutter, 2009), showing that the errors caused by the automatic parsing are not necessarily a problem for linguistic study, although a few factors (i.e. word stress patterns) could not be studied due to the nature of the annotation that can be found in a treebank of written texts.
While the manual study involved 2,390 instances of verb clusters, a sample of 411,623 clusters was gathered from the treebank.

Odijk (2015) showed that automatic annotation can even be used to study child utterances from the Dutch CHILDES corpus. This corpus was parsed with the Alpino parser for Dutch, even though this parser has not been trained on child language data. Spoken child language is a rather different domain than adult written language, making parsing errors likely. Nevertheless, a study of three near-synonymous Dutch degree modifiers that translate to ‘very’ was conducted on this data, along with an evaluation. The interesting thing about these modifiers is that two of them, *erg* and *zeer*, are used with adjectival, verbal and adpositional predicates, while one, *heel*, is only used with adjectival predicates. It is not clear how children acquire this difference. A corpus of child utterances and child-directed speech with syntactic information may reveal how much evidence there is for these constructions in child language. Useful results were obtained despite the issues, likely due to the focus on a particular linguistic phenomenon involving modifiers rather than larger syntactic structures — the two most common of the three degree modifiers were found with high accuracy.

Using the TüPP-D/Z treebank, auxiliary fronting in German three-verb clusters was studied (Hinrichs & Beck, 2013). Since three-verb clusters in subordinate clauses are a somewhat rare construction, and auxiliary fronting inside of them even more so, the massive size of the corpus was a requirement to be able to find enough instances. The authors observe what verbs participate in the construction and compare the treebank data to (much more sparse) information from diachronic corpora. For English, Lehmann & Schneider (2012) used a 580 million word dependency-parsed corpus 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, requiring massive amounts of data to study. Bouma (2017a) used the Lassy Large treebank to study an infrequent phenomenon, of which the frequency of occurrence has often been debated: long-distance dependencies. He finds enough examples of specific types of long distance dependencies to test some theoretical claims from the literature, that had not been tested using corpus data before. Automatic parsers also enable this type of research to be conducted for multiple languages — Blasi et al. (2019) studied the frequency of deep clausal embeddings, another construction that is of great importance to theoretical linguistics, in 17 different languages.

3.3 Current Approaches to Evaluation

The quality of automatically annotated treebanks is usually evaluated by testing the performance of the parser that was used to create it. Therefore, treebanks are evaluated in the same way as parsers, using an overall accuracy score such as the word-based Attachment Score. This is the percentage of words that have been assigned the correct head in the syntactic structure (sentence-based
variations or variations that include dependency labeling also exist). The Alpino parser (van Noord et al., 2006) that was used to create the Dutch Lassy Large corpus was evaluated using Concept Accuracy (the proportion of correct labeled dependencies) as a measure. A part of the corpus containing texts from various domains (e.g., books, newspaper texts) was manually verified in order to have a gold standard to compare against. This resulted in an accuracy score of 86.52%, but with clear variation across different domains (van Noord, 2009). Studies based on the corpus often report this score as a measure of quality.

However, even this is too general for the purposes of linguistic research. Rather than some domain of text, a researcher is primarily interested in one particular construction, and wants to know how accurately that particular construction was parsed in the corpus. If the parser often errs in labeling adjectives, this does not matter if one wants to investigate reflexives, but it would be a major problem for a study of adjectives. Parser errors cannot be entirely dismissed as random variation, some of the errors are likely to be systematic due to the nature of (most) syntactic adjectives being probabilistic based on statistical learning.

One obvious consequence of this is that a parser is more likely to make mistakes when parsing rare phenomena, for which there was little evidence in the parser’s training data. Phenomena that are of interest to linguists are often rare. Related to this is the fact that parsers generally perform worse on longer sentences, as shown in van Noord et al. (2006, 11, Fig. 5) for the Dutch Alpino parser. Sentence length is sometimes considered as a probabilistic processing effect in multifactorial linguistic studies, so this should also be considered. More errors occur when there is more ambiguity, regardless of whether the ambiguity is caused by semantic or syntactic factors. Multi-word units (idiomatic expressions) are also known to cause parsing errors (Nivre & Nilsson, 2004), but on the other hand, a parser may have been specifically improved to deal with multi-word units. Text types that are different than what the parser was trained on, such as the child utterances in the study by Odijk (2015), may also cause a higher error rate. Lastly, when the original text contains errors or unusual spelling, a parser is also likely to make more annotation errors.

Due to this possibility of systematic errors which may introduce more errors for certain constructions than for others, I believe that semi-automatic or manual construction-specific evaluation is necessary, using the knowledge of linguistic experts. Such an evaluation will provide insight into the quality of data gathered from automatically annotated treebanks for the purpose of linguistic study.

Some studies using automatically annotated treebanks have taken this approach. For example, Odijk (2015), in his CHILDES study, compares the parser accuracy for the specific words being studied against a manually annotated gold standard. However, such a gold standard is not always available, and the manual annotation was also found to contain errors. These errors were found by looking up the words manually, which is not possible if one is investigating more general constructions that can involve many words types, instead of particular words. Furthermore, the data set of child utterances of the constructions being
investigated was fairly small, making a thorough manual evaluation more feasible than on large automatically annotated corpora. Therefore, this approach to evaluation is not always applicable. Bloem et al. (2014) took a semi-automatic approach by manually verifying a portion of the results of their syntactic queries for verb clusters. While the precision of the results can be measured with such an evaluation, it does not address the issue of recall. Any relevant construction that was annotated incorrectly and therefore missed by the querying procedure, will not be in the sample. In other studies, i.e. Hinrichs & Beck (2013), the paper does not address the issue of construction-specific evaluation at all.

In the next section I will discuss four possible approaches to construction-specific evaluation for linguistic research using an automatically annotated corpus. I will illustrate the four approaches with examples from the Dutch verb cluster research described by Bloem et al. (2014), who conducted their research on the automatically annotated Lassy Large corpus. In listing these approaches, I am assuming that the linguist is faced with a corpus that is the end product of automatic annotation. This hypothetical researcher does not have access to, or is not able to use the tools that were used to annotate it, i.e. the methods do not require much technical knowledge. Without this restriction, other approaches could be taken, and have already been taken, such as re-training and/or evaluating the parser on an adapted text, using or creating a different annotation tool that is designed to target the construction of interest specifically, or simply parsing a large number of instances of the construction being studied and evaluating the parser’s performance on that procedure. However, it is unlikely that someone whose main interest is linguistic research would have the knowledge or motivation to perform such procedures.

3.4 Linguistically Informed Evaluation

The main difficulty of this task, evaluating some subset of a large corpus (i.e. all verb clusters), is in gaining insight into precision and recall at the same time. The four approaches discussed here have various strengths and weaknesses relating to these two measures that I will discuss. An overview of the four approaches discussed in this section and their relative benefits is shown in table 3.1.

3.4.1 Manual Evaluation of the Results

The most obvious approach is a complete manual evaluation of the results by a linguist. This involves first formulating a query that matches a specific construction, and manually inspecting the results of the search. Any result that matched the query but was not actually an instance of the construction being studied, whether it was due to an annotation error or an imprecise query, is marked as false, and others as correct. A percentage can then be calculated, which represents the precision score of the query. However, this method has several disadvantages. Firstly, it may take a lot of time and resources to evaluate
Table 3.1: Overview of the strengths and weaknesses of each approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Weaknesses</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual evaluation of the results</td>
<td>No recall, somewhat costly</td>
<td>Precision</td>
</tr>
<tr>
<td>Manual evaluation of the text</td>
<td>Extremely costly</td>
<td>Both P &amp; R</td>
</tr>
<tr>
<td>Fall back to simpler annotation</td>
<td>Misses POS-tagging errors</td>
<td>Recall</td>
</tr>
<tr>
<td>Search for particular instances</td>
<td>Hard to generalize result</td>
<td>Recall</td>
</tr>
</tbody>
</table>

all results extracted from a large corpus in this way — Bloem et al. (2014) automatically extract 411,623 verb clusters from the 145 million word Wikipedia part of the Lassy Large corpus, too many to verify manually in any reasonable time frame. A representative sample of the results would have to be used. Secondly, this method may still miss constructions that were systematically misparsed. For example, if a researcher is searching for verb clusters but verbs in a particular type of cluster have been mistagged as adjectives, a search query for verb clusters will not find those mistagged instances, and the researcher will not know of their existence. The precision of the results can be measured with such an evaluation, but not the recall.

I have tested this method on the first 10,000 sentences of the Wikipedia section of the Lassy Large Large treebank using two-verb auxiliary clusters from subordinate clauses, as shown in examples 3.1 and 3.2 as an example construction. This sample of the corpus contains 193,378 tokens, covering 0.13% of the Wikipedia section of Lassy Large. A syntactic search for the target construction yielded 315 matching verb clusters, of which five were found to arguably constitute errors — these five verb clusters all had adjectival instead of verbal participles. An example of that would be ‘He thought the door was closed’, where ‘closed’ can be an adjective as well as a verb, and these five cases were annotated as adjectives. However, the fact that they could be verbs was also available in the annotation, so these five examples may not be errors depending on your theoretical perspective. Therefore, the precision of the annotation for this part of the corpus is $\frac{310}{315} = 0.984$, or 1. From this we can conclude that two-verb clusters were likely parsed with very high precision by the Alpino parser when this corpus was created.

### 3.4.2 Manual Evaluation of the Text

To solve the recall problem, it may be possible to do a manual evaluation of the text. By reading the original corpus text rather than just the results of a search query, even instances of the construction of interest that are completely misparsed can be found by the linguist. However, this is extremely labor-intensive —
3.4. Linguistically Informed Evaluation

<table>
<thead>
<tr>
<th>Error category</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of longer cluster</td>
<td>56</td>
<td>74.67%</td>
</tr>
<tr>
<td>Parsing error</td>
<td>7</td>
<td>9.33%</td>
</tr>
<tr>
<td>Query error</td>
<td>12</td>
<td>16.00%</td>
</tr>
<tr>
<td><strong>Total differences</strong></td>
<td><strong>75</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

Table 3.2: Results of a comparison between syntactic search and POS-based search, listing the verb clusters found only by the latter one.

one will have to read a lot of text to find just one instance of a rare construction, even if only a part of the corpus is evaluated in this way. This takes away the main advantage of using a large automatically parsed treebank, and even if only a representative sample of the corpus is read, this is only feasible for common constructions. Therefore, I have chosen not to demonstrate this approach.

3.4.3 Fall Back to Simpler Annotation

Another solution is to fall back to a simpler annotation layer. It is generally the case that annotation of larger structures is more difficult. Lemmatizing and tagging (assigning a word class) only involves words, while parsing adds syntactic structure over multiple words. Queries based exclusively on word class will therefore result in fewer errors than searching on the basis of syntactic structure. For example, to retrieve verb cluster construction one would normally want to find a verb that is the head of another verb. But in this way, verbs that were attached incorrectly will erroneously be skipped. If the researcher simply searches for two verbs positioned next to each other in the linear order, these skipped verbs would also be included, at the cost of retrieving verbs that are next to each other coincidentally (i.e. as part of two different clauses) or as part of a larger structure. Comparing the result of the two procedures will produce a list of ‘suspicious’ instances, which can be evaluated manually (to be either included or excluded from the study) with less effort and better recall than when the results of a regular corpus query are manually evaluated. This does mean that the linguist will have to come up with some sort of word-class-based approximation of the construction under study using their knowledge of the language.

I have again tested this method on the first 10,000 sentences of the Wikipedia section of Lassy Large, comparing the result of a syntactic search with that of a part-of-speech (POS) based search using only the features of word class, lexical category and linear position in the sentence. The results of this are shown in
table 3.2. I identified all results that were retrieved by the POS-based search but not by the syntactic search, and manually verified them. There were 75 such results in total. In 56 cases, the query had actually matched a group of two verbs that was part of a larger verb cluster of more than two verbs. Since only two-verb clusters, not three or four verb clusters are the target construction, these are not errors. It is difficult to avoid getting results from larger clusters when using POS-based search, since the difference is syntactic. In seven cases, there was an actual two-verb cluster that had been misparsed. These cases were mostly located in very long sentences with many parsing errors. The syntactic search had missed these, indicating a recall issue. A further 12 results also contained actual two-verb clusters and were annotated correctly, but were not identified by the syntactic search. This indicates a problem with the syntactic query rather than with the annotation. They can be considered retrieval errors, not annotation errors. Most of them involved verbal particles directly before or after the verb cluster, which I did not consider when formulating the syntactic query. Detecting such errors using this method can help the researcher to refine their queries. Overall, to the 315 verb clusters that were found in the previous section, we can now add 7 + 12 new ones that were not identified by the syntactic search. This also allows me to compute the recall over this part of the corpus: 

\[
\frac{315}{322} = 0.943, \text{ or } \frac{315}{334} = 0.978 \text{ if we do not wish to consider the query errors as a recall problem.}
\]

Again, this is only an estimation of recall, calculated over a fraction of the entire corpus and using manual comparison of the results. Furthermore, this estimate does not take into account that the part-of-speech tagging may also contain errors. Automatically annotated part-of-speech tags are not perfect either, they are just more correct than the parse trees. The final approach I will discuss does not make this assumption, but instead circumvents the annotation as much as possible.

### 3.4.4 Search for Particular Instances

It is possible to search for particular instances (types) of the construction without relying on the annotation at all. The linguist can choose some representative instances of the construction and search for it directly. For example, when searching for Dutch verb clusters, one could simply search the corpus for the string 'hebben gehad' 'have had', one of many possible combinations of verbs. I will call this a 'string query', as opposed to a 'syntactic query' that one would normally perform on a treebank. This will only result in a limited number of results, but in a large automatically annotated corpus there can still be many results for a specific combination of words, even if the total number of instances of the general construction (verb-verb combinations in this case) is much larger. This string query does not rely on any syntactic annotation, and it would therefore find the verbs even if they were annotated completely erroneously, i.e. as a preposition heading a preposition. These results for the string 'hebben gehad' can then be compared to results for the syntactic structure of 'hebben gehad' with these particular words to see whether there is a recall problem:
if an example occurs in the string query but not in the syntactic query, it is annotated incorrectly in a way that makes it impossible to find with a syntactic query. If the researcher does this for various instances of the construction, they should get a clear idea of the reliability of the annotation and what sort of errors to look out for. However, it would be impossible to find all annotation errors in this way, since for most research questions it would not be possible to search for every instantiation of a construction.

One disadvantage of this method is that it requires generalization. If you measure the recall for the `hebben gehad` verbal cluster, you might assume that the recall is similar for other two-verb clusters, but this is not necessarily true — perhaps the recall is worse for less frequent words. This concern may be alleviated by sampling a variety of instantiations of the construction. An advantage of the method is that it can be used not only to evaluate the quality of the automatic annotation, but also of the annotation scheme. It has been argued, most notably by Sinclair (2004), that it is better to avoid any sort of annotation if possible, as this already imposes theory upon the data. It may be the case that the annotation scheme of the corpus makes incorrect theoretical assumptions, groups different phenomena together into one category or makes arbitrary distinctions. By performing a query that avoids the annotation altogether and comparing its results to those of a query that does make use of the annotation, such issues can be detected by a linguistic expert.

I have also applied this method to the Wikipedia data. I could not use only the first 10,000 sentences of the corpus, because there is only one verb cluster instance that occurs more than once in this sample. Instead, I took the first 300,000 sentences of the corpus and searched for the string `hebben gehad`, a common combination of common words, as well as the syntactic version: the verb cluster `hebben gehad`. The syntactic search resulted in four correct examples of `hebben gehad` verb clusters, while the string search provided 14 results, of course including the 4 correct examples from the syntactic search. Nine of the other results were actually verb clusters in main clauses, which are not the target construction, but the distinction cannot be made with just a string search because main clause verb clusters and subordinate clause verb clusters have the same form. However, the remaining string search result was indeed a valid `hebben gehad` verb cluster. On closer examination it had been parsed incorrectly, and therefore it could not have been identified by the syntactic search. It occurs in a sentence with an unusual structure that the parser apparently failed to parse completely, with the main verb `hebben` being left outside of the sentence structure. From this string search, it appears that there were actually five clusters, of which four were identified by the syntactic search. The recall here is 0.8, over this extremely limited sample for this construction.
3.5 Conclusion

In this chapter, I have discussed the issue of using and evaluating linguistic data from automatically annotated treebanks for the purposes of linguistic research. I compared four approaches to evaluation and illustrated them with examples based on a recent linguistic study. These evaluation methods may help to alleviate the concerns that linguists often have about the inaccuracies of such corpora and provide more detail than traditional measures of parsing accuracy when the goal is to study specific constructions.

Since the proposed methods all have different advantages and disadvantages, it would be best to combine them when studying a particular construction. Manual evaluation of the results can be used to determine the precision of a corpus query’s results, while searching for particular instances can be used to calculate recall over a portion of the data, determining how many examples might have been missed. Falling back to simpler annotation can be used as a verification of the syntactic annotation of the corpus, even over larger amounts of data, and provide a rough estimate of recall.

Besides estimating precision and recall, the proposed methods can also be used to identify query errors, such as the one we encountered in section 3.4.3. Formulating accurate queries for large corpora is risky and requires a full understanding of the intricacies and idiosyncrasies of the annotation scheme used, even if an user-friendly interface is available (Odijk, 2020). Examining the results of queries in these ways during the process of data extraction can help to improve the queries that are used before proceeding to study the sampled constructions further.

The proposed methods may also have broader applications: research in digital humanities may also involve searching through annotated datasets, and these methods can be used for evaluating the reliability of those searches as well. For example, philosophers have used semantic information retrieval to obtain data for a conceptual analysis (van Wierst et al., 2018). Searching for particular instances using string search can help to validate such approaches.

While the methods do require some manual annotation effort, they allow a researcher to get a clearer impression of the quality of the annotation of the particular construction they are investigating in the corpus, while still preserving the advantage of being able to obtain many exemplars with relatively little manual effort. Reporting on such a construction-specific evaluation in a large-scale corpus or treebank study makes the results easier to interpret for those who are not familiar with the errors that an automatic syntactic parser might make. Clearer quantification of the error rate for the linguistic phenomenon that is being studied will also allow researchers to make more definite claims on the basis of data from automatically annotated treebanks. In future work, a larger-scale empirical evaluation of these approaches on a wider variety of constructions and corpora could be conducted to assess them in more detail, and perhaps to create a better reason for linguists to use automatically annotated
treebanks in their studies. Furthermore, it may be interesting to investigate whether construction-specific accuracy scores can be incorporated into corpus-based statistical models of language phenomena as part of the margin of error.
Chapter Highlights

Problem Statement

- Many factors have been linked to Dutch verb cluster word order variation, but several contrasting explanations have been proposed as to why these factors are involved.

Research Questions

- Can the factor of processing complexity account for the observed variation in Dutch verb cluster order?

- If so, which order is the easier to process ‘default’ word order: main verb first or main verb last?

Research Contributions

- A variety of factors that are related to verbal cluster word order, can also be related to the processing complexity of the cluster’s context. Minimizing

4.1 Introduction

Verbal clusters in Dutch present an interesting example of grammatical variation. In clusters of two verbs, both possible word orders are grammatical, leading to optionality:

(4.1)  *Ik denk dat ik het begrepen heb*  
I think that I it understood have  
‘I think that I have understood it.’

(4.2)  *Ik denk dat ik heb begrepen*  
I think that I it have understood  
‘I think that I have understood it.’

Speakers may produce either order in similar contexts, and the choice between different variants does not appear to be determined by grammatical factors. A notable aspect of this optionality is that the difference in word order is generally assumed not to correspond to a meaning difference (e.g. von Humboldt 1836). However, when speakers choose between constructions in these situations, they do not do so randomly.

This variation cannot be explained on the basis of syntax alone. In the most widespread account of the syntactic mechanism behind verbal clusters, both constructions are of equal derivational complexity, making an explanation in terms of syntactic complexity unlikely. This account, first proposed by 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 4.1 illustrates this process. 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 4.1b) or 1-2 (Figure 4.1c) ordered verbal cluster (1 is the head, 2 is the participle). The claim that verbal clusters are formed using this special mechanism is illustrated by the fact that these clusters are very rarely interrupted by non-verbal material. Furthermore, in multiple verb constructions with more than two verbs the *Infinitivus Pro Participio* (IPP)
Verbal cluster order and processing complexity

(a) Main verb initial position
(b) Forming example 4.1
(c) Forming example 4.2

Figure 4.1: Verb raising to generate the verbal clusters of examples 4.1 and 4.2, following the analysis of Evers (1975).

 effect shows up, meaning that modal verbs that would normally be participles are infinitives instead (Wurmbrand, 2006).

Corpus studies of similar variation phenomena have shown that a large portion of the variation between near-synonymous constructions in an alternation can be statistically accounted for using multifactorial statistical models that incorporate a variety of linguistic factors beyond syntax. De Sutter (2005) and Bloem et al. (2014) found that this is also the case for Dutch verbal clusters, quantifying the impact of morphosyntactic as well as semantic factors, and properties such as sentence length and word frequency. The multifactorial models employed in these studies reveal interesting patterns in the variation and associations with linguistic factors, but to explain why all of these different factors are involved when people choose to use one word order or the other, one needs to generalize over the factors that can be found by measuring things in a corpus. Various generalizations have been suggested. Several of the observed effects regarding factors such as sentence length can be interpreted as effects of adhering to the rhythm of the Dutch language (De Schutter, 1996), evenly distributing information weight throughout a sentence (De Sutter et al., 2007), or minimizing processing complexity (De Sutter, 2005; Bloem et al., 2014).

Processing complexity refers to the amount of cognitive resources or effort required to produce or comprehend an utterance. Speakers prefer to minimize their use of cognitive resources, formulating sentences in a way that minimizes processing complexity when they have multiple grammatical ways to communicate something (see Jaeger & Tily (2011) for an overview, Levy & Jaeger (2007) and Fedzechkina et al. (2012) for experimental work).

In this chapter, we will explore the suggestion that minimizing processing complexity is a generalization that can explain the effects observed in multifactorial studies so far. De Sutter (2005) originally formulated this idea for Dutch verb cluster order variation on the basis of two factors that were found to corre-
late with verb cluster order in his study — main verb frequency, and differences between verb clusters with arguments that are complements or adjuncts. He argues that the 1-2 word order is stylistically preferred, and is therefore more likely to be used when speakers have the spare cognitive capacity to do so, while the 2-1 order is otherwise more likely to be used. De Sutter ends his dissertation with the suggestion that in future work, other factors from his multifactorial study should be interpreted from the same perspective of processing complexity. Bloem et al. (2014) take this idea and suggest that it can be tested by studying whether factors indicating ease of processing correlate with one particular verb cluster word order, which would be the ‘default’ order. They also state that this default order is not necessarily the 2-1 order, and that there are arguments for considering the 1-2 order to be the default as well. That makes the processing complexity generalization empirically testable – is there a word order that correlates with processing complexity, and if so, which one? The study in this chapter aims to answer this question.

We address the issue by gathering empirical evidence for factors that affect verbal cluster order using a large amount of corpus data. Considering the wide range of factors that have been found to be associated with verbal cluster order variation in earlier studies, we expect that semantics, information structure and processing complexity all play a role in the choice between orders. However, in this study we focus on factors related to processing complexity, while controlling for other influences as much as possible. We hypothesize that any factor that cannot be attributed to semantics (i.e. it does not affect the meaning of the utterance), should be attributed to processing. We largely follow the methodology of Bloem et al. (2014), but with more factors relating to processing complexity. Furthermore, we analyze the previously studied factors in terms of processing complexity. This approach can be viewed as analogous to that of De Sutter et al. (2007). They use a multifactorial model to study rhythm as a possible generalization over factors relating to verb cluster word order, but conclude that information weight is a more plausible generalization, suggesting the creation of another multifactorial model that includes new and more specific factors relating to information weight.

In the next section, we will consider some relevant studies and insights, firstly on the topic of word order variation, secondly on processing complexity, and lastly on Dutch two-verb clusters and the extent of the order variation there. In section 4.3, we propose two testable hypotheses on processing and the order variation. Section 4.4 discusses our methodology and the corpus that we used to test our hypotheses, and we discuss factors that relate to both verbal cluster order and processing complexity in section 4.5. In section 4.6 we present our results, followed by a discussion of the findings and a conclusion in section 4.7 and 4.8 respectively.
4.2 Related work

4.2.1 Word order variation

Word order variation in Dutch verb clusters can be viewed as a case of a syntactic alternation, or of near-synonymous constructions, where speakers have two alternative forms available that are both grammatical. Syntactic alternations have frequently been studied using multifactorial models based on corpus data, with a focus on the study of factors that can be measured in an objective way. This is the perspective from which we approach the phenomenon. While we discuss background from generative, construction grammar and functional linguistic theory to describe the observed factors, we do not intend to take a specific theoretical stance. The different theories predict that different factors shape grammar, but they do not place strong restrictions on how speakers decide between alternative near-synonymous constructions when multiple options are grammatical. To summarize some theoretical discussions on the nature of syntactic alternations, we will briefly compare Dutch verbal cluster order variation to the history of a better-known case of grammatical variation, the English dative alternation.

(4.3) He gave his friend the ticket.
(4.4) He gave the ticket to his friend.

This optionality, a choice between the double object construction shown in example 4.3 and the prepositional dative construction in 4.4, presents a similar puzzle as the Dutch verbal cluster order variation. The dative alternation was already discussed as a problem for transformational theories of language by Oehrle (1976). Moravcsik (2010) summarizes the problem that this alternation presents: it violates the principle of isomorphism, the one-to-one relation between meaning and form. This has been claimed to be a near-universal property of language (Haiman, 1980), and thus cannot easily be abandoned. Moravcsik (2010) discusses two kinds of solutions that linguists have pursued in resolving this conflict.

The first solution is to argue that the two forms actually have a single underlying structure, and are therefore actually the same thing, preserving the one-to-one relation. Larson (1988) proposed an analysis in which the two constructions are different surface representations of the same underlying structure. This approach has been applied to Dutch verbal clusters as well (Evers, 1975).

The second solution is to argue that the two forms actually have two different meanings, which are influenced by pragmatic aspects that are context-dependent. Beck & Johnson (2004) take this approach, claiming that the two constructions have different underlying forms with different argument structures, giving rise to different meanings.

From the perspective of generative linguistics, optionality between multiple near-synonymous constructions is often considered to be a question of how
objects generated by the grammar are used (Embick, 2008), and therefore outside of the scope of syntactic theory. Even in an approach where optionality is sensitive to meaning, such as in Beck & Johnson (2004), the meaning arises from the verb’s argument structure, which does not differ between the two word orders in the case of verbal clusters.

The multiple meaning approach is also common in construction grammar. Goldberg (1992) accounts for apparent usage differences between the two forms of the dative alternation by claiming that the meaning difference comes directly from the ditransitive argument structure, which carries its own semantics. Bresnan et al. (2007) later showed that information-structural properties can also account for a large portion of the variation between the two forms, using corpus evidence. In construction grammar, constructions are stored pairs of form and meaning, or form and function, in which at least some aspect of the construction’s function is not predictable from its component parts. Any difference in form therefore implies a difference in meaning or function (Goldberg, 2002).

This assertion provides some additional testable predictions about the factors that may influence verbal cluster order variation. If the two verbal cluster orders in Dutch are two constructions, they must have differences in meaning or function. However, the notion of ‘allostructions’ has also been proposed — variant structural realizations of a partially underspecified construction (Cappelle, 2006). For two-verb clusters, this would mean a verb cluster construction that is underspecified for word order, with two possible allostructions in which word order is specified. This idea formalizes the relation between two forms of an alternation, but without one form being an underlying form of the other. In this view, there does not need to be a functional or meaning difference between two alternative word orders.

Word order variation can also be influenced by functional word order principles, which have been claimed to shape grammars, for example by Hawkins (2014). This idea can be extended to word order preferences, when an alternation is present and multiple options are grammatical. Some of the factors that have been discussed in the context of verb order variation can be interpreted as consequences of such principles. This perspective is discussed by Coussé (2008).

She provides an overview of several functional word order principles, such as the adjacency principle, which states that conceptually related elements are placed together in the linear ordering of an utterance. The salience of certain positions (such as the first and last position in a sentence) is also discussed, as well as several principles of information structure, such as ‘topic-before-comment’, and complexity-related principles such as ‘light-before-heavy’. We will now address some theories that have been used to account for such complexity principles, and how they can be related to syntactic alternations.

### 4.2.2 Processing complexity

Processing complexity, more specifically the facilitation of processing by avoiding complex utterances, can be considered to be a functional motivation for word
Verbal cluster order and processing complexity

order variation. While this idea does not seem to be spelled out explicitly in theoretical generative or construction grammar literature, it has been discussed extensively in the context of psycholinguistic studies. Two main approaches to human language processing can be identified here: constraint satisfaction models, and resource-limitation (or memory-based) models (Levy, 2008). These theoretical models can make predictions regarding the complexity of structures for the purposes of human language processing.

Constraint satisfaction models consider various parallel alternative interpretations or parses of a sentence during processing, and relate processing difficulty to expectation. MacDonald (2013) presents a range of phenomena in which cognitive constraints and processing difficulty are shown to affect language production, including word order. Such processing effects have been modeled computationally by Hale (2001), using a measure of surprisal. Within this line of research, Levy (2008) proposed a resource-allocation theory of processing difficulty, using probabilistic word models. In this model, speakers would be able to choose between multiple possible parses when comprehending a sentence, preferring the interpretation with least surprisal and evenly distributed surprisal throughout the sentence, thereby minimizing processing difficulty (Jaeger, 2010). It has also been shown that structures that are dispreferred according to such probabilistic models can be produced, but are produced less fluently (Cook et al., 2009), showing that these models may be able to account for processing during production as well as comprehension. There are many different approaches to modeling surprisal though, with different representations (i.e. n-gram models or construction grammars) and different units (i.e. words or constructions) (Jaeger & Tily, 2011).

The other approach, resource-limitation models, focuses on the idea that there is some limited cognitive resource, such as memory, that imposes processing limitations. In this approach, processing difficulty is often defined in terms of dependency length — it is more difficult to process words that depend on each other when there is a larger distance between them. Gibson’s (1998) work is an example of this. In this theory, there is a memory cost for keeping track of obligatory syntactic requirements that need to be fulfilled with a dependency relation, as well as a cost for integrating an element with the current parse. This approach predicts higher processing costs for longer dependencies. This dependency length idea also corresponds to the first efficiency principle proposed by Hawkins (2014), Minimize Domains. Hawkins argues that these principles play an important role in shaping the grammar of a language. These principles are used to account for cases of language variation, including word order patterns. The basic idea of Minimize Domains can be applied to verb clusters fairly easily — the auxiliary verb is the head of the clusters and therefore dependency relations to e.g. the subject attach to it. If the auxiliary verb comes first, as in the 1-2 order, the dependency link can be made sooner. Therefore, the 1-2 order is easier to process following this approach.

These two approaches to modeling language processing do not appear to be mutually exclusive, they just have a different focus. Furthermore, they do not
deal with all domains of language, but mainly with syntax, and are therefore not complete accounts of how language processing works. As noted by De Sutter (2009, p. 226–227), it can be difficult to apply these models or principles of word ordering to the study of verb clusters, because most word order variation studies discuss interconstituent alternations (e.g. Bresnan et al. 2007; Rispens & Soto de Amesti 2016) while the two-verb cluster orders are an intraconstituent alternation (within the verb phrase).

The connection between word order variation and processing difficulty has been made before in discussions of various optionality phenomena. Generally, a multifactorial corpus study is performed, an approach pioneered by Gries (2001). These studies often show that a multitude of factors affect the choice between the two options leading to a hypothesis that generalizes over all of the factors under discussion. In the multifactorial corpus study by Gries, on optional particle movement in English, this hypothesis is called the Processing Hypothesis (PH):

The multitude of variables (most of which are concerned with the direct object NP) that seems to be related to Particle Movement can all be related to the processing effort of the utterance (Gries, 2001).

One example of a variation phenomenon where processing difficulty plays a clear role that cannot be confused with information structure, is the optionality of that in English non-subject relative clauses, as in 4.5:

\[(4.5)\] They have all the water (that) they want.

Wasow et al. (2011) explain this optionality in terms of predictability and processing. A processing account has been suggested for a similar Dutch-language phenomenon as well. In a study on the optionality of the om-complementizer in Dutch, Bouma (2017b) suggests that om-insertion is related to linguistic processing. It serves to reduce ambiguity in the contexts in which it is used, which reduces the processing complexity of the clause.

These studies show that processing complexity can be part of the explanation for optionality phenomena and grammatical variation. The models of language processing discussed in this section can be used to infer what factors might affect the processing complexity of a sentence.

### 4.2.3 Verb clusters

Verb clusters are a widely studied phenomenon in Dutch syntactic literature, both in terms of the generative mechanism behind them, and in terms of their optionality (Evers, 1975; den Besten & Edmondson, 1983; Haegeman & van Riemsdijk, 1986; Zwart, 1996; Wurmbrand, 2004). Some of the first studies of Dutch verbal clusters were focused on regional variation (Pauwels, 1953), and the extensive regional variation, as documented in the Syntactic Atlas of Dutch Dialects (Barbiers et al., 2008), is used as a dataset for testing theories
Verbal cluster order and processing complexity

55

on linguistic parameters that may govern this variation (Van Craenenbroeck et al., 2019). The mechanism of verb raising is generally used to explain the existence of verbal clusters, as illustrated in section 4.1. A broad overview of verb raising across Germanic languages, as well as different theoretical accounts of the phenomenon, is provided by Wurmbrand (2006).

While various terms have been used to describe the two orders, we will follow Stroop (1970) and call the construction in example 4.1 the descending order, and the construction in example 4.2 ascending. This is because the finite auxiliary is considered to be the verb that is highest in the syntactic structure of the clause, while the main verb is the lowest. This leads us to number the verbs accordingly: in example 4.1 the verbs follow the order 2-1, because the head comes last, and the main verb comes before it. Larger clusters can thus be adequately described, e.g. a 4-3-2-1 cluster, where 4 is the main verb, and 1 the head. We adopt this numbering convention throughout this thesis.

In this section, we provide an overview of verbal cluster constructions with two verbs. We first discuss clusters with participial main verbs, which exhibit more order variation, and then clusters with infinitival main verbs, which exhibit limited variation. Lastly, we discuss the ‘te’-ininitival cluster constructions. The discussion follows the structure of the description of verb groups in the Algemene Nederlandse Spraakkunst (Haeseryn et al., 1997, p. 946–1076) to some extent. For a more complete description of possible Dutch verb clusters, we refer to this work.

The grammatical examples in this section come from the 145 million word Wikipedia part of the Dutch Lassy Large corpus (van Noord, 2009), which we have also used as our data source. Observations on which verbs occur in the different constructions also come from this corpus, unless otherwise noted. The identifiers displayed after the translation of each example are the identifiers of the sentence in the corpus. This corpus will be discussed in section 4.4.

There are various types of two-verb clusters to consider, all of them exhibiting free variation but with different probabilities for each order. Examples 4.1 and 4.2 from the introduction show two-verb clusters with a participial main verb, and auxiliary heads. When we discuss clusters with auxiliary verbs (or ‘auxiliary clusters’), we are referring to those auxiliaries that function as auxiliaries of time, zijn ‘to be’ and hebben ‘to have’, the passive auxiliary worden ‘to be’, and zijn ‘to be’ as a copular verb. These three verbs are the most frequent auxiliary verbs that take participial main verbs, and categorizing them in this way is in line with earlier work (e.g. De Sutter 2005). There are a few constructions in which modals or other types of verbs occur with a participial main verb, but they are said to be headed by an elided infinitive (Haeseryn et al., 1997, p. 960), so we have grouped them with the infinitival clusters. This distinction between auxiliary and other clusters is made because these particular auxiliary verbs exhibit most of the order variation in Dutch two-verb clusters. We use this terminology for practical reasons, without meaning to imply any cognitively real categories or an overarching ‘auxiliary cluster’ parent construction. Most previous studies on two-verb cluster variation concern only clusters headed by
these three verbs.

Besides the clusters with participial main verbs and auxiliary heads, there are other kinds of verbs that can head a verbal cluster, though they appear to exhibit less order variation and stronger preferences for the 1-2 order. These clusters generally have an infinitival main verb. The most frequent of these are the modal auxiliary verbs, which can head a verb cluster with an infinitival main verb. Some authors even say that these verb clusters with modal auxiliaries have no optionality, because the 2-1 order is ungrammatical in them (e.g. Zuckerman 2001). However, this is empirically testable and our data shows that this is not quite true: by counting constructions in a corpus, we can observe that the 1-2 order in example 4.7 is indeed far more common than the 2-1 order in this construction, but examples of 2-1 order usage can be found. The following examples show that both the 2-1 order (example 4.6) and 1-2 order (example 4.7) are used:

(4.6) ... dat iedereen hem ongestraft _doden mocht_ ... that anyone him unpunalyzed kill may
  ‘... that anyone could kill him with impunity.’ [wik_part0021/9811-21-2]

(4.7) ... zodat hij de Jedi _kan uitroeien_ ... so he the Jedi can exterminate
  ‘... so that he can exterminate the Jedi.’ [wik_part0013/5954-13-2]

In the Wikipedia part of the Lassy Large corpus, modal clusters occur in the 2-1 order 0.5% of the time. They seem to be completely acceptable to Dutch speakers, and are considered to be grammatical in various studies. For comparison, auxiliary clusters as defined above occur in the 2-1 order 35.2% of the time in this corpus.

The modal verbs that are used in this construction are _kunnen_ ‘can’, _moeten_ ‘must’, _mogen_ ‘may’, _willen_ ‘want’, _zullen_ ‘shall’, and _hoeven_ ‘need’, as discussed in the ANS (Haeseryn et al., 1997, 984–1006), though _hoeven_ is used with the ‘te’ infinitival marker. 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. For example, in West Flemish and Afrikaans, modal clusters occur in 1-2 order while auxiliary clusters occur in 2-1 order, with exceptions in a few contexts (Wurmbrand, 2006).

Another type of auxiliary verb used with infinitival main verbs is the causal auxiliary _laten_ ‘let’, used in a 1-2 order in example 4.8:

(4.8) ... dat de luisteraar de muziek in 3-dimensionale stijl _laat_ ervaren ... that the listener the music in 3-dimensional style let experience
  ‘... that lets the listener experience the music in three-dimensional style.’ [wik_part0018 /8288-34-2]
This construction must also be grammatical in the 2-1 order at least for some speakers, although its use appears to be quite rare. In the Lassy Large corpus, only 0.114% of the clusters with *laten* are in the 2-1 order, and some of those are annotation errors. The *laten* clusters behave similarly to the modal clusters in this, and also in the fact that they take infinitival main verbs.

In addition to this, there seems to be a restricted number of non-auxiliary verbs that can occur as heads of clusters. These are known as other grouping verbs. Some of these verbs are verbs of perception (*horen* ‘hear’, *zien* ‘see’), some are stative verbs (*bijeen* ‘stay’, *gaan* ‘go’, *liggen* ‘lie’, *zitten* ‘sit’, *staan* ‘stand’), and there are various others (we refer to Augustinus, 2015 for an empirical study and overview of possible grouping verbs). These verbs are also grammatical in both orders, with 2-1 orders occurring 10.6% of the time when averaging over all of the verbs.

In the cluster types we discussed so far, the clusters with auxiliaries had participial main verbs, while the modal and other clusters had infinitival main verbs. However, there is another verbal cluster construction to consider, in which an auxiliary cluster can be made infinitival: the *te*-infinitival construction. These clusters are characterized by the fact that the main verb is marked by the infinitival marker *te*. These constructions have also been called pure-infinitival clusters, to avoid confusion with other clusters where the main verb is infinite (e.g. modal clusters). Furthermore, unlike other modal verbs, the modal and negative-polar verb *hoeven* (need) is used in this construction. Additionally, this construction also occurs with some other grouping verbs, including some verbs that cannot group into a cluster in any other way. This is also known as the second infinitival construction. Lastly, there is the ‘third construction’, in which a grouping verb is used with a *te*-infinitival marker, yet the main verb is not infinitival. These constructions are not accepted by all speakers, and have been noted by den Besten & Broekhuis (1992) to have some interesting syntactic properties.

The examples below show typical *te*-infinitival constructions with auxiliary verbs:

(4.9) \(\ldots waaraan de soort is te herkennen \)
\(\ldots by which the species is to recognize \)
\(\ldots by which the species can be recognized.' \) [wik_part0297/485044-4-1]

(4.10) \(\ldots die te maken hadden met muziek \)
\(\ldots that to make had with music \)
\(\ldots that had to do with music.' \) [wik_part0568/1276744-4-1]

These clusters have not been the focus of any quantitative corpus study, but they appear to have limitations on their optionality with a few particular verbs. When *hoeven* or a grouping verb is used, only the 1-2 order is attested in the corpus.

Next, we will summarize previous work on explaining the order variation.
4.2.4 Verb cluster order variation

There have been various accounts of verbal cluster order variation in cases where multiple orders are grammatical. The rules and mechanisms discussed in generative literature allow for a lot of optionality, as was discussed earlier, 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, 2007b). 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 is not intraspeaker variation, however. **Rhythmic** factors relate to the hypothesis that speakers may change the order of verbs to match the standard stress pattern of Dutch, though this may not be so important in written texts. De Sutter (2009) did not find a strong effect of stressed syllables near the cluster in his corpus study of written texts, except when two stressed syllables would have been directly adjacent. **Semantic** factors are described mostly in terms of lexical semantics. One finding in this area is that adjectival participles have a preference for the 2-1 order. Lastly, **discourse** factors are mentioned with syntactic priming as an example, although this could also be considered to be a processing factor. From all 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. Such a study has been conducted by De Sutter (2009), and expanded by Bloem et al. (2014), testing the association of 10 factors with various types of verbal clusters.

In a psycholinguistic study, Hartsuiker & Westenberg (2000) showed that verbal cluster orders can undergo structural priming, providing evidence that both verbal cluster orders can be considered to be distinct constructions. Under this assumption, variation between the orders may be motivated by semantic or functional differences, following Goldberg (2002).

In his dissertation, De Sutter (2005) suggests an explanation for the order variation and its correlation with various factors that he observes. At the end of his dissertation, he suggests that there may be an underlying psycholinguistic effect of processing. This suggestion of a processing effect is based on two factors, the frequencies of the orders in various modes of language, and a priming effect. He assumes the 2-1 order to be easier to process, because it appears more often in modes of communication in which the ‘production pressure’ is greater — spoken rather than written, interaction rather than monologue, spontaneous rather than edited (De Sutter, 2005, p. 321). Based on this, he suggests that the 2-1 order is the default or most economical order, produced when not enough time or resources are available to produce the 1-2 order. Under this hypothesis, producing the 1-2 order would involve some additional process of sociostylistic encoding (De Sutter, 2005, p. 322-324), a hypothesis that has been discussed...
before, although de Sutter's notion of style is fairly broad. One reason for assuming a sociostylistic process is the apparent preference for the 1-2 order in prescriptive grammars and editing guidelines. This idea is explored further in De Sutter (2007), where 5 factors that correlate with verb cluster word order, are interpreted in terms of cognitive complexity. In this article, he argues that factors that correlate with the 1-2 order have a lower processing cost.

The idea that there is a default order that is easier to process is an interesting one, which we will explore further in the next section. However, we will argue that it is not necessarily the 2-1 order. The issue discussed by De Sutter (2007) can be stated more broadly: Do either of the orders occur in contexts that are generally more difficult to process?

4.3 The processing hypothesis

De Sutter (2005) suggests that the stylistically preferred word order is the 1-2 order, shown in example 4.2. If this is true, one would expect to find this order in contexts that are less difficult to process. Speakers have to make a conscious effort to use the marked, stylistically preferred option, which is more difficult, and they will only do so when they have cognitive capacity to spare (De Sutter, 2007). Something that is more difficult will be used in an easy context, not in a context that is already difficult. We will call this De Sutter’s default 2-1 order hypothesis. Additional evidence for a default 2-1 order is the observation that young children use the 2-1 order more frequently (Zuckerman, 2001).

However, Meyer & Weerman (2016) argue that these 2-1 constructions in young children are instead a result of the base word order pattern of Dutch (Object-Verb, OV). It has been observed that Dutch children produce verb-final utterances as early as the two-word stage (Wijnen, 1997), as well as in their first subordinate clauses (van Kampen, 2010). Therefore, this seems to be a pattern or construction that is acquired early, and children use it before they produce groups of verbs or verb clusters (Wijnen, 1997).

In this view, children put the head verb in last position, strictly following the base order. The other verb, which they might also analyze as an adjective or something else non-verbal, then comes before it (a 2-1 order). Meyer & Weerman theorize that children first learn about verb raising (which forms clusters) when they acquire the 1-2 order. Before verb raising, the head verb is in final position, following the base OV order of Dutch. When raising the verb to form a 2-1 cluster, 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 create a 1-2 order results in an order that clashes with the base word order for Dutch. 2-1 orders can simply be interpreted as verb-final orders, until the learner figures out the mechanism of verb raising from the 1-2 order evidence. This idea is supported by Meyer and Weerman’s observation that Dutch children only overgeneralize the 1-2 order once they have acquired
verbal cluster constructions in general, not only the 2-1 order.

More evidence can be found in observations that were used to account for dialectal variation in Dutch three-verb clusters (Barbiers et al., 2010, 2018). It is theorized that participles can be adjectives as well as verbs in these constructions even in adult language — and in particular, that they are adjectival when used in the initial position in a verb cluster. When we follow this idea of initial adjectival participles, 1-2 orders would be the earliest form of verb raising. This may mean that this word order is more entrenched in speakers’ grammars or lexicons, and therefore easier to process. It is commonly assumed that more entrenched elements are easier to process, as De Sutter (2007, p. 322) also notes regarding participle frequency, and we will make this assumption too. If the 1-2 order is indeed acquired earlier and more entrenched, we would expect to find the 1-2 order in contexts that are more difficult to process, and the 2-1 order in contexts that are less difficult to process. We will call this Meyer & Weerman’s 1-2 order hypothesis.

An additional argument for the default 1-2 order hypothesis can be found in the literature on a related phenomenon that has undergone much psycholinguistic investigation: cross-serial dependencies in Dutch. When a verb cluster consists of multiple verbs that have arguments, the verbs and arguments are ordered with crossing dependencies in Dutch as in example 4.11, while in German, nested dependencies are preferred (4.12):

(4.11) *De mannen hebben Hans de paarden leren voeren.*
The men have Hans the horses teach feed.

(4.12) *Die Männer haben Hans die Pferde füttern lehren*
The men have Hans the horses feed teach.

‘The men taught Hans to feed the horses’

The object and subject dependency relations in the Dutch example cross each other as shown in this simplified dependency graph:

In these verb cluster constructions with embedded arguments, the Dutch clusters are in the 1-2(-3) order and the German ones in the (3-)2-1 order. An experiment by Bach et al. (1986) showed that longer Dutch verb cluster constructions (3 or more verbs) with crossed dependencies are less difficult to process than German ones, as the subjects gave better comprehensibility judgements and made fewer errors in a question answering task. They hypothesize that the processing advantage comes from being able to build ‘higher’ elements
Verbal cluster order and processing complexity

Verbal cluster order and processing complexity

of the structure first. This is also the case when the head verb comes first in a 1-2 order, and may facilitate processing of 1-2 orders. Even though Bach et al.'s (1986) study only showed such difficulties for larger clusters, it might be the case that outside of an experimental setting, with more factors contributing to processing load, there is a processing difference even for two-verb clusters, which can be mitigated by choosing the order that is easier to process. After all, there is no categorical difference, just a tendency or probability towards one of the two orders, which may be stronger when several factors that complicate processing are involved.

Under Meyer & Weerman’s 1-2 order hypothesis, speakers are expected to be more likely to use the more entrenched order in linguistic contexts that require more complex processing, as the more entrenched order is easier to process. Under De Sutter’s default 2-1 order hypothesis, speakers are also more likely to use the order that is easier to process in contexts that require more complex processing.

While both studies work from different theoretical backgrounds, both ideas of defaultness are compatible to some degree. Meyer & Weerman (2016) write “One question that remains central in verb cluster literature, and that is central in the present study, revolves around which of these orders is the default one (or, if the reader prefers, ‘unmarked’, ‘basic’ or ‘most economic’ – we take a neutral stand here), and which is derived”. While the notion of a derived form comes from generative grammar, it is clear that few assumptions on the nature of defaultness are made here. De Sutter (2007) does not explicitly define defaultness, though notes that the default order is the one that is used in situations where speakers do not have the time or the need to choose another order. Furthermore, De Sutter states that, on the basis of dialect research, one could conclude that the most frequently used word order (across dialects, presumably) in the most basic language variety in the most basic communicative situation, is the default order.

In this work, we would like to take a similar neutral stance as Meyer & Weerman (2016), taking economy or ease of processing as the main characteristic of a default word order. We will not commit to a particular theoretical framework’s notion on the nature of defaultness beyond what is mentioned here. This ‘ease of processing’ definition is compatible with De Sutter’s functional definition (assuming that economical equals ‘easy to process’), though it is not necessarily compatible with his second point we paraphrased above: the most frequently used word order across dialects is not necessarily the most economical and therefore the default. For Dutch verb clusters, it appears to be the case that the 1-2 order spreads more easily in contact situations. It is more frequent in the western parts of the Netherlands which experienced more language contact. Furthermore, it has rapidly spread in the Frisian language area once Dutch-Frisian bilingualism became the norm there, even though Frisian was previously a language with the 2-1 order (Meyer et al., 2015b; Hoekstra & Versloot, 2016). Therefore, it may be the case that the 1-2 order is easier to process for Dutch speakers. In a comparison between English, German and Dutch a similar trend
is evident — English, the language with the most language contact, only allows 1-2 orders. German, with relatively little language contact, only allows 2-1 orders, while Dutch is in between, both in terms of contact and in terms of verb cluster word order (Meyer et al., 2015b). We do agree that the most basic language variety is probably the one that is easiest to process, however, it is not clear whether the 2-1 order is still the most prevalent in spoken Dutch, because the data used in Stroop’s (2009) study of spoken Dutch did not take speaker age into account. In a recent study by Olthof et al. (2017), a change towards the 1-2 order in spoken language was observed in younger speakers in the Corpus of Spoken Dutch. It may be the case that only older speakers still have a 2-1 order preference in informal spoken Dutch.

While the definitions of Meyer & Weerman (2016) and De Sutter (2007) are comparable, both relating to ease of processing, these two approaches lead to opposite predictions regarding the default cluster order. De Sutter’s hypothesis implies that a rather high-level process, sociostylistic encoding, is the main factor in deciding between these word order variations, a choice that speakers do not seem to be aware of. While this is possible, we believe that an explanation in terms of more basic cognitive mechanisms would be more plausible, such as an entrenched 1-2 order suggested by the acquisition evidence and the work on cross-serial dependencies. Therefore, in addition to searching for factors of processing complexity, we will also compare our observations to these two default order hypotheses. There is of course a third possibility, that there is no default or easier to process verb cluster word order at all. This would be a possible conclusion of the study if we find that some factors support the default 1-2 order hypothesis and others support the default 2-1 order hypothesis, i.e. if the ‘easier’ conditions of the different factors do not clearly correlate with one of the two orders. In this case, there would be no evidence that one particular order is easier to process.

4.4 Method and data

To be able to discuss the effect that different factors have on verbal cluster order variation, we have created a multifactorial model of the variation from written corpus data. We model verbal cluster order as a binary variable, in which the orders can be 1-2 (ascending) or 2-1 (descending). This is the dependent variable. This variable is modeled in terms of the factors of the multivariate model, the independent variables, which will be discussed in the next section. In this, we follow the methodology of Bloem et al. (2014), but with some additional or differently operationalized variables, and an expanded data set.

Verbal cluster order variation is one of many language variation phenomena where it has been established that multiple factors contribute significantly to the observed variation. Such phenomena are best studied using multifactorial models, rather than testing each factor one by one. The statistical power of a multifactorial model is greater. Testing each factor while applying the necessary
corrections for running multiple tests, also increases the chance of type II errors, a failure to reject a false null hypothesis. Starting with Gries (2001), such multifactorial models have been successfully used in the study of language variation. These models allow the researcher to examine how much a factor contributes to choosing one construction, while controlling for the other factors. In this way, the possible causes of a particular variation can be quantified in probabilistic terms, and the relative importance of each factor can be seen. Being corpus studies, these quantitative studies generally also emphasize evidence from larger samples of language and operationalize their factors in a precise way.

To obtain a large sample of verbal clusters, we used the Lassy Large corpus (van Noord, 2009) as our source of automatically annotated Dutch language data. This is a corpus of written Dutch that has been automatically annotated with full syntactic dependency trees by 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 use of this corpus limits the factors that can be investigated to those available in the annotation scheme of the corpus, which excludes accent and rhythm-related factors for example, as there is no prosodic annotation. From the corpus, we obtained the verbal cluster constructions of interest by defining them at the level of dependency syntax, extracting all matching constructions. This process avoids subjectivity, beyond the definition of the construction.

There are two main reasons why we chose to study this phenomenon using such a large corpus. Firstly, we are investigating probabilistic effects. In Dutch verbal cluster order variation, almost anything goes, but some combinations of factors may go together better than others. Probabilities and probabilistic associations are more accurately estimated when there is more data. Secondly, a large corpus allows us to control for information structure. Information-structural factors, such as givenness, are an important class of factors that are often invoked when explaining variation phenomena. However, they are not easily annotated, especially not automatically. Because they are not included in most corpora, the next best thing is to control for them. One way to do that is by having a very large corpus, in which many instances of the construction of interest will occur, in many different contexts with many different information-structural properties. We expect this to balance out the effects of these properties.

We extracted verbal clusters from the Wikipedia part of the Lassy Large 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, 827,709 two-verb verbal clusters were extracted, and 75.03% of these are in the 1-2 order. We chose to use this part of the Lassy Large corpus as it can be taken to represent ‘average’ standard Dutch. As discussed earlier, there is clear regional variation in verb cluster order preference, so this needs to be taken into account when choosing a corpus. While this corpus

---

1A longer discussion on using automatically annotated corpora to study verbal cluster variation and an overview of some linguistic variation studies using such corpora is provided by Bloem et al. (2014).
4.5. Measures of processing complexity

In this section, we will examine some factors that are empirically measurable for verbal cluster constructions, discuss how they relate to processing complexity, and how we operationalize them in our corpus study. Most of the factors have been discussed in previous literature as being relevant to verbal cluster order, or to other word order variation phenomena. Furthermore, most of the factors were already tested for associations with verbal cluster order by De Sutter (2009) and validated by Bloem et al. (2014). Only factors for which the values could be automatically extracted from a syntactically annotated treebank were selected.

4.5 Measures of processing complexity

In this section, we will examine some factors that are empirically measurable for verbal cluster constructions, discuss how they relate to processing complexity, and how we operationalize them in our corpus study. Most of the factors have been discussed in previous literature as being relevant to verbal cluster order, or to other word order variation phenomena. Furthermore, most of the factors were already tested for associations with verbal cluster order by De Sutter (2009) and validated by Bloem et al. (2014). Only factors for which the values could be automatically extracted from a syntactically annotated treebank were selected.
Verbal cluster order and processing complexity

We do not aim to provide a complete overview of studies linking these factors to verb cluster order variation — such background can be found in De Sutter (2005) for most of the factors. Rather, our focus here is on establishing connections between these factors and the notion of processing complexity by discussing empirical, mostly psycholinguistic work in which the factors, or similar factors, are discussed.

Processing complexity refers to the amount of cognitive resources or cognitive effort required to produce or comprehend an utterance. Processing complexity can be indirectly quantified using measures from psycholinguistics, such as reaction time or reading time. Faster times indicate faster processing, which indicates lower complexity. Evidence for ease or complexity of processing ideally comes from such psycholinguistic studies. However, there isn’t always evidence from psycholinguistic studies specifically target at a factor of interest, so we also discuss theoretical work that may predict higher or lower processing effort for particular factors. Examples of such theoretical work are the resource-limitation and constraint satisfaction models discussed in section 4.2.2. These models make predictions regarding the complexity of processing certain kinds of constructions.

Effects of cognitive effort can be found throughout different domains of language. In the domain of pragmatics and discourse, information structure seems to affect the processing efficiency of utterances. For example, reaction times to tasks involving words that are in focus, are faster. Wang et al. (2014) summarize various processing effects in this domain. Syntactic structure tends to be more explicitly marked in contexts that are more difficult to process, which is known as the complexity principle (Rohdenburg, 1996; Pijpops et al., 2018). In morphology, morphological complexity has been used as a basis for constraints on possible combinations of morphemes, and has also been related to processing cost (Plag & Baayen, 2009). And in the lexicon, lexical access speed can be seen as an indication of complexity, for example in the theory of Gibson (1998).

There are many models of processing complexity, focusing on different aspects of language processing. Our summary in section 4.2.2 discussed only a few. We will not define processing complexity in terms of any particular model of sentence processing, because as we discussed in the previous section, verbal cluster variation has already been shown to be influenced by factors related to various domains of linguistics. We are not aware of any single model of processing complexity that covers all of these domains, from word frequency to discourse integration, and there is also the aforementioned problem that most work has been done on interconstituent alternations, while we are dealing with an intraconstituent alternation. Instead, we will start from these factors which were previously identified, and some additional ones commonly used in grammatical variation studies that relate to complexity, and discuss studies that empirically measure processing complexity in their domain.
Syntactic priming  Priming is the idea that units of language get activated when they are perceived or used, and some traces of this activation remain for a while, making easier and/or faster to re-activate them. The phenomenon was first considered at the level of words, but was later applied to other linguistic units as well. This effect is often measured in a priming task, where participants will have a faster reaction to a word or pattern that they have seen before, even if they did not consciously perceive it (as in the case of masked priming).

A syntactic priming effect was first tested experimentally by Bock (1986). In a primed sentence production task, she found that particular syntactic forms were more likely to be used when that form had been perceived in the priming sentence, even though other connections between the prime and the sentence to be uttered were controlled for and minimized. The two alternations tested in the experiment were transitives in active and passive voice, and the dative alternation. Syntactic priming for the same constructions was later found in Dutch as well (Hartsuiker & Kolk, 1998).

A large number of studies have been conducted to show that this priming effect is not lexical, but rather syntactic or constructional. It cannot be entirely attributed to other factors such as priming of thematic roles (see Pickering & Ferreira 2008 for an overview of evidence). Syntactic priming in verbal clusters was studied by Hartsuiker & Westenberg (2000), who showed that verbal cluster orders can be primed. They chose verbal cluster constructions to try to exclude any effects from semantics from their study, an effect they call ‘conceptual priming’. They make the assumption that the order variation is meaningless because the auxiliary verb of the cluster does not have intrinsic meaning. However, some semantic associations with verbal cluster order have been found (Bloem, 2016b), and it can be argued that meaning arises from such associations. In addition, if phenomena like the past tense are viewed as a construction, auxiliary verbs express this meaning.

Nevertheless, their experiment showed that verbal cluster order is affected by a priming effect in spoken language. Later, De Sutter (2005) also observed such an effect in a corpus of Dutch newspaper text, though he discusses the possibility that it was a stylistic choice instead. He did not find a significant effect of priming distance, meaning how much text there was between the first instance of a cluster, the prime, and the second cluster. Such an effect might be expected due to gradual decay of the level of activation of the construction, though it has been observed in structural priming studies that activation from priming can last for a longer time. Bock & Griffin (2000) found no difference in the strength of priming between conditions where the prime and the target were separated by 0, 1, or 10 sentences, suggesting no decline within this timeframe. Nevertheless, De Sutter (2007) notes that this syntactic persistence effect he found in the newspaper corpus can be linked to processing ease, as re-using a recently activated pattern can reduce processing costs.

We consider priming as a factor related to complexity, because verbal clusters with a primed word order are easier to process. People react faster to primed stimuli and are more likely to reproduce them. If we observe sequences of the
same verbal cluster order, a priming effect probably contributed to this, and this suggests that processing difficulty does indeed affect verbal cluster order.

In operationalizing this factor, we distinguished three cases. When a verbal cluster is the first verbal cluster of the text (a text is a Wikipedia article from the corpus), we consider it unprimed, and it is categorized as ‘none’. When a verbal cluster is not the first verbal cluster of the text, we looked back and checked whether the previous verbal cluster was in the 1-2 or 2-1 order. These are the two possible ‘primed’ states. Because Wikipedia articles may be edited by multiple people, there is no guarantee of an actual priming effect — perhaps one cluster was produced, and subsequently another cluster was inserted before it by a different author. In such a situation, one cannot claim that the second cluster was primed by the first. Unfortunately this shortcoming cannot be resolved with this data set. It would not be helpful to exclude the variable either, because the model may attribute the orders affected by priming to some other factor. Furthermore, there is independent evidence of verb cluster priming effects (Hartsuiker & Westenberg, 2000).

**Morphological structure of the main verb** The Dutch language has a variety of morphologically complex verbs. Some types of complex verbs appear to be affected by syntactic processes, thus making these verbs a complicated phenomenon on the boundary of morphology and syntax. We want to distinguish two types of complex verbs, which differ in complexity. Firstly, there are inseparable complex verbs such as *volbrengen* (to complete), where *vol* is a distinct morpheme meaning ‘full’. Secondly, there are also separable complex verbs, which contain a particle that may appear in a different place in the sentence in some contexts, such as *volhouden* (to endure, to keep up). One notable difference between them is that in the first type the stress is on the second syllable, but in the second type the stress is on the first syllable (when unseparated), the separable particle. This fact has been used to explain this factor as a rhythmic factor in previous work. However, there is a link with processing complexity as well.

The following examples illustrate the difference between the two types of complex verb:

(4.13)  Ze vragen Anne of zij de wedstrijd *volhoudt.*

They ask  Anne if she the match  endures.

‘They ask Anne if she can play out the match.’

(4.14)  *Anne volhoudt de wedstrijd.*

(4.15)  *Anne houdt de wedstrijd vol.*

Anne endures the match  PART.

‘Anne endures the match.’
4.5. Measures of processing complexity

(4.16) Ze vragen Anne of zij de opdracht volbrengt.
They ask Anne if she the assignment completes.
‘They ask Anne if she will complete the assignment.’

(4.17) Anne volbrengt de opdracht.
Anne completes the assignment.
‘Anne completes the assignment.’

(4.18) * Anne brengt de opdracht vol.

In Dutch subordinate clauses, the finite verb appears at the end, with its particle directly before it (written together in Dutch orthography), as in example 4.13. In main clauses, the finite verb has to appear in second position due to the V2 effect, while the particle still appears at the end, as in 4.15. Having the particle in second position as well is ungrammatical (4.14). Inseparable complex verbs show the opposite behavior — leaving the morpheme in final position is ungrammatical (4.18). This factor has been associated with verb cluster ordering for a long time, i.e. by Meeussen & Vanacker (1951), who claims that separable verbs only occur in the 2-1 order when there is no stress previous to it in the sentence.

According to Booij (2010), inseparable complex verbs should be considered as words, while separable complex verbs are neither words nor phrases. The principle of Lexical Integrity states that syntactic rules cannot operate on parts of words, therefore separable verbs cannot be words. Yet these verbs also have word-like properties — in verbal clusters, they may occur in unseparated form and must have undergone some syntactic process as a syntactic unit. Their word stress patterns also follow those of single words (Booij, 2002). Booij considers them to be something in between words and phrases, and calls them ‘constructional idioms’. Similar claims have been made about particle verbs in German — Müller (2002) argues that they are syntactic combinations, not just morphological objects.

Particle verbs such as *afwassen* (to wash up) are one type of separable complex verb, which also can also be found in English. Other types of separable verbs include adjectives, e.g. *schoonmaken* (clean-make), nouns, e.g. *pianospel* (piano-play), or even morphemes that do not occur independently (any more), e.g. *teleurstellen* (disappoint) (Booij, 2002). We group all of these under the class of separable complex verbs. Separable complex verbs should be more difficult to process than other types of verbs, complex or non-complex, due to their phrase-like properties. Inseparable complex verbs behave like simplex verbs syntactically — besides being inseparable, they also cannot have inflectional markers intervening between the prefix and the verb root (Behrens, 1998, p. 704). While both separable and inseparable complex verbs may need to undergo morphological processing, only the phrase-like separable verbs need to undergo syntactic processing as well. Using evidence from priming studies, it has been argued that German complex verbs are morphologically decomposed
in processing (Lüttmann et al., 2011; Smolka et al., 2015). Lüttmann et al. (2011) note that this is the case both for separable and inseparable complex verbs. However, these priming studies do not present the verbs in contexts in which they might be separated, presenting only single words or word pairs, and therefore these psycholinguistic studies do not tell us whether separable complex verbs are more difficult to process than nonseparable complex verbs. We are not aware of any study that tests this directly. However, we conjecture that processing complexity is higher in the case where verbs are complex and separable.

In our study, this factor is operationalized such that any cluster where the main verb is annotated as a particle verb in the corpus, or any cluster with a separate particle in it, is considered a cluster with a separable verb, while others are inseparable.

**Length of the middle field** The linear distance between a syntactic head and its complement is considered to be a source of processing complexity, particularly in resource-limitation models of language processing. In these models, resolving a dependency between a head and a complement comes with an ‘integration cost’, and this cost depends on the length of the dependency (Gibson, 1998). We therefore consider the length of the middle field of the clause that the verbal cluster occurs in, to be a measure of processing complexity.

We follow the definition of De Sutter (2005, p. 153-154) in saying that the length of the middle field is the number of words between the connective of the start of the clause (or if there is none, just the start of the clause) and the verbal cluster, not counting elements such as particles or infinitival markers at the start of the cluster. De Sutter (2005) interprets this factor as word stress related: a longer middle field increases the chance of an accented morpheme appearing directly before the verbal cluster, causing it to clash with the accented syllable in the main verb. Speakers would then prefer putting the main verb in the final position to avoid this clash, causing 1-2 orders.

De Sutter (2007) discusses this factor in terms of processing, and interprets it in terms of a probabilistic model of processing: a longer middle field is more likely to provide specific cues on the nature of the upcoming main verb, leading to a more accurate prediction and a lower processing cost of the main verb. However, an opposite prediction can be made in terms of resource-limitation models of language processing: a longer middle field is more likely to lead to long dependency lengths and more dependencies between various parts of the middle field and the verb cluster, leading to higher integration costs when those dependencies are resolved and attached to the verb cluster head (or main verb) when the verb cluster is processed. The larger number of words and constructions that need to be processed in a longer middle field should also add to the processing cost. It is also not necessarily the case that longer middle fields allow better prediction of the main verb — if there is just a single word in the middle field that has a high association strength with the main verb, this
will be more informative than a large middle field that has a low association strength with the main verb. This is something that could be empirically tested with a statistical language model, but we are not aware of any such work. The idea was not verified empirically by De Sutter (2007).

We conclude that the length of the middle field is a measurement of complexity, and we will assume that longer sentences cost more resources to process.

In our study, this factor is operationalized by taking the position in the sentence of the first verb in the cluster, and subtracting the position of the start of the subordinate clause.

Structural depth The processing complexity of sentences is often linked to their syntactic structure. The notion of center-embedding has been essential in discussions on the computational complexity of human language, starting with the claim that English is context-free and computationally a non-regular language, because it contains center-embedded structures that cannot be produced by a regular language (Chomsky, 1957). Limits on the complexity of these structures in corpus data were attributed to working memory limitations.

We consider the structural depth of a verb cluster to be the number of nodes in the syntactic tree of the sentence between the main verb of the verbal cluster and the head of the entire sentence. This corresponds to the notion of syntactic depth discussed by Yngve (1960), and it is similar to the notion of ‘structural distance’ (O’Grady, 1997), except that structural distances are measured between two particular words or constituents, rather than to the root node of a syntactic tree.

This factor strongly depends on the dependency grammar formalism used in the Lassy Large corpus, where each head-complement relation between constituents generally constitutes a level of depth. The dependency analyses used in this corpus do not necessarily correspond to linear word order, i.e. an extraposed constituent and a non-extraposed constituent may have the same depth because they have the same head-complement relations.

In the processing literature, this factor is particularly important in constraint satisfaction models of language processing. While resource-limitation models focus more on linear distance, constraint satisfaction models focus more on expectation and prediction, often taking hierarchical structure into account to determine whether something is expected in a given context. It can therefore also be considered a measure of processing complexity.

Of course, the importance of this measure depends on how much of the apparent hierarchical syntactic structure of sentences is cognitively real. It has also been argued that language is not (or hardly ever) processed hierarchically, but rather sequentially. Frank et al. (2012) outline such a model, citing evidence from cognitive neuroscience, psycholinguistics and computational models. In this model, linguistic input is processed as a stream of components that have a linear order, rather than a tree of constituents. Depth is therefore not a relevant notion in this model, and cannot be related to processing effects. In
this view, deeper structures would not lead to a higher processing load. On the basis of generative theories of language, where embedding, caused by recursive processes, is a basic property of grammar, this point can also be argued. In this view, deeper syntactic structures do not necessarily lead to a higher processing load, unless some kind of extrinsic, non-linguistic performance limit is hit and grammatical processing breaks down, as initially suggested by Miller & Chomsky (1963). From this, and from the observation that recursion is central to many human cognitive processes, it follows that more verb clusters that are deeper in the syntactic structure would not be more difficult to process. Despite these possible objections, we will for now assume that processing deeper structures is more complex.

In our study, this factor is operationalized by counting how far down in the syntactic tree of the entire sentence the head verb of the cluster is located. Each branching point of the tree is counted as one level. The trees of the Lassy Large corpus use a dependency tree format which is not binary-branching (van Noord, 2009). A visualized example of such a tree can be found in van Noord (2009, p. 119).

Information value of the last preverbal word  This factor concerns the word class of the word that is used directly before the verbal cluster. The main reason why we chose to test only the last word before the verbal cluster is to allow comparisons to De Sutter (2005), who did the same thing, but it may also be interesting for reasons of processing. Some late processing may take place on this word, affecting verbal cluster processing as well, particularly when function words are involved. In an ERP study, Hoen & Dominey (2000) found a left anterior negativity (indicating working memory load) between 400 and 700 ms after presentation of a function word. This was hypothesized to be an effect of predicting the next constituent, which is easier after a function word.

De Sutter (2005) found that the information value of the last preverbal word affects the order of the verbal cluster, linking it to prosody. De Sutter (2007) linked it to probabilistic processing, in the sense that a more informative word before the verb cluster will make prediction of the main verb, and therefore processing, easier. We do not see a particular reason why a highly informative word (such as a noun) before the cluster facilitates prediction more than a highly informative word earlier in the clause, i.e. an animate subject noun that might only occur with certain main verbs. De Sutter (2007) provides the following examples, where 4.19 shows a preverbal word with a high information value (a noun) and 4.20 shows a preverbal word with medium information value (an adverb):

\[(4.19) \quad \ldots \text{dat grote stukken landbouwgrond en infrastructuurwerken zijn} \]
\[
\text{... that large tracts of farmland and infrastructure were} \]
\[
\text{verwoest} \]
\[
\text{destroyed} \]
\[
\text{‘... that large tracts of farmland and infrastructure have been destroyed.’} \]
4.5. Measures of processing complexity

De Sutter (2007, Ex. 29)

(4.20)  
\[ \text{... dat ze al gebruikt zijn} \]
\[ \text{... that they already used were} \]

‘... that they have already been used.’ De Sutter (2007, Ex. 30)

However, the following constructed combination of those sentences is also perfectly normal:

(4.21)  
\[ \text{... dat grote stukken landbouwgrond en infrastructuurwerken al zijn verwoest} \]
\[ \text{... that large tracts of farmland and infrastructure have already been destroyed} \]

Even though the last preverbal word has medium information value here, we would argue that the clause is more informative than example 4.19, because it contains more information overall. Therefore, an observed effect of information value of the preverbal word does not necessarily mean the verb cluster was more predictable — it could have been predicted by something that came earlier in the clause, too.

Alternatively, this factor can be interpreted in terms of processing of the preverbal word itself. If that particular word requires more processing, or late processing because a functional projection needs to be created, this may affect the processing of the verb cluster. If the preverbal word is difficult to process, there may be fewer resources available to process the verb cluster. However, the literature is inconclusive on whether high-informative open-class words or low-informative function words are more difficult to process.

In psycholinguistics, differences have been observed in lexical access to open-class words (nouns, verbs) and closed-class words (function words). Bradley (1978) described a difference between the two categories: lexical decision response times to open-class words depend on their frequency, while those to closed-class words do not. This was taken as evidence for different processing mechanisms. More recently, Segalowitz & Lane (2000) argue against this model and the lexical decision methodology, and show that function words are accessed more quickly in general. This appears to be an unresolved issue, so while we can clearly state that this is a factor relating to processing complexity, it is not so clear which of the factor’s levels is actually the more complex one.

We operationalize this factor in the same way as De Sutter (2007) and Bloem et al. (2014), in terms of three classes: high informational (nouns, verbs, numerals), intermediate informational (adjectives and adverbs) and low informational (pronouns, conjunctions and prepositions). The word class is taken from the annotation of the corpus.
We also investigate another property that these preverbal constituents can have: definiteness. This factor has been discussed in previous work, but has not been included in a multifactorial model before. De Sutter et al. (2007) viewed it as another measure of information weight, closely related to the aforementioned ‘information value’ factor. In our view, this property can also be related to processing rather than ‘information value’, as it indicates discourse-givenness of the referent. However, it only applies to noun phrases. In this work, we will not directly test givenness since this is very difficult to determine using automatic annotation methods. Givenness involves long-distance referencing in the discourse, across sentence or even paragraph boundaries, and we are not aware of any reliable automatic method for detecting it in a large corpus.

A definite noun phrase has usually been previously given in the discourse and is assumed to be already accessible, facilitating processing. Givenness also leads to phonological reduction, a further indication that given items are easier to process. However, it is often assumed that definite constituents have to be linked to a previously given discourse element, while indefinite constituents do not (e.g. by Heim 1982). This may be a complex process. Crain & Steedman (1985)’s discourse-based theory of NP attachment makes predictions along these lines, stating that indefinite NPs should be processed more easily because no additional sets of discourse objects are presupposed. This idea was later experimentally confirmed by Gibson et al. (1996), although this study only used a questionnaire, rather than a psycholinguistic measure.

Furthermore, studies of impaired populations of speakers suggest that definites are more difficult to process: Polite et al. (2011) tested English-speaking children with specific language impairment (SLI) and showed that they produce fewer definite articles than typically developing children in definite contexts, while there was no significant difference in indefinite article use. Testing Dutch children, Schaeffer et al. (2014) find a strong effect in this direction in children with High Functioning Autism (HFA), as well as those with SLI. Experimental results by Schaeffer et al. (2018) show that the difficulty that SLI and HFA children have, have different underlying causes: for HFA children, the issue appears to be related to a pragmatic impairment, while for the SLI children, who perform worse on tests of morphosyntax and working memory, it is a problem in maintaining discourse continuity. While the problems that HFA children have may not be indicative of general processing difficulties, but instead a pragmatic impairment, the findings for SLI children are relevant. These findings show that discourse integration, a higher-level cognitive process, takes place in the processing of definites, and that a population with fewer cognitive resources in the linguistic and working memory domains (children with SLI) has particular problems with it. We can therefore take this as evidence that definites are more difficult to process than indefinites.

In the present study, this factor is operationalized with three possible levels: a preverbal word is considered definite when it is a noun that is in a dependency relation with a definite determiner. When it is a noun in a dependency relation...
with an indefinite determiner, it is considered indefinite, and in any other case, it is considered ‘unknown’.

**Frequency of the main verb** Psycholinguistic theories generally assume that more frequent words are more easily activated, making them easier to process. This assumption can be found in the theory of Gibson (1998) mentioned in the section on structural depth: “high frequency lexical items require fewer energy resources to become activated than low frequency lexical items do”. Therefore, more frequent words are less complex to process. The frequency of the main verb is operationalized by counting the frequency of the verb lemma in the entire corpus (not only in verb clusters), and taking its logarithm.

**Extraposition** Extraposition of the prepositional object, also called PP-over-V, is another source of variation that can occur around verbal clusters in Dutch, and one that may be linked to complexity as well. Even though Dutch embedded clauses are normally verb-final, when the main verb of the verbal cluster has a prepositional object, the object can optionally be positioned after the verbal cluster:

(4.22) ... *om medeklinkerclusters aan het begin van woorden te vermijden*. ... REL consonant clusters at the start of words to avoid.

‘to avoid consonant clusters at the start of words.’ [wik_part0694/1649193-9-2]

(4.23) ... *wat werd veroverd door Hettieten*. ... that was conquered by Hittites.

‘that was conquered by Hittites.’ [wik_part0010/4008-17-2]

Example 4.23 shows the extraposition. Similar to verbal cluster order variation, there is no clear meaning difference between the two orders. This type of extraposition has been linked in previous work to various complexity-related factors that we also discussed in this study: inheritance, morphological complexity, definiteness and heaviness (length), as well as habituality, information pattern, intonation pattern and register (Jansen, 1978; Canegem-Ardijns, 2006). Willems & De Sutter (2015) performed a multivariate corpus study of PP placement studying the effect of the weight of the middle field, the postfield and the extraposed PP itself. They found that the weight of the PP is the main factor, and that extraposition into heavy postfields is less likely. Weight was operationalized in terms of length and complexity (embedded structures), similar to our factors length of the middle field and structural depth. This shows that complexity affects PP extraposition in some ways as well, and that the order variation caused by extraposition may be comparable to verbal cluster order variation. Willems & De Sutter (2015) state that their model requires more predictor variables because its performance leaves room for improvement,
perhaps some of this remaining variation can be accounted for using other factors of processing complexity.

One interesting observation by Willems & De Sutter (2015) is that they find extraposition to be more common than non-extraposition. They find 58.4% extraposition based on 1,850 clauses from the Dutch Parallel Corpus of journalistic texts. If extraposition is such a common process, it might be expected that its processing difficulty is not very high — if it was difficult, the alternative of non-extraposition would be used instead. Furthermore, a commonly used construction is likely to be more entrenched, and according to constraint satisfaction processing models, more predictable and therefore easier to process.

These studies show that extraposition is used to make a subordinate clause less difficult to process, but perhaps a direct link between extraposition and ease of processing can be found. We are not aware of any psycholinguistic study of PP-over-V extraposition, but a study by Levy & Keller (2013) on processing German verb-final structures may provide some insight. German is verb-final and has verbal clusters that are theorized to have the same underlying mechanisms as Dutch verbal clusters. What Levy & Keller (2013) found is that the verb is more difficult to process when both an adjunct and a dative NP are present in a relative clause, compared to conditions where only one of them was present. Importantly, they only tested verb-final clauses in their relative clause experiment (Experiment 2), and controlled for sentence length. They did not test other subordinate clauses besides relative clauses.

In Dutch, PP adjuncts and dative NPs are elements that can be extraposed. If the processing difficulty observed by Levy & Keller (2013) is due to the presence of arguments before the verb, we would expect that subordinate clauses with extraposition are easier to process than those without. When there is extraposition, arguments that would have been before the verb come after the verb instead. Levy & Keller (2013) did not have a condition with arguments after the verb, as this is not possible in German. Nevertheless, if we assume that Levy & Keller’s (2013) findings also apply to Dutch, this would match what van Haeringen (1956) already hypothesized. He claimed that extraposition may alleviate some of the ‘tension’ involved with not producing the verb until the end of the sentence, when the sentence is long. It would also help explain why extraposition was found to be associated with various complexity-related variables, e.g. by Willems & De Sutter (2015).

Levy & Keller’s (2013) study discussed PP adjuncts, but in Dutch, PP complements can also be extraposed. When De Sutter (2009) tested the effects on extraposition on verbal cluster order, he included a distinction between extraposition of adjuncts and extraposition of complements. De Sutter (2007) argues that complements are more costly to keep in memory than adjuncts due to their higher information weight, making a complement more difficult to process when extraposed — it has to be remembered for longer. One might also expect a tendency for the main verb and the complement to be closer together, following the adjacency principle. Complements are more strongly related to the main verb semantically, which might cause additional processing
4.5. Measures of processing complexity

difficulty if there is a great distance between them. For adjuncts, this is not the case. Example 4.23 shows an extraposed adjunct, example 4.24 contains an extraposed complement, as evidenced by the fact that the extraposed constituent is required to complete the meaning of the verb.

(4.24) ... *dat* hij *zijn lievelingspaard Incitatus wenste aan * te stellen
    ...
that he his favourite horse Incitatus wished PART to appoint
    *als consul*.
    as consul.

‘that he wished to appoint his favourite horse Incitatus as consul.’

In summary, Levy & Keller’s (2013) findings are the most closely related psycholinguistic evidence that we could find, so we will take this to mean that processing verbal clusters with extraposed elements is easier than processing clusters without extraposition. However, it should be noted that the evidence comes from a somewhat different context.

The extraposition factor is operationalized as follows. When the verb cluster is the last element in the linear order of the sentence, we automatically label this as ‘nothing’. When an adjunct follows the verb cluster in the clause, we consider this an extraposed adjunct. When a complement follows the verb cluster, this is considered an extraposed complement. When there is some other material after the verb cluster, such as another subordinate clause or part of the main clause, this is labeled as ‘Other continuation’. When there is punctuation after the verb cluster, followed by some other material, this is considered ‘Punctuation and continuation’. These distinctions are mainly made for technical reasons.

Multi-word units  A multi-word unit (MWU) is a lexical unit, consisting of multiple words that together carry a different meaning than that of its constituent parts. They are also known as formulaic sequences. Collocations and idioms are types of multi-word units. Evidence from psycholinguistic studies, such as the self-paced reading experiment of Conklin & Schmitt (2008) and the eye-movement study of Underwood et al. (2004), suggests that MWUs are processed faster and remembered better, even when they span across syntactic boundaries (Tremblay et al., 2011). However, there are a few things to consider when multi-word units include (parts of) verbal clusters.

In his thesis on verbal cluster order, de Sutter discussed the similar concept of ‘inherence’, preverbal constituents that have a strong semantic connection with the main verb. In his 2005 study, he hypothesizes an association between this phenomenon and the 1-2 order, for reasons of prosody. An inherent preverbal constituent is assumed to have an accent, and the main verb has an accent too. The main verb would then be more likely to be uttered in the final position, to avoid disrupting the sentence rhythm. He finds that this is not true for regular verb phrases that fit the definition of inherence, but that there is an effect for fixed expressions, which he defines as any collocation between a verb and one or more non-verbs in the verb phrase. Collocations are defined using
log-likelihood ratio tests, using statistical significance at \( \alpha = 0.05 \) as a cutoff point for collocation-hood, and only collocations between two words (one verb and one non-verb) were tested. De Sutter (2007) also discusses this factor, again interpreting it in probabilistic processing terms — hearing the first part of a multi-word unit before a cluster will make it easier to predict what the main verb of the cluster will be.

We also interpret this finding as a complexity-related one, given the clear evidence for a relation between MWUs and processing difficulty discussed earlier. In this study, we will test several types of MWU constructions involving verbal clusters for associations with the 1-2 or 2-1 orders. Firstly, we can make a distinction between constructions with extraposable PP arguments and those without. MWU constructions without a prepositional object involve at least one verb of the cluster, and optionally something before the cluster:

\[
\text{(4.25) Suske en Wiske willen hem helpen met de campagne, om zo ook een oogje in het zeil te kunnen houden.}
\]

‘Suske and Wiske want to help him with the campaign, to be able to keep an eye on things too.’ [wik_part0869/2279152-4-5]

MWU constructions can also take a prepositional phrase as their object, which is not part of the multi-word unit:

\[
\text{(4.26) Omdat Neerlinter openlijk partij had gekozen tegen de bisschop van Luik, was de gemeente geëxcommuniceerd.}
\]

‘Because Neerlinter had openly sided against the bishop of Luik, the township was excommunicated.’ [wik_part0832/2143613-6-3]

In this example, the prepositional object \( \text{de bisschop van Luik} \) is an open slot. It could be replaced with a wide variety of things that can be prepositional objects, such as \( \text{zijn moeder} \) (his mother), a conjunction like \( \text{zowel Nederlandstaligen als Franstaligen} \) (both the Dutch and French speaking) or even other multiword units such as \( \text{alles en iedereen} \) (everything and everyone).

Because these constructions have a prepositional object, extraposition as discussed in the previous section can also take place here. Example 4.26 is in fact an example of it. Both of these options are grammatical:

\[
\text{(4.27) ... dat hij partij had gekozen tegen de bisschop van Luik}
\]
\[
\text{(4.28) ... dat hij tegen de bisschop van Luik partij had gekozen}
\]

We have included this distinction in the model as well for the sake of completeness, and in order to find possible effects of continuous and discontinuous sequences.
In summary, we assume that clusters with MWUs and transitive verbs, where there is no prepositional object to extrapose, are easier to process. When there is a prepositional object, we expect more processing difficulty since some syntactic processing will have to take place.

The LASSY Large corpus includes annotation of multi-word units, so we were able to extract this information without having to compute which words form multi-word units. The way multi-word units are annotated in the corpus is described by Odijk (2013). The factor is operationalized based on this information from the corpus. If one of the verbs was annotated as a multi-word part, it was categorized as a case of **multi-word units**, with different values marking where the rest of the multi-word expression went (before the cluster, after the cluster, or there was no non-verbal component).

### 4.6 Results

In this section, we present two types of results. Firstly, Table 4.1 lists all of the explanatory variables (predictors) used in the model, along with their effect size. The model uses these variables to predict the binary dependent variable of verbal cluster word order, which can have 2 values: 1-2 or 2-1 order. Secondly, Table 4.2 presents the same variables, sorted by information gain. This allows us to see which of the variables contribute the most towards statistically explaining the observed variation. Unless noted otherwise, all of our factors account for differences significant at the $\alpha = 0.001$ level. When working with huge, automatically-annotated corpora, sample sizes are so large that p-values become practically meaningless and even effects with a very small size easily become statistically significant, as P-values depend on sample size. We therefore choose to focus on effect sizes and information gain instead.

Table 4.1 allows us to explore the effect of each factor on verbal cluster order. Most of the explanatory variables are categorical: they have several possible values which are not necessarily ordered. For example, the variable **type of auxiliary** has different (categories of) verbs as possible values, such as grouping verbs and modal verbs, but we cannot say that grouping verbs are ‘greater’ or ‘smaller’ than modal verbs. For these variables, one of the possible values was taken as the reference value, or baseline, to which all the other values are compared. For the variable **type of auxiliary**, the auxiliary verb *zijn* was taken as the baseline. The odds ratio for this baseline category is therefore always 1.00, and this is indicated in bold in the table. As an example, the effect size listed for the verb *hebben* then indicates the odds of a 1-2 order in case there is the auxiliary verb *hebben* instead of *zijn*, all other conditions being the same. The effect size of each variable is given as an odds ratio. An odds ratio further from 1 in either direction indicates a stronger effect.

For example, these odds ratios show that in our data, *hebben*-clusters are two times more likely to be in the 1-2 order than *zijn*-clusters, while controlling for the other factors. Note that this also implies the reverse: *zijn*-clusters are two
times more likely to be in the 2-1 order than *hebben*-clusters, because there are only two possible orders. It is simply a matter of which baseline we choose and which order (1-2 or 2-1) we decided to report on. The baselines were selected to be either the same as those in previous work, the most frequent value, or the most basic value (i.e. ‘none’). We will be reporting values for the 1-2 order (ascending or participle-final), because De Sutter (2009) also did this. As a result, odds ratios > 1 indicate an effect of association with the 1-2 word order, odds ratios < 1 indicate 2-1 order.

The variables main verb log-frequency, structural depth and length of middle field are not categorical, but continuous. In these cases, an odds ratio cannot be interpreted in the same way. Instead, we show the standardized $\beta$ estimate, a different measure of effect size. It represents an effect on the dependent variable, cluster order, in terms of standard deviations.

Table 4.2 gives us a ranking of the variables, in order of their contribution to the model. This is expressed as an Akaike Information Criterion (AIC) value, which is a measure of information loss. The model with the lowest AIC loses the least information, compared to the data it models, and accounts for the most variance.

The ranking was produced through stepwise regression. In this procedure, we start with a model with no independent variables (explaining nothing). Such a model has the maximum possible information loss (AIC). We then calculate which of the variables, when added to the model, would cause the greatest decrease in AIC. This variable will be ranked 1, and added to the model. Then we repeat the procedure. Out of all the remaining variables, we calculate which one would cause the greatest decrease in information loss, and add it. This variable will be ranked 2. We repeat the procedure until we ran out of variables, or the remaining variables do not decrease the information loss any more. In our case, we ran out of variables — all of the variables contribute to the model.

We will start with some general observations. The 1-2 order is more frequent overall in our dataset. Even if we take only the clusters with the auxiliary verb ‘zijn’ (to be), which are relatively more likely to be in the 2-1 order, the 1-2 order is still the most common in terms of absolute frequency, with 57.39% of such clusters occurring in the 1-2 order for this construction. The reason for this relative preference is that the verb ‘zijn’ commonly functions as a copula verb, in which case the 2-1 order is more likely (according to De Sutter 2009) or even required (according to prescriptive grammars).

In the remainder of this section, we discuss the effect sizes of all the factors from section 4.5, listed in Table 4.1.

Firstly, there are two factors included in the model that were included as controls, and we will not focus on them here: type of auxiliary and ‘te’-infinitive. While these factors are associated with verbal cluster order variation, and therefore should be included to have an accurate model of verbal cluster variation, they have a clear meaning component, and can be considered to be different constructions on the basis of that. They control for the different verbal cluster constructions discussed in section 4.2.3 — type of
<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Odds ratio</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of auxiliary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zijn</strong></td>
<td>1.00</td>
<td>191.811</td>
<td></td>
</tr>
<tr>
<td><strong>worden</strong></td>
<td>1.47</td>
<td>277.531</td>
<td></td>
</tr>
<tr>
<td><strong>hebben</strong></td>
<td>2.00</td>
<td>97.969</td>
<td></td>
</tr>
<tr>
<td>Grouping verb</td>
<td>9.65</td>
<td>55.157</td>
<td></td>
</tr>
<tr>
<td>Modal verb</td>
<td>159.27</td>
<td>189.470</td>
<td></td>
</tr>
<tr>
<td>Causal verb</td>
<td>1135.14</td>
<td>14.855</td>
<td></td>
</tr>
<tr>
<td><strong>‘te’-infinitive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1.00</td>
<td>764.312</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.29</td>
<td>63.397</td>
<td></td>
</tr>
<tr>
<td><strong>Priming</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1.00</td>
<td>228.227</td>
<td></td>
</tr>
<tr>
<td>Primed for 1-2</td>
<td>1.24</td>
<td>444.950</td>
<td></td>
</tr>
<tr>
<td>Primed for 2-1</td>
<td>0.59</td>
<td>154.482</td>
<td></td>
</tr>
<tr>
<td><strong>Morphological structure of the main verb</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inseparable</td>
<td>1.00</td>
<td>810.401</td>
<td></td>
</tr>
<tr>
<td>Separable</td>
<td>7.06</td>
<td>17.308</td>
<td></td>
</tr>
<tr>
<td><strong>Length of middle field</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per additional word</td>
<td>1.0390</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td><strong>Structural depth</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per additional level of embedding</td>
<td>0.95</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td><strong>Information value</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>416.663</td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.00</td>
<td>139.715</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.79</td>
<td>240.170</td>
<td></td>
</tr>
<tr>
<td><strong>Definiteness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indefinite</td>
<td>1.00</td>
<td>22.657</td>
<td></td>
</tr>
<tr>
<td>Definite</td>
<td>1.04</td>
<td>70.231</td>
<td></td>
</tr>
<tr>
<td>Unknown/NA</td>
<td>1.27</td>
<td>734.821</td>
<td></td>
</tr>
<tr>
<td><strong>Frequency of main verb (log)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per additional order of magnitude</td>
<td>1.06</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td><strong>Extraposition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nothing</td>
<td>1.00</td>
<td>329.211</td>
<td></td>
</tr>
<tr>
<td>Extraposed adjunct</td>
<td>0.75</td>
<td>144.673</td>
<td></td>
</tr>
<tr>
<td>Extraposed complement</td>
<td>0.65</td>
<td>75.807</td>
<td></td>
</tr>
<tr>
<td>Other continuation</td>
<td>0.98</td>
<td>139.610</td>
<td></td>
</tr>
<tr>
<td>Punct. and continuation</td>
<td>0.91</td>
<td>138.408</td>
<td></td>
</tr>
<tr>
<td><strong>Multi-word units</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1.00</td>
<td>811.588</td>
<td></td>
</tr>
<tr>
<td>PP object, before</td>
<td>1.88</td>
<td>8.373</td>
<td></td>
</tr>
<tr>
<td>PP object, after</td>
<td>2.79</td>
<td>1.225</td>
<td></td>
</tr>
<tr>
<td>Other/no object</td>
<td>0.37</td>
<td>6.523</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Effect of different variables on the likelihood of 1-2 verbal cluster orders
Table 4.2: Decrease in information loss when adding factors to the model in a stepwise regression procedure

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>AIC</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(none)</td>
<td>463279</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>Type of auxiliary</td>
<td>382538</td>
<td>80741</td>
</tr>
<tr>
<td>2</td>
<td>Priming</td>
<td>378185</td>
<td>4353</td>
</tr>
<tr>
<td>3</td>
<td>‘te’-infinitive</td>
<td>374378</td>
<td>3807</td>
</tr>
<tr>
<td>4</td>
<td>Extraposition</td>
<td>371413</td>
<td>2965</td>
</tr>
<tr>
<td>5</td>
<td>Length of middle field</td>
<td>369817</td>
<td>1596</td>
</tr>
<tr>
<td>6</td>
<td>Frequency of the main verb</td>
<td>368744</td>
<td>1073</td>
</tr>
<tr>
<td>7</td>
<td>Information value</td>
<td>367806</td>
<td>938</td>
</tr>
<tr>
<td>8</td>
<td>Morphological structure of the main verb</td>
<td>366870</td>
<td>936</td>
</tr>
<tr>
<td>9</td>
<td>Multi-word units</td>
<td>366162</td>
<td>708</td>
</tr>
<tr>
<td>10</td>
<td>Structural depth</td>
<td>365674</td>
<td>488</td>
</tr>
<tr>
<td>11</td>
<td>Definiteness</td>
<td>365461</td>
<td>213</td>
</tr>
</tbody>
</table>

**Auxiliary** refers to the type of auxiliary verb used in the construction, and ‘**TE’-INFinitive** encodes whether the construction was a *te*-infinitive construction. Following Bloem et al. (2014, p. 1979), auxiliary types with similar word order preferences and similar function are grouped. As we discussed in the introduction, it is not controversial that constructions, which are pairs of form and function, can be distinguished on the basis of their meaning. Such constructions can have different word order preferences simply for historical reasons (Van de Velde, 2017) or due to frequency effects, such as the modal verb clusters which are known to occur almost exclusively in the 1-2 order (and this is what we also found). Besides a control variable for the type of auxiliary, one might also expect a control variable for the main verb, or type of main verb. While this would probably improve the model, there were too many different main verbs in the data for the model to handle. However, effects related to the main verb have been observed (Bloem, 2016b), and therefore it would be good to find a method of incorporating these effects in future work.

In Table 4.2, we can see that the two control variables are quite important in explaining the observed variation. The factor **Type of Auxiliary** is ranked first, with a large decrease in AIC. This follows from the fact that this factor distinguishes between different constructions with different meanings, which sometimes have clear word order preferences. A large part of this can be attributed to the difference between modal clusters and other clusters — as noted, modal clusters occur in the 1-2 order 99.5% of the time, and they are frequent. This is a large deviation from the average, which is 75.03%. The factor **Type of Auxiliary** accounts for this, and is therefore an integral part of
4.6. Results

the model. Bloem et al. (2014, p. 1981) note that there are no great differences between a model with and without the modal clusters, which means that the factor is controlled for adequately.

The factor finiteness is ranked third. While it also has word order associations, te-infinitive clusters are quite rare compared to other constructions, so the overall impact of this factor is smaller.

What we are interested in here is the fact that even when we control for relevant factors, there is still a lot of variation remaining. For example, among verbal clusters with the auxiliary ‘have’ and a finite main verb, there are still clusters in both the 1-2 and 2-1 orders, and this variation can be linked to the processing complexity-related factors that we discuss below. We are more interested in these factors that cannot be explained away as semantic or lexical differences. Word order differences are less likely to be stored preferences in this case and more likely to hint at processing effects.

When using such a large multifactorial model, there is the risk that the different factors are not independent, and partially explain the same portion of the variation. For example, it might be the case that complex particle verbs are generally also less frequent. Statistically, we can test for this by computing the variance inflation factor (VIF) of the variables in the model. This is a measure of multicollinearity — correlation between predictor variables. In our model, all of the VIFs are very low (< 2). The factors therefore do not appear to correlate with each other, are independent, and do not statistically account for the same parts of the variation.

Syntactic priming The results in Table 4.1 show that 1-2 clusters are more likely to be preceded by other 1-2 clusters (1.24) and less likely to be preceded by 2-1 clusters (0.59), compared to clusters that are unprimed. Table 4.2 shows that the role of syntactic priming in explaining the observed variation is large — the factor is ranked second. This confirms the importance of processing effects in the model, as syntactic priming is a clear processing effect.

Morphological structure of the main verb In the data, separable verbs are 7.06 times more likely to occur in the 1-2 order, shown in Table 4.1. De Sutter (2009) got an effect size of 3.87 here, and Bloem et al. (2014) observed 4.92. It is a robust observation that separable verbs are more likely to occur in the 1-2 order. In a study of main verb order associations in verb clusters, it was also noted that many of the verbs that are most strongly associated with the 1-2 order are separable verbs (Bloem, 2016b).

While the effect size is large, the importance of this factor for the overall model is not as large. It is ranked eighth in terms of information gain, as shown in Table 4.2. This is because most verbs are not separable complex verbs, so the factor is relevant to a limited portion of the data.
Length of the middle field  According to Table 4.1, clusters preceded by longer middle fields are more likely to occur in the 1-2 order. This matches the result of De Sutter (2007). This is 1.039 (3.9%) times more likely per additional word in the middle field. This effect is fairly important in the overall model, ranking fifth in Table 4.2.

Structural depth  In the data, clusters that are more deeply embedded in the sentence structure are less likely to occur in the 1-2 order. From Table 4.1 we can see that the effect size is 0.95 (-5.3%) per additional level of depth. This effect is fairly small, and as a result, this is one of the factors of lesser importance to the model. It is ranked tenth in Table 4.2.

It is interesting that the effect direction here is the opposite from that of the length of the middle field. Intuitively, one might expect that clause length and depth correlate with each other: longer clauses have bigger trees. Looking into this more closely, we found that the two factors only correlate moderately, with \( r = .32 \), too little to consider one to have a confounding effect on the other. Sentences with clusters and long middle fields are not nearly always deeply embedded. This demonstrates that the two factors need separate explanations.

Definiteness of the last preverbal constituent  Table 4.1 tells us that clusters preceded by a definite constituent are 1.04 (4%) times more likely to occur in the 1-2 order. It is our only finding that is not significant at the \( \alpha = 0.001 \) level — for definite constituents compared to indefinite, \( p = 0.042 \). It then comes at no surprise that it is also ranked last in Table 4.2. From the frequencies we can see that only 11% of verbal clusters were directly preceded by a noun marked for definiteness in our study, perhaps this is not enough to draw conclusions from. This result also contrasts with De Sutter et al.’s (2007) finding, who found that indefinites are more associated with the 1-2 order, rather than definites. However, this was not tested in a multifactorial model.

Information value of the last preverbal constituent  In the data of Table 4.1, clusters that are preceded by a word of high information value (or an open-class word) are 0.79 times more likely to occur in the 1-2 order, which means they are 1.27 (0.79^{-1}) times less likely to occur in the 1-2 order, or 27% less. Interestingly, this result is different from those of De Sutter (2009) and Bloem et al. (2014), even though those studies defined the factor in the same way. The factor is ranked fairly low in Table 4.2 (7th) even though it affects all clusters; perhaps the effect is not important enough to measure accurately across different models.

Frequency of the main verb  We estimated verb frequencies by counting the number of occurrences of the verb’s root in the entire Lassy Large corpus (thereby counting the occurrence of all forms of the verb). According to the model, clusters that contain main verbs of higher frequency are more likely to
occur in the 1-2 order. The effect size is 6% per order of magnitude of frequency, so for each magnitude (e.g. 10,000 or 100,000 occurrences), the 1-2 order is 1.06 times more likely in our data, as shown in Table 4.1. It is not a large effect, but it is robust, and a similar effect was found by De Sutter (2009). In Table 4.2, this factor is ranked 6th, so it is one of the more important processing-related factors. One might think that more frequent main verbs occur in clusters with modal verbs more frequently, causing a difference, because modal verbs occur in the 1-2 order relatively more often. The observation is true: the average token frequency of main verbs in modal clusters is 340.199, while the average token frequency of main verbs in other clusters is 104.379. However, because we are using a multifactorial model, this factor is already taken into account, and should not influence the result much. We created a model of only the non-modal clusters to verify this, and the effect size was also 1.06.

**Extraposition**  In the data, clusters with extraposed constituents are generally less likely to be in the 1-2 order, 0.75 for adjuncts and 0.65 for complements (Table 4.1). In Table 4.2 this factor is ranked fourth, and therefore explains a relatively large portion of the variation.

**Multi-word units**  In the results for this factor (Table 4.1), we can see that clusters that involve a fixed expression are more likely to occur in the 1-2 order when they have a prepositional object, but less likely to occur in the 1-2 order when they do not. The information gain of the factor MWU is fairly low, ranked 9 in Table 4.2. This can be attributed in part to the low frequency of multi-word units.

4.7 Discussion

In the preceding section, we have seen that all of the factors that we have investigated contribute in some way to a model that statistically accounts for the verbal cluster order variation that we found in the data. Furthermore, in section 4.5 we have learned that these factors, apart from the control variables, can be related to processing complexity. This allows us to conclude that processing complexity clearly plays a role in this order variation.

With this in mind, we can start to address the question of what the processing difference between the two constructions might be. Do either of the two orders correlate with contexts that are difficult to process? In section 4.5, we discussed for each factor which value is the most difficult to process. For the factor INFORMATION VALUE, we could not learn from the literature whether open-class or closed-class words are more difficult to process. For the other factors, it was fairly clear which value was more difficult to process (e.g. longer sentences rather than shorter sentences, separable verbs rather than inseparable ones). There are three possible answers to the question posed here:
**Default 1-2 order hypothesis**  This is the right answer when we find the 2-1 order in contexts that are less difficult to process, and the 1-2 order in contexts that are more difficult to process. This hypothesis is supported by ideas from Meyer & Weerman’s (2016) work on acquisition of verbal clusters.

**Default 2-1 order hypothesis**  This is the right answer when we find the 1-2 order in contexts that are less difficult to process, and the 2-1 order in contexts that are more difficult to process. This hypothesis is supported by ideas from De Sutter’s (2005) thesis, where he proposes that sociostylistic processing may play a large role.

**Neither**  This is the right answer when we do not find a clear association between processing difficulty either of the word orders. There are still processing-related factors, but they do not clearly point to one order occurring in contexts that are easier to process than the other.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>Supports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Type of auxiliary</td>
<td>(control)</td>
</tr>
<tr>
<td>2</td>
<td>Priming</td>
<td>(both)</td>
</tr>
<tr>
<td>3</td>
<td>‘te’-infinitive</td>
<td>(control)</td>
</tr>
<tr>
<td>4</td>
<td>Extraposition</td>
<td>1-2</td>
</tr>
<tr>
<td>5</td>
<td>Length of middle field</td>
<td>1-2</td>
</tr>
<tr>
<td>6</td>
<td>Frequency of the main verb</td>
<td>2-1</td>
</tr>
<tr>
<td>7</td>
<td>Information value</td>
<td>(unclear)</td>
</tr>
<tr>
<td>8</td>
<td>Morphological structure of the main verb</td>
<td>1-2</td>
</tr>
<tr>
<td>9</td>
<td>Multi-word units</td>
<td>1-2</td>
</tr>
<tr>
<td>10</td>
<td>Structural depth</td>
<td>2-1</td>
</tr>
<tr>
<td>11</td>
<td>Definiteness</td>
<td>1-2</td>
</tr>
<tr>
<td>-</td>
<td>Overall word order frequencies</td>
<td>2-1</td>
</tr>
</tbody>
</table>

**Table 4.3: Summary of the factors that were tested, and which default order hypothesis they support**

In Table 4.3, we have summarized our results from the previous section with respect to the default word order hypotheses. The table lists the word order association for the more difficult to process condition of each factor (i.e. for length of the middle field, a longer middle field). For most factors, we observed that the value that is more difficult to process, is associated with the 1-2 order, thereby supporting the default 1-2 order hypothesis. However, there are a few exceptions. We will now briefly discuss each factor with respect to these hypotheses, starting with the overall frequency of the orders.

The 1-2 order is more frequent for all of the verbal cluster constructions we tested in our data, which might be seen as support for the default 1-2 order
hypothesis. However, it should be noted that this corpus only contains written data. Some studies have found that spoken Dutch contains more 2-1 orders. For example, Stroop (2009) finds 63% 2-1 orders in the Corpus of Spoken Dutch (CGN) among clusters with a participle. This goes against the default 1-2 order hypothesis, since it is generally assumed that producing spoken language involves more processing difficulty, because it is more immediate. De Sutter (2005) also notes that text genres with more production pressure include relatively more 2-1 orders. These findings from spoken data rather support the default 2-1 order hypothesis, although it can be argued that there is more to entrenchment than just raw token frequency. Furthermore, it has been suggested that younger speakers do not show this 2-1 order preference in spoken Dutch (Olthof et al., 2017).

Next, we have some factors that do not really affect the hypotheses. We did find the priming factor as operationalized in our study and that of De Sutter (2007), but as noted by De Sutter, it applies equally to both word orders, so it has no bearing on the default word order hypothesis. The factor information value is not associated with a hypothesis in the table, because it was not clear to us which value is more difficult to process. However, De Sutter (2007) considered clusters preceded by a word of high information value to be easier to process. If this is true, our results would support the default 1-2 order hypothesis, as the more difficult condition of lower information values is more associated with the 1-2 order.

It is notable that medium-informative words (adverbs and adjectives) show the same effect as function words, while it seems unlikely that they are processed in similar ways. However, this is compatible with the views of Cinque (1998, 2010), who takes adverbs and adjectives to be specifiers of functional projections, thus making them quite parallel to functional elements. In this theory, the low and medium-informative words would all be specifiers, while the high-informative words are heads. But it could also be the case that this finding is simply not robust across different models or data sets. De Sutter (2009) also discussed this factor and in his study it was found to be the weakest effect. Furthermore, in De Sutter’s study the effect was in the other direction: words of high information value were 1.92 times more likely to occur in the 1-2 order, and medium information value words were 1.41 times more likely. The cause of this is either a difference between the corpora of the two studies (the Wikipedia domain versus the newspaper domain), or methodological differences. Furthermore, Bloem et al. (2014) found that words of high information value were 1.11 times more likely to occur in the 1-2 order, and medium information value words were 1.21 times more likely, again a different result, indicating that this finding may not be very reliable.

Several factors clearly support the default 1-2 order hypothesis. This is the case for morphological structure of the main verb (separable participles are more difficult to process and associated with the 1-2 order), length of the middle field (clusters with longer middle fields are more difficult to process and associated with the 1-2 order) and definiteness
Verbal cluster order and processing complexity

(definites are more difficult to process and associated with the 1-2 order).

The factor Multi-word units also matches the default 1-2 order hypothesis. Multi-word unit verbal clusters are easier to process and associated with the 2-1 order, except when they have an open slot for a PP. In that case, the association is the opposite, presumably because they will have to undergo some sort of syntactic processing. One part of the construction undergoes cluster formation and the other part optionally undergoes extraposition, even though both parts are stored as a unit. This may be a demanding task, though this has not been psycholinguistically tested, as far as we are aware. It is interesting to note that De Sutter (2007) finds an opposite effect for this factor: in his study, clusters in fixed expressions are 2.26 times more likely to be in the 1-2 order. Therefore, he interprets this as supporting a default 1-2 order. However, this study does not distinguish between fixed expressions with and without PPs, and the fixed expressions were not annotated in the same way, which may have led to different results.

For the factor extraposition, we can say that the overall result matches the default 1-2 order hypothesis, since extraposition eases processing, and extraposition is associated with the 2-1 order. Non-extraposition would be the more difficult condition, and that is associated with the 1-2 order.

Since we included adjunct and complement extraposition separately in the model, we can now also address the idea of De Sutter (2005), who suggested that complements might be more likely to be positioned closer to the main verb, due to their semantic relatedness. In the case of extraposition, that would predict 1-2 orders. However, extraposed complements are the condition that is the most likely to be in the 2-1 order in our results, so an adjacency effect seems unlikely. Therefore, it is more likely that both complement and adjunct extraposition depend on the processing complexity of the context. De Sutter (2007) suggested that complements might actually be more difficult to process when extraposed, rather than less difficult. Under that assumption, one could expect that the default order is more likely to be used in the case of complement extraposition. In our data, the ‘extraposed complement’ is associated more with the 2-1 order than the ‘extraposed adjunct’ condition. Under De Sutter’s (2007) assumption, this actually matches the default 2-1 order hypothesis, though the effect is small. Furthermore, we are not aware of any psycholinguistic evidence suggesting that complement extraposition is more difficult to process, nor is it clear that complement extraposition would be more difficult than no extraposition.

Another link that could be investigated is the one predicted by van Haeringen (1956), between extraposition and clause length. We did not include the length of the extraposition in our model, and there was no clear correlation between extraposition and length, but nevertheless this idea would be interesting to include in a multifactorial study. The idea was previously tested by De Sutter et al. (2008), though not in a multifactorial model. They did find a trend of longer extraposed elements correlating with more 1-2 orders. If long extraposed elements make the cluster more difficult to process, this supports the default 1-2 order hypothesis. It could be that the length of the extraposition and the
length of the middle field, when taken together, show a clearer processing effect, or that the results are different when extraposition is measured in a way that controls for the total clause length. It has already been shown by Willems & De Sutter (2015) that the length of the PP affects extraposition, perhaps this is a factor of processing complexity that relates to the verbal cluster the PP is extraposed over. This would be an interesting factor to include in future work.

One particularly interesting consideration here is that the PP-over-V extraposition in itself is a grammatical phenomenon showing variation that may be related to processing complexity, in the same way as the verbal cluster order variation. As we discussed, it has been linked to some factors related to processing in previous work. Willems & De Sutter (2015) explained some of the variation in PP placement in terms of complexity, and they state that their model requires more predictor variables because its performance leaves room for improvement. Perhaps some of this remaining variation can be explained using other factors of processing complexity that we just discussed. There could also be an interaction between the two phenomena, involving a balance of processing difficulty. For example, if a verbal cluster is produced in a ‘more difficult’ ordering, then the prepositional object may be produced in an ‘easier’ position, or the other way around. However, answering this question would require a larger corpus study of extraposition constructions and verbal cluster constructions at the same time, with a methodology similar to the current study.

Lastly, two of our factors appear to match the default 2-1 order hypothesis, rather than the 1-2 one. This was the case for the factor structural depth, a factor that was not studied by De Sutter (2007). It is unexpected that this factor and length of the middle field show opposite results — deeper syntactic structures are associated with the 2-1 order, but long middle fields are associated with the 1-2 order. One might expect that longer sentences also have more complicated structures with deeper structures, but interestingly, we found that structural depth and length of the middle field only correlate weakly. Therefore, the two factors appear to have distinct effects on verb cluster order. An explanation such as the one proposed by Baumann (2014) might account for this — in this experimental reading time study, a distinction is made between ‘structural distance’, which is like structural depth, and ‘linear distance’, which is like length. Baumann shows that ‘structural distance’ better predicts reading times, and has a clearer processing cost. In this view, the factor length of the middle field is not important for processing and the default 2-1 order hypothesis is supported. However, there is also a lot of evidence in support of resource-limitation models of sentence processing, in which ‘linear distance’ plays a larger role. Another possible explanation might be found in the theories that we mentioned in section 4.5, i.e. on non-hierarchical processing (Frank et al., 2012), which claim that syntactic hierarchy isn’t very relevant for processing effort. Since this factor also ranked second-to-last in terms of information gain in the model (Table 4.2), its importance as a processing effect may be limited.

The other factor that we found to be associated with the default 2-1 order
hypothesis is frequency, which showed more frequent words being associated with the 1-2 order. This factor was also studied by De Sutter (2005) and his results are consistent with ours. This appears to be counter-evidence to the default 1-2 order hypothesis. There are some possible issues with using this factor in a regression model. When including frequency in the model, there are likely to be more instances of the frequent verbs in the dataset of verb clusters, which may amplify the effect of the frequency factor in the regression model. Alternatively, it might be the case that some verbs are relatively more frequent in verbal cluster contexts, though this seems unlikely. As far as we are aware, any verb can be used in a verbal cluster construction. The consequence of this would be that these verbs are more easily activated in verbal cluster contexts only, resulting in a construction-specific frequency effect rather than a general one. Further testing for construction-dependent frequency effects may be helpful here to exclude this possibility.

Nevertheless, it appears that most of the processing-related factors that we investigated are compatible with the default 1-2 order hypothesis, as summarized in Table 4.3. These findings lead to an explanation based on Meyer & Weerman's (2016) acquisition study — the 1-2 word order is more entrenched in standard Dutch, as it is acquired as a construction earlier than the 2-1 order. Because the 1-2 order is more entrenched and more easily activated, it is used relatively more often in contexts that are more difficult to process.

4.8 Conclusion

In this work, we have aimed to provide a linguistic explanation for the observation that many different factors are associated with verbal cluster order variation. We have shown that for this particular case of grammatical optionality, Dutch verbal cluster order, it is necessary to go beyond semantic and syntactic factors in search of an explanation. Various factors that are associated with this word order variation were linked to processing complexity in this study. We claim that processing complexity is another motivation for variation, in addition to the more commonly discussed factors of semantics and information structure, which might apply to other alternations of near-synonymous constructions as well. Minimizing processing complexity is a plausible explanation for various facts that we observed in the corpus data.

All of the factors we discussed have a significant association with either of the word orders. The direction of the effect is fairly consistent. For most factors, it is the case that the more difficult condition is associated with the 1-2 order. We view this as evidence for a hypothesis we proposed based on Meyer & Weerman’s (2016) acquisition study — these factors are evidence that the more entrenched and more frequent 1-2 word order is used in contexts that are more difficult to process. However, not all of the factors we investigated support this idea, and the study was limited to written corpus data. Earlier work has shown a greater proportion of 2-1 orders in spoken Dutch, a fact that is difficult to
reconcile with this hypothesis.

It is notable that one of the factors we found to affect verbal cluster word order, the extraposition factor, is in itself also a grammatical phenomenon showing variation that could be investigated using the same methodology that was used in this study. More generally, it would be interesting to invoke processing complexity as a factor to explain other instances of word order variation among near-synonymous constructions. Using large amounts of annotated linguistic data and multifactorial models, it is possible to test this empirically.

Furthermore, it would be interesting to investigate additional factors related to processing complexity to obtain more evidence for or against the default 1-2 word order hypothesis. One possible factor, which was not included in the present study, is the number of arguments of the verb. It has been found that verbs that are more associated with the 1-2 order are more often ditransitive, while verbs associated with the 2-1 order are more often intransitive (Bloem, 2016b). This could be related to processing complexity as well. Using our methodology we are limited to factors that can be extracted from this automatically annotated corpus however, unless we analyze a part of the data manually. This would need to be done to look at most discourse or information-structural factors, such as ‘givenness’ or ‘focus’. The corpus also does not allow the inclusion of rhythm-related factors, which have been shown to affect verb cluster order in previous work, although these are unlikely to be related to processing complexity.

It might also be interesting to implement a model of language processing more directly, and test whether complexity in this model predicts verb cluster word order on corpus data. For example, a constraint satisfaction model of language processing would predict that a probabilistic language model can measure processing complexity. However, this is difficult because there are many possible ways in which such a model might be implemented, e.g. at the level of words, constructions, word classes, or a combination thereof. A further test would of course be a psycholinguistic experiment, in which the complexity of verb cluster contexts could be manipulated more directly, and processing difficulty could be indirectly measured.

Besides processing effects, there may be semantic or constructional factors that play a role in the choice between word orders that our model did not control for. Particular verbs or semantic classes of verbs can show a preference for one order or the other (Bloem, 2016b). It would be interesting to incorporate this into a multifactorial model in future work, although this is difficult due to the large number of possible main verbs.
CHAPTER 5

Testing the Processing Hypothesis of word order variation using a probabilistic language model*

Chapter Highlights

Problem Statement

- Previous work has indicated that processing complexity affects verb cluster word order choice, but processing complexity is a rather broad notion that is not easy to test empirically, as it can be operationalized in many ways.

Research Questions

- Can a specific theory of human language processing, surprisal theory, account for Dutch two-verb cluster word order variation?

- How much variation can we account for using a simple version of this single theory, compared to taking a wide range of factors that are related to processing complexity?

5.1 Introduction

According to functionalist theories of language, the way humans process language has shaped the grammars of our natural languages (Hawkins, 2014). While it is not always clear whether a particular grammatical rule or construction can be viewed as a consequence of general language processing mechanisms, there is certainly evidence suggesting that processing efficiency plays a role — speakers may choose to use different constructions in more complex contexts. This is particularly clear in contexts where grammatical variation is possible. Sometimes a speaker 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: [SUBJ gave DO to IO] or [SUBJ gave IO DO]. When a ditransitive verb is used, a speaker can almost always choose between those two constructions. For this particular alternation, and others like it, many studies have shown that a wide range of factors affect the choice (Gries, 2001; Bresnan et al., 2007; Colleman, 2009; Wasow et al., 2011), including factors related to language processing, and that the choice is not random.

These near-synonymous constructions are a particularly interesting case for the study of language processing, because other factors that may affect linguistic form, such as (most aspects of) meaning and grammaticality, are the same across both constructions. Nevertheless, usage differences can be observed between the two alternatives, even when produced by the same speaker. What remains to explain these differences are factors such as information structure, other pragmatic factors or (relative) processing complexity. To be able to take such factors into account, near-synonymous constructions are often studied using (large) text corpora and multifactorial statistical models. A range of variables that are considered to be empirical operationalizations of relevant factors (e.g. a factor such as definiteness, which can be related to information structure or processing complexity) are measured for each instance of the construction in the corpus, and modeled statistically. The model can then show how much each of those variables contributes to explaining the variation. This approach was first taken by Gries (2001) for English optional particle movement, studying
Testing the Processing Hypothesis of word order variation

the alternation between constructions where the particle ‘up’ is placed before 
or after the noun phrase:

(5.1) John picked up the book.
(5.2) John picked the book up.

The dative alternation was also studied using this method, by Bresnan et al. 
(2007). The variables that are found to be significant predictors in these multi-
factorial corpus studies are often related to language processing. Finding that 
construction 5.1 is preferred in contexts that are more difficult to process, Gries 
(2001) proposed the Processing Hypothesis for particle movement:

The multitude of variables (most of which are concerned with the 
direct object NP) that seems to be related to Particle Movement 
can all be related to the processing effort of the utterance. (Gries, 
2001)

However, the definition of processing effort or processing complexity used in 
these studies is generally quite broad. A wide variety of measures and features 
that can be linked to processing complexity are used, as well as theoretical 
notions applying to various domains of language. While the results of this 
approach are interesting, it is difficult to generalize over the factors discussed 
in such studies when so many different things constitute processing complexity. 
An alternative is to more precisely define processing complexity by taking a 
more specific theory of language processing that is internally consistent. While 
such theories may not cover all domains of linguistic complexity, they have been 
used to account for a range of phenomena and they help to make the notion of 
processing effort more directly quantifiable. This means that they can be used 
as a single measure, that can therefore be tested empirically on large corpora.

In this work, we test such a specific theory. We test a basic implementation 
of constraint satisfaction models of language processing by applying an n-gram 
language model to a case of grammatical variation between near-synonymous 
constructions. We use this n-gram model as a measure of surprisal, which, 
according to constraint satisfaction models of language processing, is a measure 
of processing complexity. This particular case of variation, Dutch verb clusters, 
has previously been studied using the type of multifactorial statistical model 
just described, and significant effects of processing complexity were found in 
these studies (De Sutter, 2007; Bloem et al., 2017b). By comparing our results 
to the results of these studies, our study can serve as a test of n-gram language 
models as a measure of processing complexity, and perhaps even of the surprisal 
theory it is based on.

We will start by introducing our case study of Dutch verb clusters in sec-
tion 5.2. Section 5.3 will address models of language processing and how lan-
guage processing has been argued to affect grammatical variation in previous 
work. Section 5.4 describes our data, in section 5.5 we describe our language 
model, and in section 5.6 we present our results. The results are discussed in 
section 5.7.
5.2 The case of Dutch two-verb clusters

Just like other Germanic languages, Dutch expresses properties such as tense and aspect by means of auxiliary verbs. As Dutch is (mostly) verb-final, these verbs end up clustered together at the end of the sentence. But unlike in other Germanic languages, these verb clusters allow a high degree of word order variation. Even in two-verb clusters, both logical word orders are possible in almost all cases:

(5.3) *Zij zei dat ze het gelezen had*

She said that she it read had

‘She said that she has read it.’

(5.4) *Zij zei dat ze het had gelezen*

She said that she it had read

‘She said that she has read it.’

The difference in word order is generally assumed not to correspond to a meaning difference, so we can consider these constructions as near-synonymous. As in other instances of near-synonymous constructions, a wide variety of factors has been shown to correlate with this alternation (De Sutter et al., 2007) and several generalizations over these factors have been proposed: sentence rhythm (De Schutter, 1996), information weight (De Sutter et al., 2007) and also minimizing processing complexity (De Sutter, 2005; Bloem et al., 2017b). Bloem et al. argue that the order in example 5.4, called the ascending order, is easier to process than the alternative order 5.3, called the descending order, because it correlates with features that are considered to be more difficult to process. This is similar to how Gries (2001) argued that the construction in example 5.1 is easier to process. Additional evidence comes from the claim that the ascending order is also acquired earlier by children (Meyer & Weerman, 2016).

In Bloem et al.’s (2017b) study, factors that are expected to correlate with the verb cluster word order variation are tested using a multifactorial model, and it is argued that those factors relate to processing complexity (besides the ones that mark different constructions). As an example, a factor relates to processing complexity when some psycholinguistic study has measured that a particular factor is more difficult to process. A set of such factors can be viewed as an a measure of processing complexity. However, another approach to measuring processing complexity is also possible and has been used in other corpus studies of grammatical variation phenomena: to implement a theoretical model of language processing, and test that on instances of the constructions of interest extracted from a corpus. The next section will elaborate upon these two methods of measuring processing complexity, and discuss studies that have used them.
5.3 Processing complexity

Processing complexity, from a human subjects perspective, refers to the amount of cognitive effort required to produce or comprehend an utterance. Speakers prefer to minimize their use of cognitive resources, formulating sentences in a way that minimizes processing complexity when multiple ways to express something are available. Listeners seem to process complex sentences more slowly and make more comprehension errors (Jaeger & Tily, 2011). This human subjects definition of complexity has also been called ‘relative complexity’, in contrast to ‘absolute complexity’ which is the formal complexity of the linguistic system (i.e. grammar) being used. Generally, only relative complexity is invoked in studies of grammatical variation.

There are at least two ways in which the notion of processing complexity can be invoked to account for grammatical variation in a corpus. Firstly, one can take a theoretical model of language processing, and apply it to instances the constructions under study from a corpus. The model might predict that one construction is more difficult to process than the other, or perhaps only in certain contexts. Secondly, one can use empirical measures of processing complexity, based on psycholinguistic experiments. If experiments have shown that people exhibit slower reading times or make more errors in sentences with feature A than with feature B, this can be taken to mean that feature A is more difficult to process. One can then test in a corpus whether the constructions under study occur more with the ‘easy’ feature or the ‘complex’ feature. This section will discuss these two approaches.

5.3.1 Theoretical models

Among theoretical models of language processing, two main approaches can be identified: constraint-satisfaction models, and resource-limitation (or memory-based) models (Levy, 2008).

Resource-limitation models focus on the idea that there is some limited cognitive resource, such as memory, that limits people’s capacity to process and produce language. Gibson’s (1998) Dependency Locality Theory is a prominent example of this approach. In this theory, among other constraints, longer-distance dependencies are dispreferred because they require more memory, and are therefore considered more difficult to process. Another such model, which is frequently referred to in linguistics, is formed by the efficiency principles of Hawkins (2004; 2014). The first principle in his theory is Minimize Domains, which states that dependency relations in the smallest possible domains are the most efficient. These principles are argued to play an important role in shaping what is grammatical, though they can be applied to the study of grammatical variation as well. Wiechmann & Lohmann (2013) applied this theoretical model in their multifactorial corpus study of prepositional phrase ordering, an alternation in English where the order of a verb’s two PP arguments (an adjunct and a complement) is free:
The astronomer gazed [into the sky] [through his telescope].

One factor they derive from the theory is that a shorter PP argument might prefer the first position, following the principle of Minimize Domains (the phrasal combination domain would be shorter with that ordering). Their model did not have a very high predictive accuracy over the corpus data. This is a common finding in these studies, as not every factor can be included in the model — factors such as prosody and information structure are difficult to test using a standard annotated corpus. Nevertheless, they found that the constraints theorized by Hawkins (2004) held for the corpus data they studied. However, they do not compare the effect of these constraints to other factors that often affect variation, such as empirical measures of processing complexity. Only the additional factor ‘information status’ is discussed.

Furthermore, these principles cannot easily be applied to every case of grammatical variation. The Wiechmann and Lohmann study discusses a case of interconstituent alternations, involving the ordering of constituents. However, in our case study of Dutch two-verb clusters, the alternation takes place within the verb phrase domain, and is therefore an intraconstituent alternation. As noted by De Sutter (2009, p. 226–227), principles like Minimize Domains do not necessarily apply here. So, we will look to the other main approach to modeling language processing.

The other approach, constraint satisfaction models, uses information from various domains of language (i.e. lexical, pragmatic) to consider various parallel alternative interpretations or parses of a sentence during processing. Furthermore, they relate processing difficulty to expectation, which is often grounded in probability theory (Jurašky, 2003) or relatedly, measures of surprisal (Hale, 2001; Levy, 2008). Therefore, this has also been called the Surprisal framework. In Hale’s surprisal theory, log-probability is considered a measure of the difficulty of a word. More surprising sequences of words or structures (that have lower probability) are considered to be more difficult to process and therefore more complex. These measures have been used to make various predictions about processing complexity that were verified using empirical data from psycholinguistic experiments. The concept of minimizing surprisal has also been called uniform information density (UID). This term is frequently used in linguistic studies, for example by Levy & Jaeger (2007). This UID measure measures the same thing as the perplexity measure, which is often used by computational linguists to evaluate language models. Levy & Jaeger (2007) studied it in the context of syntactic reduction, namely the possible omission of ‘that’ as a relativizer, which can also be considered a form of grammatical variation. In their study, an n-gram language model is a significant predictor of relativizer omission, as well as more syntactic features that are considered to have predictive power. This n-gram model was trained on a version of the Switchboard corpus in which all optional relativizers were omitted. However, no comparison with empirical measures of processing complexity is made. The UID measure has also been found...
Testing the Processing Hypothesis of word order variation

5.3.2 Empirical measures

We have just seen some examples of corpus studies of grammatical variation in which a particular theoretical model of processing complexity is used as the basis of the analysis, but usually, processing complexity is defined more broadly. An example of such a definition of processing complexity can be found in the first multifactorial corpus study of grammatical variation, where Gries (2001) states: “My idea of the notion of processing effort is a fairly broad one: it encompasses not only purely syntactic determinants, but also factors from other linguistic levels”. He lists phonologically indicated processing cost, morphosyntactically determined processing cost, semantically conditioned processing cost, and discourse-functionally determined processing cost.

In De Sutter’s (2007) variational corpus study of verb clusters, he interprets five factors that have previously been linked to verb cluster order variation in terms of cognitive cost. For example, the factor ‘frequency’ is interpreted as an indicator of cognitive cost, since psycholinguistic studies (i.e. reaction time studies) have shown that lower-frequency words are processed more slowly. In a subsequent corpus study, Bloem et al. (2017b) provide an overview of nine such factors that correlate with the word order variation in a large corpus. Just as other corpus studies of variation, this study is operationalized as a logistic regression model predicting which of the two orders is likely to be used, given the factors as predictors or independent variables. These factors are shown in Table 5.1. In this table, they are ranked by their information gain as measured in the stepwise regression procedure performed by Bloem et al. (2017b). In this procedure, one starts with an empty model, and adds the most informative factor each time, measuring the information gain. This measure is expressed as an Akaike Information Criterion (AIC) value, which measures information loss. A higher AIC means that more information is lost by the model, compared to the original data set. Therefore, the highest-ranked factors account for the largest amount of variation.

Factors 1 and 3 are control variables. Using a different auxiliary verb changes the meaning of a verb cluster construction and different auxiliary verbs have different word order preferences, so this factor obviously predicts word order in this kind of model, even though it is not a processing complexity factor. For the other factors, Bloem et al. discuss how they can be linked to results from psycholinguistic studies in which the factors, or similar ones, are measured, as well as to verb cluster order variation. Several of these factors are the ones that De Sutter (2007) also discussed. Frequency is also included here (6th in the table), as well as syntactic Priming (2nd), which is argued to ease processing on the basis of priming studies. The Length of the middle field of the
5.3. Processing complexity

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>AIC</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(none)</td>
<td>463279</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>Type of auxiliary</td>
<td>382538</td>
<td>80741</td>
</tr>
<tr>
<td>2</td>
<td>Priming</td>
<td>378185</td>
<td>4353</td>
</tr>
<tr>
<td>3</td>
<td>‘te’-infinitive</td>
<td>374378</td>
<td>3807</td>
</tr>
<tr>
<td>4</td>
<td>Extraposition</td>
<td>371413</td>
<td>2965</td>
</tr>
<tr>
<td>5</td>
<td>Length of middle field</td>
<td>369817</td>
<td>1596</td>
</tr>
<tr>
<td>6</td>
<td>Frequency of the main verb</td>
<td>368744</td>
<td>1073</td>
</tr>
<tr>
<td>7</td>
<td>Information value</td>
<td>367806</td>
<td>938</td>
</tr>
<tr>
<td>8</td>
<td>Morphological structure of the main verb</td>
<td>366870</td>
<td>936</td>
</tr>
<tr>
<td>9</td>
<td>Multi-word units</td>
<td>366162</td>
<td>708</td>
</tr>
<tr>
<td>10</td>
<td>Structural depth</td>
<td>365674</td>
<td>488</td>
</tr>
<tr>
<td>11</td>
<td>Definiteness</td>
<td>365461</td>
<td>213</td>
</tr>
</tbody>
</table>

Table 5.1: List of factors in the Bloem et al. (2017b) model of verb cluster order variation, ranked by information gain.

The sentence (5th) is also discussed, where a longer middle field is argued to be more difficult to process due to longer dependencies. The factor extra-position (4th) indicates whether a prepositional phrase was extraposed and positioned after the verb cluster, which has been argued to ease processing, and the factor information value (7th) measures the information value of the word before the verb cluster (i.e., whether it is a function word or content word). The factor morphological structure of the main verb (8th) refers to separable verbs, such as afdessen ‘wash up’ — such verbs appear to have a strong preference for the ascending order. The multiword unit factor (9th) indicates whether the verb cluster is (part of) a fixed expression, and structural depth (10th) refers to the depth of the verb cluster in the syntactic tree of the sentence. Lastly, as for the factor definiteness of the last word before the verb (11th), definiteness is argued to be more difficult to process on the basis of a study with language-impaired children, among other work. More detailed descriptions of the factors, their link to the notion of processing complexity and their effect on word order are provided by Bloem et al. (2017b).

All of the factors listed in Table 5.1 are statistically significant predictors of verb cluster word order, and they are all linked to processing complexity. In the present study, we will use this study as a basis of comparison for our probabilistic language model based on the constraint-satisfaction theory of language processing. Outside of the world of multifactorial corpus studies, processing complexity is also often defined in terms of empirical psycholinguistic measures, as evidenced by Bach et al.’s (1986) study on the processing complexity of larger
verb clusters, where processing complexity is measured in terms of error rate and comprehensibility judgements.

5.4 Data

For reasons of comparison, we use the same corpus that was used by Bloem et al. (2017b), which is the Wikipedia section of the Lassy Large corpus (van Noord, 2009). This corpus consists of a 145 million word dump of the Dutch-language Wikipedia in August 2011, and among these words, we can find 827,709 two-verb verbal clusters in total. The corpus has been automatically annotated with full syntactic dependency trees by the Alpino parser for Dutch (van Noord et al., 2006). While we do not need the annotation to train our language model, we do need it to automatically find and extract verb cluster constructions — extracting any sequence of two verbs is not sufficient. Furthermore, the annotation was used to extract the empirical measures of processing complexity used by Bloem et al. (2017b), used as factors in their model. The corpus was split into a training set (90%) and test set (10%), and from each set, the verb clusters and the factors were extracted. We also extracted plaintext, but tokenized, versions of the training and test sets for creating the language model.

5.5 Language model

To model the surprisal or predictability of a verb cluster, we trained a trigram language model on the plaintext corpus. We used Colibri Core (van Gompel & van den Bosch, 2016) to implement the language model efficiently. Colibri Core’s compression and counting algorithms enabled the modeling of this fairly large corpus without requiring excessive amounts of memory. The model was trained by having Colibri count n-grams and storing them as an unindexed pattern model. We used 3 as the maximum construction length \((n = 3)\) and no minimum length (to get trigrams, bigrams and unigrams), and no skipgrams. The construction threshold was set to 2, i.e. n-grams that only occur once are not included in the language model. Because we use an automatically annotated corpus, including such hapax legomena would be likely to result in the inclusion of many tokenization errors at the cost of higher memory usage.

A Colibri unindexed pattern model stores frequencies, but not probabilities. We perform maximum likelihood estimation (MLE) on the model over the training data to obtain probabilities during the test procedure. When testing, we iterate through all verb clusters extracted from the test set portion of the corpus, and estimate their probability and perplexity using frequency counts from the Colibri pattern model. For each of the two verbs in a cluster, we use linear interpolation to include trigram, bigram and unigram construction counts in the estimate. Furthermore, we use generalized additive smoothing, over the unigram constructions only, to account for out of vocabulary words in
100

5.6. Results

the test set. Therefore, our maximum likelihood estimation for a single verb is performed as follows:

\[
P(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) \\
+ \lambda_2 P(w_n|w_{n-1}) \\
+ \lambda_3 P(w_n) \quad \text{where} \quad P(w_n) = \frac{\delta + c(w_n)}{\delta|V| + c(N)}
\]

(5.1)

\(\lambda_1, \lambda_2 \text{ and } \lambda_3\) refer to the interpolation weights of trigram, bigram and (smoothed) unigram probabilities respectively. \(\delta\) is the smoothing parameter. \(V\) refers to the vocabulary size (or, the number of types in the language model), \(N\) to all tokens, and \(c(\cdot)\) refers to counts. We can now use perplexity per word (PPW) as a measure of surprisal for individual verbs or verb types. Perplexity is a measure of predictability or surprisal and it is generally used to compute how well a language model predicts a word in a sequence, or a sequence of words. We compute PPW as follows:

\[
PP(W) = 2^{-\log_2(p(V_1)p(V_2))}
\]

(5.2)

\(p(V_1)\) and \(p(V_2)\) are the estimated probabilities of the two verbs in the cluster (as estimated in Equation 5.1).

However, we would also like to have a measure of the predictability of the verb cluster as a whole. To this end, we also compute the perplexity per word over both words in the verb cluster:

\[
PP(C) = 2^{-\frac{1}{2}\log_2(p(V_1)p(V_2))}
\]

(5.3)

The log probabilities of \(P(V_1)\) and \(P(V_2)\) are multiplied by 1/2 because the verb cluster can be regarded as a sequence of length 2.

As noted, this measure is similar to Uniform Information Density as defined in previous work. We did not evaluate our language model in detail, because the goal of this study is not to have an optimal model for natural language processing. Rather, it is an attempt to see whether a basic form of a constraint-satisfaction model of language processing can account for verb cluster order variation, so we aim to make as few assumptions about the nature of the language model as possible.

5.6 Results

We ran the testing procedure to obtain perplexity values for each verb cluster with a range of parameter settings. We decided on a set of parameters to use based on two criteria. Firstly, regarding the construction length, we wanted the linear interpolation weights to be somewhat balanced between unigram, bigram and trigram probabilities in order to have a representative trigram model that does not rely too much on unigram or bigram probabilities. Longer construction lengths seem more cognitively plausible. Secondly, even though we
do not consider a well-performing language model to be essential for this study, we chose parameter settings that resulted in a low overall perplexity per word, computed over all verbs within the clusters. This resulted in a model with the interpolation weights set at $\lambda_1 = 0.3$, $\lambda_2 = 0.45$, $\lambda_3 = 0.25$, and a smoothing parameter of $\delta = 0.5$. We take perplexity to be an indicator of complexity following (Hale, 2001) who took log-probability as an indicator of complexity, as well as subsequent work in constraint-based models of language processing.

As a reminder, we repeat examples 5.3 and 5.4, showing the two possible word orders. Example 5.7 is in the descending order, where the main verb comes first and the auxiliary verb, in this case hebben ‘to have’, comes last. Example 5.8 is in the ascending order, which is the opposite:

(5.7) Zij zei dat ze het gelezen had
She said that she it read had
‘She said that she has read it.’

(5.8) Zij zei dat ze het had gelezen
She said that she it had read
‘She said that she has read it.’

<table>
<thead>
<tr>
<th>Word order</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending order clusters</td>
<td>1681.2</td>
</tr>
<tr>
<td>Descending order clusters</td>
<td>1675.8</td>
</tr>
<tr>
<td>Overall PPW</td>
<td>1679.8</td>
</tr>
</tbody>
</table>

Table 5.2: Perplexity per word (PPW) results for the two word orders, over all test-set clusters.

Table 5.2 shows that the perplexity per word of clusters in the ascending order, which is the more frequent order, is slightly higher than that of clusters in the descending order. At first sight, this seems to confirm the processing hypothesis of Bloem et al. (2017b) — the ascending order occurs in contexts of higher surprisal, and therefore lower predictability. This would confirm their idea that the ascending order is easier to process — to minimize surprisal and maintain uniform information density, one would use the less complex construction in the more complex context.

The difference seems small though, only 5.4 units of perplexity. We can test the predictive power of this measure for predicting word order by defining a logistic regression model over this data. In this model, word order is the dependent variable, a binary outcome variable that can be either ‘ascending’ or ‘descending’, and verb cluster perplexity (as defined in Equation 5.3) is the predictor variable. In this model, the perplexity factor is significant ($p < 0.05$), with a z-score of 2.2. As for the effect size, according to the model, for a one unit increase in perplexity, the log odds of the cluster being in the ascending order
5.6. Results

Information density is less uniform

Information density is more uniform

Figure 5.1: Visualisation of average perplexity per word results for both word orders.

Increases by 0.00000000035 ($3.5 \times 10^{-10}$). In other words, the effect size is tiny. For confirmation, we also tested the predictive power of the model by computing the concordance index (c-index). A c-index of 0.5 indicates chance level prediction, while 1 is perfect. This model’s c-index, based on 100 bootstrap repetitions, is 0.493, while multifactorial models of verb cluster order variation had c-indexes of 0.803 (De Sutter, 2005) and 0.765 (Bloem et al., 2014). Therefore, we cannot consider this result to be reliable.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear position</td>
<td>First verb</td>
<td>2264.9</td>
</tr>
<tr>
<td></td>
<td>Second verb</td>
<td>1245.8</td>
</tr>
<tr>
<td>Verb type</td>
<td>Auxiliary verb in cluster</td>
<td>714.0</td>
</tr>
<tr>
<td></td>
<td>Main verb in cluster</td>
<td>3952.0</td>
</tr>
<tr>
<td>Position and type</td>
<td>Auxiliary verb in descending</td>
<td>178.2</td>
</tr>
<tr>
<td></td>
<td>Main verb in descending</td>
<td>15763.5</td>
</tr>
<tr>
<td></td>
<td>Auxiliary verb in ascending</td>
<td>2445.8</td>
</tr>
<tr>
<td></td>
<td>Main verb in ascending</td>
<td>1155.6</td>
</tr>
<tr>
<td>Overall PPW</td>
<td></td>
<td>1679.8</td>
</tr>
</tbody>
</table>

Table 5.3: Perplexity per word (PPW) results for various conditions, over all test-set clusters.

To analyze this negative result, we can look at the perplexity per word values of individual verbs, for different verb types and verb positions. These different conditions are listed in table 5.3. We can distinguish two conditions here: the position of the verb in the linear order of the sentence (does it come first or last),
and whether the verb is an auxiliary verb or a main verb. These are essentially two features of the two word orders: in the ascending order, the first verb is the auxiliary verb while the second verb is the main verb, while the reverse is true for the descending order.

As for the linear position, we can observe that the first verb of a cluster is less predictable than the second verb. This seems logical, because in a two-verb cluster, the first verb is always followed by another verb. As for the verb type, we observe a bigger difference in perplexity between auxiliary verbs and main verbs — auxiliary verbs are much more predictable than main verbs. This can also be expected, because there is a limited number of auxiliary verbs (including any verbs that select another verb in a cluster), while main verbs can be anything and may include unknown words. This shows that linear position and verb types are both confounding factors in computing verb cluster surprisal. However, these observations do not control for the fact that the ascending order is more frequent, and therefore main verbs more often occur in the second position in linear order.

Therefore, it may be more informative to look at perplexity values for both linear position and verb type, as visualized in figure 5.1. This shows that perplexity is distributed quite differently in both orders. In the descending order, the main verb comes first. The perplexity for this is very high — the descending main verb is very surprising both because main verbs are more surprising, and because verbs in the first position are more surprising. Conversely, the auxiliary verb, which comes second, has very low surprisal. In the ascending orders, the two factors balance each other out — the auxiliary verb (low surprisal factor) comes first (which is a high surprisal factor). The main verb (high surprisal) comes last (low surprisal). In other words, instances of the ascending order have a more uniform information density.

Based on this result, we define a measure of verb cluster information density, which is the absolute difference between the log probabilities of both verbs in the cluster:

\[
UID(C) = |\log_2(P(V_1)) - \log_2(P(V_2))| \quad (5.4)
\]

Again, \(P(V_1)\) and \(P(V_2)\) are the estimated probabilities of the two verbs in the cluster (as estimated in Equation 5.1). Putting this factor in a logistic regression model gives us a c-index of 0.686 according to the procedure described for the previous model (except that the measure from Equation 5.4 is used, instead of that in Equation 5.3). This is of course a lot better than 0.493, especially for a model with a single predictor. The effect of the factor is also highly significant. We can now test whether this UID-effect holds when we also include the nine empirical measures of processing complexity (and the control variables) from the study of Bloem et al. (2017b). This would tell us if our cluster-UID-measure measures the same thing as the empirical measures from that study.

This can be done by adding the UID measure from Equation 5.4 to the
multifactorial regression model of Bloem et al. as a predictor variable. To do so, we created a regression model that includes all of the factors listed in Table 5.1 as predictors, as well as our UID measure, and that has word order as the dependent variable. We found that adding UID to the Bloem et al. model significantly improves it. A global comparison of the original model and the model with the UID factor using the $\chi^2$-test shows that the residual deviance drops from 54880 to 48795, and this is statistically significant ($p < 0.001$). We also observe that the UID-measure is highly significant in this model, with an odds ratio is 0.788, indicating a decrease in the odds of observing an ascending order when the UID-measure is higher (which is when the density is less uniform). Furthermore, if we perform stepwise regression with this model to measure information gain, the UID factor is ranked second after the control factor TYPE OF AUXILIARY. In Table 5.1, which lists information gain for the Bloem et al. model, it would be listed second. It is therefore the most informative factor related to processing complexity in the new model. However, the predictive value of the model does not improve — the original model has a predictive value of $c = 0.7897$, and adding our UID-measure gives us $c = 0.7896$, a negligible difference.

Surprisingly, there is no multicollinearity in this model. The variance inflation factor (VIF) for each factor is very low ($< 1.2, 1.203$ for the UID factor). This indicates that the UID-measure does not correlate with the factors from the Bloem et al. model, but is complementary to them and accounts for a different portion of the statistical variance.

5.7 Discussion

Our results show that perplexity-per-word as a measure of surprisal does not predict word order variation in two-verb clusters, even though it has been argued that processing complexity predicts word order. The perplexity values computed on the basis of the probabilities from the n-gram language model that we used cannot be considered a measure of processing complexity. However, a derived measure of uniform information density (UID), which measures a difference in surprisal within the verb cluster construction on the basis of the same language model, does predict the word order variation. We furthermore showed that this UID measure improves upon a previous model that was based on empirical measures of processing complexity. Therefore, our UID measure can be viewed as complementary to these measures when accounting for word order variation in two-verb clusters.

This result indicates that part of the variation that the empirical measures account for, is also accounted for by the UID measure, but not all of it. Furthermore, not all of the variation that the UID measure accounts for, is accounted for by the empirical measures. More broadly, human subject measures and a computational measure of complexity were combined, and this combination lead to an improvement in explanatory power for this grammatical variation. We chose to use a measure in the Surprisal framework, or constraint satisfaction modeling
Testing the Processing Hypothesis of word order variation

approach, because Dependency Locality Theory is not so clearly applicable to verb cluster order variation, which is an intraconstituent alternation.

Our analysis also showed that verb clusters in the ascending order generally have a more uniform information density than verb clusters in the descending order. This is because both linear position and the type of verb affect the predictability of a verb, and in the ascending word order, these two factors balance each other out. Under the assumption made by Levy & Jaeger (2007) that uniform information density indeed facilitates processing, our findings seem to support the processing hypothesis for Dutch verb cluster order. However, the direction of the effect is not clear - it can either be argued that the ascending order is easier to process because it has a more uniform information density, or it can be argued that the ascending order is more difficult to process, because speakers use it in more predictable contexts (that are less difficult to process). In future work, this ambiguity could be clarified by studying the information density of not only the verbs themselves, but also the words before and after the verbs.

While we believe that Dutch verb cluster word order is a typical case of near-synonymous word order variation, this raises the question of whether these findings would hold for other cases of grammatical variation. Our result does not necessarily mean that surprisal measures are not representative of processing complexity in general. Surprisal has been shown to be informative in other studies of other phenomena, for example to discover contexts in which a relativizer is preferred (Levy & Jaeger, 2007). Furthermore, surprisal can be and has been measured in a variety of ways. We implemented it in a very basic way. Hale (2001) measured surprisal using a probabilistic parser rather than an n-gram language model. Perhaps a measure of surprisal that takes more structure or syntax into account would be more predictive of verb cluster order variation or other alternations. The measure can and has been implemented on the basis of other structural elements rather than words, such as constructions, part-of-speech sequences, topic models, or any other level of structure that one could train a language model over. In future work, it would be interesting to try computing surprisal in the same way as Hale (2001), to compute it at a different level of structure, or to use more elaborate language models containing larger chunks or skipgrams. A delexicalized n-gram model could be used to make the measures we used more sensitive to structure. For our particular case it would also be interesting to define our prediction task in a different way — to learn more about the word order variation, it would be interesting to adapt the language model such that it only predicts the order of the cluster, rather than the specific words in it. This might tell us more about how predictable the cluster orders are, regardless of the specific lexical items involved.

Nevertheless, our findings do provide evidence that uniform information density may be a better operationalization of constraint satisfaction models of language processing than plain surprisal, when one is studying an alternation involving multiple words. Uniform information density should be considered as a measure of processing complexity, particularly in multifactorial corpus studies.
of grammatical variation.
Lexical effects on Dutch verbal cluster order*

Chapter Highlights

Problem Statement

- It is claimed that there is no meaning difference between the Dutch verb cluster orders, but various theories of language assume isomorphism: one form corresponds to one meaning. These theories predict a meaning difference between the two orders.

Research Questions

- Can associations between a verb and a verb cluster order be found, even when controlling for other factors?
- Can those associations be explained in semantic terms?
- How can we control for the effect of other factors, when we are studying lexical associations using collostructional analysis?

6.1. 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 test for lexical associations with a particular word order using collostructional analysis, and discuss whether the two word orders are semantically different.

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:

(6.1) \textit{Ik denk dat ik het begrepen heb}\ \\
I think that I it understood have \ \ \\
‘I think that I have understood it.’

(6.2) \textit{Ik denk dat ik heb begrepen}\ \\
I think that I it have understood
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 6.1 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 6.2 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. von Humboldt 1836, Bader et al. 2009), although it has been argued that there is a difference in interpretation (Pardoen, 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 6.1 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 6.1b) or 1-2 (Figure 6.1c) ordered verbal cluster.

These analyses account for the grammaticality of both orders, but do not specify which one to use when. 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 between these two options 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 particular verb cluster orders 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.

The present study is relevant to two factors that may cause speakers to use one word order over the other, but have not yet been included in multifactorial models of the phenomenon: the effect of associations with related constructions, and the effect of semantic differences between the two orders.

In usage-based approaches to the study of language, frequency has emerged as an important factor influencing language processing. Frequent usage helps to entrench linguistic elements in memory, and reinforces the associative connections between them. Linguistic elements that often occur together in language use, develop a stronger association in memory (Diessel & Hilpert, 2016). The mechanism underlying this is known as entrenchment (Langacker, 1987). For verb clusters, this may mean that verbs that often occur in, let’s say, the 1-2 order, become associated with it in memory, although this relies on the assumption that each verb cluster order is a distinct linguistic element. This implies that speakers could choose to use the 1-2 order with a particular verb simply because they have often seen this combination of verb and word order before, which may account for part of the variation that has been observed.

When these associations involve words within constructions, they are often assumed to have a semantic motivation: the words should be semantically compatible with the construction (Goldberg, 1995). However, as mentioned, for verb clusters the existence of such semantic associations is not at all clear. Furthermore, studies have pointed out that the associations between words and constructions are not fully predictable from general semantic criteria (e.g. by Boas 2003), yet speakers still know them (Diessel & Hilpert, 2016). Therefore, it seems prudent to investigate both the possibility of semantic differences between the verb cluster orders, and lexical associations of specific verbs with verb cluster orders, as factors that may affect the choice between the 1-2 order and the 2-1 order.

6.1.1 Previous work

These associations between main verbs and verb cluster orders have previously been studied by De Sutter (2005, 242-248), though they receive only a fairly
brief mention in his dissertation. De Sutter is mainly interested in this topic to test the hypothesis of Pardoen (1991), which states that the 1-2 order leads to a dynamic interpretation of verbs. In a dataset of 2390 verb clusters, De Sutter finds 53 main verbs (5.8% of total) that are significantly associated with the 2-1 order, and 22 main verbs (2.4% of total) that are significantly associated with the 1-2 order. After examining the verbs, De Sutter tentatively agrees with Pardoen’s hypothesis, but also finds some exclusively stative verbs that are associated with the 1-2 order. Furthermore, De Sutter claims that the 2-1 order is more lexically specific, as more main verbs are significantly associated with it.

What does this study tell us about semantic or lexical associations? De Sutter (2005) is unable to draw any clear conclusions regarding semantic patterns associated with either order. Furthermore, relatively few words with a statistically significant association for either word order are found. There are always at least two possible reasons for a lack of statistically significant result: either these lexical associations do not exist, or the sample size was too small and there was simply not enough evidence available to prove an effect for each verb. I will argue that this second reason is the most likely explanation. De Sutter’s dataset consisted of 2390 verb clusters, so there were not likely to be many main verbs occurring in more than five clusters. Taking a similarly sized sample of verb clusters from another corpus demonstrates this.

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. To find statistically significant 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 verbs, despite the overall size of the corpus. This is probably the reason why lexical associations with an open class of words such as verbs or nouns, with no selectional restrictions, are not often investigated, and why De Sutter (2005) did not get many results. There are too many possible words that could be used, and most of them are relatively infrequent: specifically, as the main verb of a verb cluster, any Dutch verb can be used.

However, De Sutter extracted these verb clusters from a manually annotated 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 the present study, I use the Wikipedia section of the Lassy Large corpus,
one of the largest syntactically annotated corpora 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.

Using this large dataset of verb clusters, it becomes possible to look for
lexical and semantic associations of verb cluster orders by focusing on the
usage patterns of main verbs in verb clusters in a corpus of Dutch. Apart from
De Sutter (2005), 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).

The rest of the paper is structured as follows. First, I discuss the concepts of
lexical and semantic associations in section 6.2. I discuss how lexical associations
can be measured in section 6.3. Section 6.4 discusses related work on Dutch
verb clusters, and section 6.5 describes the corpus data I used for this study, as
well as the way in which the analysis was performed. Section 6.6 presents the
results of the study, followed by a discussion in section 6.7 and a conclusion
(section 6.8).

6.2 Lexical and semantic associations

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 processing complexity, prosody, or other factors mentioned in the
introduction to this chapter. It might be the case that the construction prefers
a larger class of verbs, and a particular verb associated with the construction
is just one of many. 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
(Bloem et al., 2014). Because these effects are already known, I have not studied them in detail, but I have tried to control for them in this study, both by excluding them and by adapting the collostructional analysis method. Besides these ‘other’ factors, as mentioned in the introduction, there are two factors that may cause us to observe lexical associations in a corpus. Firstly, observed lexical associations can indicate an effect of entrenched associations with related constructions in the grammar of the writer/speaker, and secondly, they can indicate an effect of semantic differences between the two orders. In this section, these two possibilities will be discussed in more detail.

6.2.1 Associations with related constructions: Lexical preferences

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 develop an entrenched association with this word order in the mental grammar of the speaker. It will then be produced more often in that order as well, 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). When this association cannot be attributed to other, more general factors, it must be a lexical preference. A lexical association can be observed in a corpus, but to establish that it is a lexical preference of the speaker or writer, other factors must be excluded. Conversely, if lexical preferences do not exist, one would expect that no statistically significant evidence for any 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 lexical 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.

\[1\] The 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.
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 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 Pardoen (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. However, De Sutter (2005) did not find clear evidence for this hypothesis. Nevertheless, for this reason and for the reason that lexical associations are often assumed to have a semantic motivation, this study includes tests for semantic preferences of the verb cluster word orders, including stative and dynamic.

### 6.2.2 Semantic differences between the orders

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, as the two orders differ in form. 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 do not have a clear hypothesis on what sort of semantic preferences might be found among the verbs associated with a particular order. The only two hints of semantic differences that are available are the dynamic interpretation hypothesis of Pardoen (1991), and the possibility of adjectival interpretation of participles in the 2-1 order, discussed by De Sutter (2005), among others, which I will discuss later. To keep all options open, in this study I will follow a broad semantic categorization of the associated verbs in an attempt to find semantic associations.

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, described in the next section.

6.3 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; Colleman 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 $\chi^2$, 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). This is the method that is typically used to measure lexical associations.

There is a downside to the use of collocation-based methods, which is that they do not control for other factors that may affect the lexical association. Only frequency is controlled for. For verb clusters, there are many other relevant factors that have been identified in the literature. For example, particle verbs are strongly associated with the 1-2 order construction. This is unlikely to be the case due to semantic reasons, but rather for reasons of processing (Bloem et al., 2017b) or prosody (De Sutter et al., 2007). Therefore, I also explore two other measures that do involve multiple control factors.

6.3.1 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 O1 Od]. In this case, the verb give is filled in, while the other elements are open slots that can be filled, e.g. Od 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 6.1 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).?

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

---

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.
6.3.2 Gain Ratio

As mentioned, the association measures that are calculated in a collostructional analysis do not control for factors that may have caused an association between a word and a construction to be observed. In an attempt to quantify the effect of lexical associations on the variation in comparison to other factors, such as the particle verb factor mentioned above, we applied a measure of informativity that can quantify the informativity of each factor in an information-theoretical sense, including a ‘main verb’ factor, on the variation. More informative factors can account for more of the observed variation between word orders in a multifactorial statistical model. A ranking of the informativity of factors affecting verb cluster order variation has also been calculated by Bloem et al. (2017b) using the information gain measure, but this ranking did not take the lexical factor into account.

The measure of informativity that I used for this is the Gain Ratio measure (Quinlan, 1993). This measure is computed over a training set of the verb clusters we extracted from the Lassy corpus, annotated for their order and for various factors that are known to affect verb cluster order (taken from Bloem et al. (2014), the same factors are used in a multifactorial regression model in this study, but without a main verb factor). First the Information Gain (IG) of the ‘main verbs’ factor is computed, the difference in entropy between a model of the training data without and with knowledge of the main verbs. In this way, the other annotated factors are controlled for. The IG is then normalized for the number of possible values per feature to obtain the Gain Ratio, because there are a lot of possible values for the main verb, it could be any verb. We used TiMBL to compute this value (Daelemans et al., 2007). This gives us a measure of the effect size of the lexical association factor compared to the control factors, but does not provide any information on specific verbs.

This may sound like a more cumbersome way of calculating the effect of a random factor in a multifactorial regression model, and it is. The normal way to measure the effect of a variable with many possible values, such as the main verb of a cluster (where any Dutch verb can be used) would be to include it as a random effect in a mixed-effects model. However, due to the large size of the data set and the large number of verbs, such models fail to converge. Therefore, TiMBL was used instead, which uses an exemplar-based method to perform what is called memory-based learning. Exemplar-based methods are better suited for modeling a factor with so many possible values (i.e. all Dutch verbs), as they do not attempt to generalize. Generalization to a linear term, as in a linear regression mode is an impossible task when there are many unique verbs — you can’t create a line from a single data point.

However, this measure cannot quantify the effect of lexical associations per verb, only the effect of lexical associations of main verbs in clusters in the dataset in general. To obtain a measure per verb, I tried to use the residuals of
6.4. 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 6.3 (repeated from the introduction) is often called the ‘groene volgorde’ (green order), and the 1-2 order of example 6.4 is called the ‘rode volgorde’ (red order):
Lexical effects on Dutch verbal cluster order

(6.3) *Ik denk dat ik het begrepen heb*  
I think that I it understood have  
‘I think that I have understood it.’

(6.4) *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 6.3 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 6.3 and 6.4 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, 2007b). 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 collocational 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.

6.5 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 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. For this analysis, it is necessary 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.

6.5.1 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
6.5. Method and data

<table>
<thead>
<tr>
<th></th>
<th>horen</th>
<th>¬ horen</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>406</td>
<td>367,181</td>
<td>367,587</td>
</tr>
<tr>
<td>¬ 1-2</td>
<td>974</td>
<td>198,876</td>
<td>199,850</td>
</tr>
<tr>
<td>totals</td>
<td>1,380</td>
<td>566,057</td>
<td>567,437</td>
</tr>
</tbody>
</table>

Table 6.1: An example contingency table for the verb *horen* ‘to hear’, used in collostructional analysis

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 6.4 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 6.1. The value $c$ is the frequency of co-occurrence of a potential collexeme, such as the verb *horen* with the 1-2 order construction: in other words, the frequency of *[horen AUX]*. 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 *horen* in all verb clusters. The value $cn$ is the frequency of occurrence of the 1-2 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. Once we have these four counts, the other values in the contingency table can be calculated automatically based on
them, and the table can be used to compute a measure of association. In this case, we compute an odds ratio of 4.4 for *horen* and the 2-1-order (indicating that it is 4.4 times more likely to occur in the 2-1 order), and compute a p-value of $9.69 \times 10^{-164}$ using Fisher’s Exact Test. The frequencies in the table already hint at this strong effect (406 uses in the 1-2 order and 974 uses in the 2-1 order) but association measures also control for the frequency of each condition in the corpus. For example, if the 2-1 order was twice as common as the 1-2 order, the above numbers of 406 and 974 would probably not constitute a significant association, as one would expect the verb to be found in more 2-1 orders when there are more 2-1 orders.

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

6.5.2 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 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 tools, one of which allows users to search the treebank using XPath, 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 API, 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 6.4. 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 6.5.1.

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). In addition, I compute odds ratios as a measure of effect size. The associations were calculated using the fisher_exact function from the SciPy library for Python, which provides odds ratios as well as p-values from the test.

In his verb cluster study, De Sutter (2005) used the log-likelihood ratio test as a measure of association to compute both a p-value and an effect size (the ratio). 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 α = 0.05). It

---

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/
does not allow ranking the verbs, while ranking is essential for a collostructional analysis. Furthermore, there is no 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.

6.6 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.

6.6.1 Initial distinctive collexeme analysis

Following the method of distinctive collexeme analysis outlined in section 6.5.1, 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 6.2 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
### 6.6. Results

<table>
<thead>
<tr>
<th>1-2 order</th>
<th>2-1 order</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collexeme</strong></td>
<td><strong>Odds R.</strong></td>
</tr>
<tr>
<td>toebehoren (76:0)</td>
<td>inf ***</td>
</tr>
<tr>
<td>‘to belong to’</td>
<td></td>
</tr>
<tr>
<td>malen (56:0)</td>
<td>inf ***</td>
</tr>
<tr>
<td>‘to grind’</td>
<td></td>
</tr>
<tr>
<td>toeleggen (212:6)</td>
<td>19.38 ***</td>
</tr>
<tr>
<td>‘to focus’</td>
<td>‘to trust’</td>
</tr>
<tr>
<td>opzeggen (70:2)</td>
<td>19.21 ***</td>
</tr>
<tr>
<td>‘to cancel’</td>
<td>‘to necessitate’</td>
</tr>
<tr>
<td>plaatsvinden (1004:30)</td>
<td>18.32 ***</td>
</tr>
<tr>
<td>‘to take place’</td>
<td>‘to require’</td>
</tr>
<tr>
<td>kwijtraken (171:8)</td>
<td>11.73 ***</td>
</tr>
<tr>
<td>‘to get rid of’</td>
<td>‘to befriend’</td>
</tr>
<tr>
<td>binnenkomen (63:3)</td>
<td>11.52 ***</td>
</tr>
<tr>
<td>‘to come in’</td>
<td>‘to oblige’</td>
</tr>
<tr>
<td>opsommen (55:3)</td>
<td>10.06 ***</td>
</tr>
<tr>
<td>‘to sum up’</td>
<td>‘to make use’</td>
</tr>
</tbody>
</table>

Table 6.2: Top eight collexemes of the 1-2 and 2-1 verb cluster orders

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. This is in stark contrast to De Sutter’s (2005) results, who finds 53 verbs (5.8% of total) that are significantly associated with the 2-1 order, and 22 verbs (2.4% of total) that are significantly associated with the 1-2 order, at a threshold of $\alpha = 0.05$. However, De Sutter does not use any frequency cutoff, which may affect the percentage positively (if many low-frequent verbs only occur in one order) or negatively (if many low-frequent verbs occur in both orders enough to have no evidence for a significant association), so it is difficult to compare these numbers exactly.

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 $\chi^2$-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 more strict condition, 531 verbs (43.2%) have significant associations.

To assess the robustness of the effect, it is also interesting to see whether the more frequent verbs have word order associations in verbal clusters. In order to do so, we ran another analysis with only the verbs that occurred more than 1000 times as verbal cluster main verbs in the data set. This left us with the 110 most frequent verbs. Out of this set, 90 (82%) are significantly associated with one of the two word orders at $\alpha = 0.05$. This demonstrates that our results are affected by some kind of frequency effect. Since p-values depend on sample size, this is not too surprising. To avoid this effect, we also look at effect sizes for this overview.

We take an odds ratio of 1.5 (or 0.66) as a threshold of effect size: we consider a word to be associated with an order, if its estimated chance of being used in that order is $> 1.5$ times more likely than being used in the opposite order. 60 (54.5%) of the highly frequent verbs have an effect size greater than this threshold. For the larger set of verbs, 777 (63.1%) of the verbs have an effect size greater than this threshold. This shows that these word order associations are not a marginal phenomenon that only occurs with uncommon verbs, though they are slightly more common among lower-frequency verbs. It also indicates that the associations we find are not just due to frequency effects of null hypothesis significance testing.

### 6.6.2 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 6.4. 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)

As it is not clear to me how the third measure would contribute, 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 6.2 (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 6.2) 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, as found by Bloem (2016a). 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 6.6.4 I will discuss an analysis where adjectivity is controlled for.

### 6.6.3 Particle verb factor

One notable result of the first analysis 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 certain contexts. In Dutch orthography, 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 6.2, 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 6.2 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 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%. As for effect size, 346 verbs (45.1%) reach our previously established effect size threshold of 1.5 (previously 63.1%). The reduction in terms of effect size shows that the lower proportion of significantly associated verbs in this condition is not just a frequency effect due to the smaller size of this data set.

These results indicate 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.

6.6.4 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 auxiliary verb. This also controls for the adjectivity factor I discussed in section 6.6.2, 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 6.3 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%. As for effect size, 109 verbs (50%) reach our previously established effect size threshold of 1.5 (previously 45.1% with all non-particle verbs). 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.

6.6.5 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
Table 6.3: Top twenty collexemes of the 1-2 and 2-1 verb cluster orders with hebben ‘to have’, excluding particle verbs
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 Pardoen (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 6.3). The results are summarized in Table 6.4. 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.\(^8\)

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

---

\(^8\)I noticed a few possible errors, for example, the first sense of slapen ‘to sleep’ is annotated as having negative polarity
Lexical effects on Dutch verbal cluster order

<table>
<thead>
<tr>
<th>Property</th>
<th>Values</th>
<th>1-2 order</th>
<th>2-1 order</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valency</strong></td>
<td>Mono</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Di</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Tri</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td><strong>Transitivity</strong></td>
<td>Intransitive</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Transitive</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Ditransitive</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>+control</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>-control</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td><strong>Dynamic</strong></td>
<td>+dynamic</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>-dynamic</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td><strong>Attributive</strong></td>
<td>+attributive</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-attributive</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td><strong>Spatial</strong></td>
<td>+spatial</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-spatial</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td><strong>Cognition</strong></td>
<td>+cognition</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>-cognition</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td><strong>Polarity</strong></td>
<td>Negative</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.4: Frequency count of semantic properties of the top 20 words in each word order

that there are no meaning associations of those features with a particular word order. Based on Pardoen’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 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 6.3 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 6.4 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 6.3, 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 *dranken* ‘to drink’. It appears that at least some semantically similar verbs have very similar cluster order associations.

### 6.6.6 Gain Ratio

Besides controlling for other factors by excluding them from the analysis, I have also explored alternative ways of measuring lexical effects. I used the Gain Ratio measure of informativity to calculate the importance of the lexical effect compared to other factors that affect verb cluster order. The effect of the main verb was compared to the other factors affecting verb cluster order choice discussed by Bloem et al. (2017b). The measure indicates that the factor ‘main verb’ is among the most informative factors for modeling cluster order, as shown in Table 6.5. The results differ from those calculated by Bloem et al. (2017b) because they include more types of verb clusters, such as those with modal auxiliary verbs, which greatly increases the informativity of some factors, such as TYPE OF AUXILIARY.

Other measures of feature informativity produce somewhat varying results, but assign a high rank to this factor as well. The Gain Ratio measure has been criticized for having a bias towards features with more values, and the main verb factor has many more values than the other factors (White & Liu, 1994). TiMBL (Daelemans et al., 2007) implements a $\chi^2$-based measure as an alternative, where different numbers of degrees of freedom can control for the number of values. With this measure, the main verb lexical effect factor is actually ranked first, and the main verb frequency factor is ranked second. However, this measure does not work well for low-frequency features, which is likely to be the case for these main verb factors, as many verbs occur only a few times and there are many possible main verb frequencies. This $\chi^2$-based result is therefore unreliable.
Table 6.5: Gain Ratios of factors affecting verb cluster order variation, as calculated by TiMBL (Daelemans et al., 2007)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>Gain Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘te’-infinitive</td>
<td>0.0862</td>
</tr>
<tr>
<td>2</td>
<td>Morphological structure of the main verb</td>
<td>0.0349</td>
</tr>
<tr>
<td>3</td>
<td>Multi-word units</td>
<td>0.0139</td>
</tr>
<tr>
<td>4</td>
<td><strong>Main verb lexical effect</strong></td>
<td><strong>0.0123</strong></td>
</tr>
<tr>
<td>5</td>
<td>Type of auxiliary</td>
<td>0.0104</td>
</tr>
<tr>
<td>6</td>
<td>Frequency of the main verb</td>
<td>0.0100</td>
</tr>
<tr>
<td>7</td>
<td>Priming</td>
<td>0.0091</td>
</tr>
<tr>
<td>8</td>
<td>Definiteness</td>
<td>0.0063</td>
</tr>
<tr>
<td>9</td>
<td>Length of middle field</td>
<td>0.0060</td>
</tr>
<tr>
<td>10</td>
<td>Extraposition</td>
<td>0.0046</td>
</tr>
<tr>
<td>11</td>
<td>Information value</td>
<td>0.0041</td>
</tr>
<tr>
<td>12</td>
<td>Structural depth</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

6.6.7 Residuals of logistic regression

Similar to the association values produced by collostructional analysis, our other controlled measure, the averaged logistic regression residuals per verb, can also be used to rank the verbs. Higher ranking verbs are more associated with a particular order than predicted by the logistic regression model. A positive residual indicates that the association with the 1-2 order is stronger than predicted, and a negative residual indicates that the association with the 2-1 order is stronger than predicted. This result in a ranking of verbs that is surprisingly similar to that of the collostructional analysis, as shown in Table 6.6. This table shows the same ranking as Table 6.2 on the left, and on the right it shows the verbs ranked by the new measure, average residuals of the logistic regression model from Bloem et al. (2014).

We can see that, of the 8 verbs most associated with the 1-2 order in the collostructional analysis, 5 are also among the top 8 1-2 order verbs ranked by residuals. If we look at the top 30 instead of the top 8, 50% of the verbs occur in both lists. However, we can see that the morphologically complex verbs, with particles, still complicate the issue. Even though the residuals measure controls for complex verbs statistically, as this factor is included in the model, many still show up as strongly associated with the 1-2 order. It is not clear whether this is due to actual lexical preferences involving these verbs or due to an improper fit of the regression model for this particular factor. It might be the case that this factor is not weighed strongly enough by the regression model, as particle verbs are relatively rare. If this is the case, this would also mean that the actual
### 6.6. Results

#### 1-2 order: Collostructional analysis

<table>
<thead>
<tr>
<th>Collexeme</th>
<th>Odds R.</th>
<th>1-2 order: Regression residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>toebehoren (76:0)</strong></td>
<td>inf ***</td>
<td><strong>malen (56:0)</strong> 0.911</td>
</tr>
<tr>
<td>‘to belong to’</td>
<td></td>
<td>‘to grind’</td>
</tr>
<tr>
<td><strong>malen (56:0)</strong></td>
<td>inf ***</td>
<td><strong>opsommen (55:3)</strong> 0.889</td>
</tr>
<tr>
<td>‘to grind’</td>
<td></td>
<td>‘to sum up’</td>
</tr>
<tr>
<td><strong>toeleggen (212:6)</strong></td>
<td>19.38 ***</td>
<td><strong>opgaan (330:30)</strong> 0.887</td>
</tr>
<tr>
<td>‘to focus’</td>
<td></td>
<td>‘to go up’</td>
</tr>
<tr>
<td><strong>opzeggen (70:2)</strong></td>
<td>19.21 ***</td>
<td><strong>overlopen (57:5)</strong> 0.876</td>
</tr>
<tr>
<td>‘to cancel’</td>
<td></td>
<td>‘to run over’</td>
</tr>
<tr>
<td><strong>plaatsvinden (1004:30)</strong></td>
<td>18.32 ***</td>
<td><strong>kwijtraken (171:8)</strong> 0.868</td>
</tr>
<tr>
<td>‘to take place’</td>
<td></td>
<td>‘to get rid of’</td>
</tr>
<tr>
<td><strong>kwijtraken (171:8)</strong></td>
<td>11.73 ***</td>
<td><strong>binnenkomen (63:3)</strong> 0.860</td>
</tr>
<tr>
<td>‘to get rid of’</td>
<td></td>
<td>‘to come in’</td>
</tr>
<tr>
<td><strong>binnenkomen (63:3)</strong></td>
<td>11.52 ***</td>
<td><strong>binnenvallen (118:8)</strong> 0.851</td>
</tr>
<tr>
<td>‘to come in’</td>
<td></td>
<td>‘to invade’</td>
</tr>
<tr>
<td><strong>opsommen (55:3)</strong></td>
<td>10.06 ***</td>
<td><strong>toeleggen (212:6)</strong> 0.831</td>
</tr>
<tr>
<td>‘to sum up’</td>
<td></td>
<td>‘to focus’</td>
</tr>
</tbody>
</table>

Table 6.6: Top eight collexemes of the 1-2 order (repeated from Table 6.2), and the top eight verbs with positive residuals. Verbs in **bold** occur in both lists.

The effect of the particle verb factor is stronger than what has been measured by Bloem et al. (2014) and De Sutter (2005). It is also possible that the 1-2-order preference for particle verbs becomes exaggerated for some verbs due to an effect like preferential attachment, which would result in effects greater than the one that is controlled for showing up in the data.

To circumvent this issue, we can examine the top 30 of 2-1 order verbs instead. Of the 30 verbs most associated with the 2-1 order in the collostructional analysis, 73% are also among the top 30 2-1 order verbs ranked by residuals, and the non-overlapping verbs are mostly near the bottom of the list. Furthermore, the two measures strongly correlate ($r_s = .972, p < 0.001$). This suggests that the odds ratios of the collostructional analysis and the residuals of the regression model measure the same thing. The fact that the model residuals show overlap and a strong correlation with the top collexemes indicates that the collostructional analysis method is reliable and that lexical associations are present, even when other factors are controlled for.
6.7 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. The results of the Gain Ratio measure indicate that there are lexical associations that are not accounted for by other factors in the model. Furthermore, 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. I take these preferences to be entrenched associations between main verbs and verb cluster word orders, arising from familiarity (i.e. having encountered this pattern, the combination of verb and word order, before in the input). Such a preference has to be stored in the mental lexicon or constructicon in some way, whether it is as a property of the verb or of the construction that selects it. These lexical preferences contribute to the choice of a word order, along with other factors.

The word orders also appear to have some semantic associations, with the 2-1 order showing an association with negative verbs and cognition verbs in my data. Lexical-semantic considerations generally do not come into play in previous, purely syntactic accounts of verb cluster order variation. 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 the verb cluster 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. Yet in the present study, I did observe an effect of apparent lexical-semantic properties of the main verb on the cluster’s word order. While the effect could also be explained as a semantic selection restriction or dispreference of verbal heads on verbal complements in certain word orders, it is not possible for such restrictions to be conditional on word order in an account where word order is only assigned during PF.

The observed 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 already implies that the word orders are distinct constructions, as in this case each order would have its own form and its own meaning. If a word order is chosen partly due to an association with a specific verb, rather than on the basis of general features of the context (such as its processing difficulty), this must be stored in the grammar or lexicon. It could be stored either as a preference of a word order construction for that verb, or a preference of the verb for a
word order construction. The semantic associations also suggest the existence of differences in semantic preferences of the orders. Therefore, the results suggest that the word orders are distinct linguistic units or constructions, rather than different realizations of an underlying unit.

Of course, this is not the only factor affecting the choice. Another possibility that this study did not explore is that there may also be a functional difference between the two word order constructions, related to information structure or facilitation of linguistic processing: one order could be easier to process than the other (De Sutter, 2009; Bloem et al., 2017b,a).

It is difficult to exclude the possibility that some of the variation my analysis attributed to specific verbs was actually caused by another factor. I excluded various factors from the analysis, as well as performing multifactorial analyses, 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 inherence 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.

Much like De Sutter (2005), I found generally stronger associations of verbs with the 2-1 order, as well as a larger number of associated verbs with the 2-1 order. De Sutter explained this as evidence that the 2-1 order is more lexically specific. However, I believe that it is simply a frequency effect. The 1-2 order is more frequent, therefore the threshold is higher for verbs to be significantly associated with the 1-2 order — it requires a higher frequency. Similarly, the threshold for being significantly associated with the less frequent 2-1 order is lower, and more verbs are significantly associated with it. This observation on the absolute number of significantly associated verbs per word order is therefore not very interesting, in my opinion.

### 6.8 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 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).

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. As collostructional analyses are concerned with syntactic structure, a dependency-based model would be a good avenue to explore. These models have been shown to perform better than distributional semantic models that do not take syntax into account at predicting brain imaging data associated with specific words (Abnar et al., 2018), though nouns were used in that study.

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.

Alternative measures that control for other factors seem to correspond with these results. The results of the collostructional analysis strongly correlated with averaged regression residuals per verb of a multifactorial regression model of the variation, indicating that the collostructional analysis method was fairly reliable even without controlling for other factors. Of course these analyses can also not exclude all other factors that might contribute to an association between a verb and a word order, but the use of a large data set and the inclusion of a comprehensive set of control factors in our analysis provides stronger evidence than previous work on lexical associations.

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.

This study showed that verb cluster orders can have their own associations with other constructions. It might therefore be interesting to explore whether there are other types of associations that might affect verb cluster orders, such as analogical links. Perhaps some subordinate clause verb cluster orders are preferred in some contexts because they are similar to other constructions in those contexts. Examples of this could be subordinate clause clusters that resemble verb groups in main clauses, three-verb clusters that contain verb combinations that are highly frequent in two-verb clusters, or participle-auxiliary verb clusters in the 2-1 order that resemble adjective-auxiliary constructions, where the verb has to come after the adjective.

Lastly, the fact that semantic associations with verb cluster word orders were observed in this study also indicates that it can be interesting to apply collostructional analysis to other phenomena with an unclear form-function mapping, even though the method was intended for studying semantic differences.
An agent-based model of a historical word order change*

Chapter Highlights

Problem Statement

- While verb cluster constructions in West-Germanic share a common history, we currently see a wide variety of possible word orders. It is not clear why English and German developed opposite verb cluster word orders from the same Proto-West-Germanic starting point.

Research Questions

- What factors could have pushed different West-Germanic languages to have different verb cluster word orders?
- Can we create a model of this language change that makes few assumptions and is informed by historical data?

Research Contributions

- We discuss and model changes in the West-Germanic languages that may have affected their verb cluster order. The present word order in German may have developed due to increased use of subordinate clauses, while the English order is predicted to be influenced by the grammaticalization of the verb *to have*.

- We demonstrate the use of agent-based modeling techniques with minimal assumptions as a tool for historical linguistic study, by incorporating historical data both at the starting and ending point of the change.

7.1 Introduction

Agent-based modeling is a method for simulating the behaviour of individual agents (i.e. a speaker of a language) in a larger community of agents (i.e. all speakers of the language). While agent-based models have been successfully used as tools in the field of evolutionary linguistics to study how linguistic structures could have emerged, they have not yet spread to the field of historical linguistics, which is more interested in describing and modeling change in existing natural languages. Both fields are concerned with modelling language change, although the starting assumptions and context of the modeling are different. In historical linguistics there is data available about structures in earlier and more modern states of the language, while in evolutionary linguistics the structures have to emerge from the implemented mechanisms. Nevertheless, the mechanisms described, such as grammaticalization, are often similar and lend themselves to study using similar methodology.

In the field of evolutionary linguistics, agent-based models are used to model language as a complex dynamic system, whose structure depends on the interactions of its speakers. An early overview of such work is provided by Steels (1997), who emphasizes the possibilities of modeling various aspects of language in this way. Among this work is a study by Briscoe (1997) on the default word order of languages, though it assumes a framework of universal grammar in which learning consists of setting parameters. Subsequent work included the application of this method to specific domains of linguistics, such as the emergence of vowel systems (De Boer, 2000) and the development of agent-based models specific to language, such as the iterated learning model of Kirby & Hurford (2002). Language change was often only discussed in terms of the emergence of new structures, and lacked comparisons to historical data (de Boer & Zuidema, 2009). Alternatively, studies used artificial languages, as noted by Choudhury et al. (2007), whose own work is an exception. A few other studies that relate to historical linguistics can be found. Daland et al. (2007) and Van Trijp (2012)
An agent-based model of a historical word order change

model some apparent idiosyncrasies in inflectional paradigms of natural languages, Daland et al. (2007) doing so with a model that includes social structure, and Van Trijp (2012) using the Fluid Construction Grammar framework. A further example is Landsbergen et al. (2010)’s study that models some mechanisms of language change from the perspective of cultural evolution. Overall, agent-based language studies informed by historical data are not widespread, and often involve many assumptions or dependence on a framework. A recent exception to this is a study by Pijpops & Beuls (2015) on Dutch regular and irregular verbs.

Our emphasis in this chapter is on creating an agent-based model that makes minimal assumptions, in order for the presented methodology to be useful for any theory of language that allows for functionalism in language change. Our case study, the historical development of verbal cluster order in Germanic languages, involves a word order variation in which multiple constructions are grammatical. This kind of phenomenon has not been investigated with an agent-based model before. Besides syntactic analyses (Evers, 1975), recent work on verb clusters has also discussed non-syntactic factors influencing word order, using frequency-based methods (De Sutter, 2005; Arfs, 2007b; Bloem et al., 2014) and historical data (Coussé, 2008). We follow up on this line of work with our agent-based model, in which a functional bias induces language change. Using this model, we show how the current orders of verb clusters in modern West-Germanic languages might have developed and diverged from the proto-Germanic cluster orders.

In the next section, we briefly outline the phenomenon of verbal cluster order variation. We then describe the methodology of the simulation and its initial state, followed by the results and a discussion of those results.

7.2 Verbal clusters

Many verbal cluster word orders are attested in different Germanic languages (Wurmbrand, 2006). We will illustrate this with an example from Dutch, a language where the ordering of these verbs is relatively free. In two-verb clusters, the finite verb can be positioned before or after the infinitive:

(7.1)  *Ik denk dat ik heb begrepen.*
       I think that I it have understood
       ‘I think that I have understood it’

(7.2)  *Ik denk dat ik heb begrepen heb.*
       I think that I it understood have
       ‘I think that I have understood it’

In the literature, construction 7.1 is called the 1-2 order (ascending order or red order), and construction 7.2 is called the 2-1 order (descending order or green order). Both orders are grammatical in Dutch, and express the same
7.3 Methodology

Our simulation consists of a group of agents that can function as speakers and recipients of verbal cluster utterances. Each agent has its own instance of a probabilistic language model that stores and produces such utterances (and nothing else). We will first describe the language model and the linguistic features of verbal clusters that it stores, and then we will explain what happens when the simulation is run and the agents interact.

To find linguistic features that may be associated more with one order than with the other, we rely on synchronic corpus studies of Dutch, a language in which both orders are possible. Associations have been found with a variety of factors, including contextual factors such as regional differences between speakers (Coussé et al., 2008). When creating a language model for an agent, we are only interested in factors that may cause a particular speaker (or agent) to
choose a particular word order. It has been shown that verbal cluster order variation correlates with both constructional factors, such as the use of a particular linguistic form, and processing factors, such as sentence length (Bloem et al., 2014). We examine only the constructional factors, because those are likely to be stored in the lexicon with their own associated word order preferences. The most important of these are the main clause / subordinate clause distinction (there are more 2-1 orders in main clauses in present-day Dutch), and the type of auxiliary verb (there are more 2-1 orders when a copula verb is used in a cluster). These two factors not only are associated with different order preferences in modern Dutch, but have also undergone historical changes that may have triggered our word order change. Regarding the first factor, subordinate clauses have become more prevalent, and it has long been assumed that syntactic subordination arose in Old Germanic varieties (Kellner, 1892, p. 51–56) or Proto-Germanic, later continuing to develop in most Germanic languages such as Middle Dutch (Burridge, 1993, p. 43–51). Regarding the second factor, one type of auxiliary verb, *to have*, grammaticalized during the time period we are interested in: Besch & Wolf (2009) note that *to have* auxiliaries to express the perfect tense are not attested in the oldest Old High German texts. Salmons (2018, p. 168) also discusses the rise of the present perfect, and notes that passive auxiliary verbs might have grammaticalized in the same way, as Gothic did not have them but Old High German did.

We will assume that the two factors, clause type and auxiliary type, are stored as features, each with their own word order preferences. This way of storing features is based on the bidirectional model of Versloot (2008), though our models learn by interacting rather than iterating.

Table 7.1 shows all of the possible combinations of feature values a verbal cluster can have in our model. Our model assumes two clause types (main and subordinate) and three different types of auxiliary verbs, reflecting the historical sources of verb clusters:

1. Clause type feature
   (a) Main clause context
   (b) Subordinate clause context

2. Auxiliary type feature

<table>
<thead>
<tr>
<th></th>
<th>mod+inf</th>
<th>have+PP</th>
<th>cop+PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>main</td>
<td>97%</td>
<td>50%</td>
<td>10%</td>
</tr>
<tr>
<td>sub</td>
<td>80%</td>
<td>50%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 7.1: Reconstructed proto-Germanic probabilities for the 1-2 order.
3. Methodology

(a) modal + infinitive: the origin of verb clusters in Germanic

(b) ‘to have’ + participial main verb (PP): arose only later in history to extend the possibilities of expressing temporal and aspectual features

(c) copula + PP: originally a passive, predicative, construction — not purely verbal, rather adjectival.

A cluster can have either of two word orders: the 1-2 and the 2-1 order.

The simulation consists of a language agents, each starting out with n exemplars of verbal clusters, stored in the agent’s language model. An agent’s language model contains the type of information shown in table 7.1: for each possible combination of feature values, exemplars are stored. In addition to their features, exemplars have the property of being either in the 2-1 or 1-2 order. A percentage of 2-1 or 1-2 orders can be calculated over all the exemplars in an agent, or over all the exemplars in all agents, as in table 7.1. The agents’ language models do not contain any other structures or features. We did not use an existing framework in order to have as few parameters and assumptions as possible. The simulation was implemented in the Python programming language.

When the model is run, each run consists of $a \times n \times i$ interactions, in which $i$ is another parameter that governs the number of interactions per agent and exemplar. In an interaction $i$, a random agent from $a$ is picked as the speaker and another random agent as the recipient. The speaker agent generates a verbal cluster based on its language model, and the recipient agent stores it as an exemplar. When a speaker agent generates a verbal cluster, it picks the features of a random exemplar from its language model, and then assigns a word order based on the word order probabilities of both of the exemplar’s features individually. For example, a 1-2 (ascending) realization of a modal subordinate clause cluster may be produced according to the following:

$$P(asc|x) = P(asc|x_{sub}) \cdot P(asc|x_{modinf})$$

(7.1)

where $x$ is a set of feature values. $P(asc|x_{sub})$ is the probability of a subordinate clause being in 1-2 order, and $P(asc|x_{modinf})$ for the modal+infinitive construction type. These probabilities are calculated from the stored frequency of the features in 1-2 contexts:

$$P(asc|x_{sub}) = \frac{F(sub,asc)}{F(sub)}$$

(7.2)

So, the probability of a modal subordinate clause cluster being expressed in the 1-2 order depends on how many exemplars the agent has stored in which a subordinate clause cluster was in the 1-2 order (relative to 2-1), as well as the number of exemplars in which a modal cluster was in the 1-2 order (relative to 2-1). Example 7.3 is a (Dutch) example of a modal subordinate clause cluster in 1-2 order, though our language model is more abstract and does not use actual words, only the features.
An agent-based model of a historical word order change

(7.3)  *Ik denk dat ik het wil horen.*

I think that I it want hear

‘I think that I want to hear it’

After producing an exemplar, the agent normally deletes it from its own storage, because we do not want the relative frequencies of the various feature values (i.e. the number of copular verbs) to vary randomly. We are only interested in the word order. Furthermore, this prevents an endless growth of the agents’ language models. Only when a growth factor applies, this deletion does not happen.

The simulation includes two growth factors $g_{\text{have}}$ and $g_{\text{sub}}$ to simulate the historical changes relevant to the two features in our model. As mentioned previously, ‘to have’ grammaticalized as an auxiliary verb, and the use of subordinate clauses increased, and this is simulated in the following way. When these growth factors are set to 1, after every $a$ interactions, an exemplar with the relevant feature is kept where it otherwise would have been deleted from the language model. A growth factor of 2 doubles the rate to $2 \times a$. $g_{\text{have}}$ applies as long as there are fewer *have*-clusters than clusters of either of the other types in the model, and applies $g_{\text{sub}}$ as long as there are fewer subordinate clause clusters than main clause clusters.

When an agent is the recipient of a verbal cluster exemplar, it simply stores it in its language model, including the word order that it came in. So, when example 7.3 is perceived, the 1-2 order production probability of subordinate clause clusters and that of modal clusters will go up (separately) in the language model of the recipient agent, as that probability is calculated over all relevant features of the stored exemplars. A critical learning bias is simulated here: the tendency to decompose an utterance into features and storing information about the features, rather than storing it as a whole. This is the only assumption we make about the language faculty in this model, and it is a functional one. It simulates the fact that people do not perfectly copy a language from each other.

We initialize each experiment with 30 agents ($a = 30$), and $i = 5000$ to simulate a long time course in which simulations will almost always stabilize in the end. With fewer agents, some agents lose all of their exemplars during the simulation: by random chance, they may be selected to produce exemplars (in which case that exemplar gets deleted from their storage) much more often than they are selected to receive them (in which case they receive an exemplar into their storage). Each agent starts with a language model of 73 exemplars ($n = 73$) that follows frequency patterns as reconstructed for 6th century Germanic, based on a comparison of verb cluster frequencies in Old English, Old High German and Old Frisian texts. These figures are also summarized in table 7.1. For any unattested combination of features and word order a single exemplar is included to simulate noise.

2 The recipient does not delete any exemplar, this only happens in production.
7.4 Results

Figures 7.1 and 7.2 show example results of the agent-based model simulation, with different parameter settings. The graphs show the results of 50 different simulation runs overlaid, each run being a possible language, and the reported proportion is calculated over all agents in the simulation. The X-axis represents time (in number of interactions) and the Y-axis represents the proportion of 1-2 orders, a value between 0 and 1. This proportion is calculated over all of the agents in the simulation. When the simulation is run for long enough, it will always stabilize into a situation where a language either has only 1-2 or only 2-1 orders, though some feature combinations stabilize faster than others. Due to space constraints, we only show results for subordinate clause clusters (with any auxiliary type), but the general patterns are similar for all of the features, though some change sooner than others. We can observe that the model correctly predicts languages with dominant 1-2 orders such as English, as well as languages with dominant 2-1 orders, as in German.

However, a model that predicts everything is not very interesting. We would like to know when a language in the model becomes a 1-2 order language or a 2-1 order language, in other words, English-like or German-like. We can do this by changing the two growth factors: the rise of subordinate clauses ($g_{sub}$) and of to have ($g_{have}$). Figure 7.1 shows simulations in which to have grammaticalizes faster, while in Figure 7.2, the frequency of subordinate clauses rises more quickly. A clear difference can be observed — Figure 7.1 shows more languages gaining English-like 1-2 orders (35% 1-2, 56% 2-1 and the rest had not stabilized yet), while Figure 7.2 shows more German-like 2-1 orders (92% 2-1, 7% 1-2). Different speeds of grammaticalization of to have and growth of subordinate clauses result in different dominant word orders.

<table>
<thead>
<tr>
<th>%descending</th>
<th>mod+inf</th>
<th>habba+PP</th>
<th>cop+PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>main</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>sub</td>
<td>100%</td>
<td>33%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Table 7.2: Descending output probabilities from early-Modern Frisian text of around the year 1550.

<table>
<thead>
<tr>
<th>%descending</th>
<th>mod+inf</th>
<th>habba+PP</th>
<th>cop+PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>main</td>
<td>100%</td>
<td>92%</td>
<td>70%</td>
</tr>
<tr>
<td>sub</td>
<td>98%</td>
<td>33%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Table 7.3: Descending output probabilities of the model, some time before complete convergence.
An agent-based model of a historical word order change

Figure 7.1: 50 runs with faster growth of have+pp constructions ($g_{have} = 2$, $g_{sub} = 1$)

Figure 7.2: 50 runs with faster growth of subordinate clauses ($g_{have} = 1$, $g_{sub} = 2$)
We compared the model to frequency figures for early-Modern Frisian (ca. 1550) once the proportions main clause–subordinate clause and the proportions between the three constructions were comparable to those in our Frisian dataset (table 7.2 and 7.3). As such, the results are promising. Tuning of the model — e.g. by a slower or quicker rise of the amount of subordinate clauses — shows that in the long run it tends to produce 100% ascending or 100% descending realizations for all feature sets. The situation with 100% descending realizations reflects basic word order in German and Frisian, although then V2-movement is needed to get the finite verb in the second position in main clauses. We assume V2 to be a grammaticalised side effect of the asymmetries as reconstructed for Proto-Germanic, where ascending orders were dominant in the combination of modal+inf, which happened to occur more often in what we call main clauses from the modern perspective than in subordinate clauses. The current model is probably too crude to model more complex word orders.

7.4.1 Dutch

This model cannot yet account for the current state of the Dutch language, which first moved towards mainly 2-1 orders like German, and then shifted towards 1-2 orders again starting from the 16th century (Coussé, 2008), a change that is still in progress (Olthof et al., 2017). There is evidence that the 1-2 order has become the default order (Meyer & Weerman, 2016), and this second change was likely caused by a factor outside the scope of our basic model, such as language contact. Clearly, it is possible to simulate language contact in an agent-based model. One naive way of doing this would be to add, at a certain point in the simulation, a group of new agents with different word order preferences. Adding this feature to our simulation gives us the result shown in figure 7.3 for all types of subordinate clauses. The graph shows, again, the 1-2 order probabilities computed over all subordinate clause clusters of all the agents in each of 50 simulations, over time (in number of interactions). To create the graph, the simulation was sampled 20 times throughout its running time, and the lines are drawn between these 20 samples.

In these 50 simulations, after 1500 * a interactions, we spawn \( \frac{2}{5} \) additional agents that have a ‘neutral’ language model. With this neutral language model, agents have a 50% chance of using the 1-2 or 2-1 order for any feature in the model. In terms of the model, they have 5 exemplars of each possible combination of features. In figure 7.3, these agents appear at the sampling point marked with the asterisk (*). The sample before the marked point already shows some simulations that have converged to a language with 0% 1-2 orders in the subordinate clause, but after the additional agents are spawned at the marked point, none of the simulation runs has 0% 1-2 orders, as a population with 50% 1-2 orders was just introduced. Subsequently, the simulations continue to stabilize. We can see that even with this naive approach to language contact, there is an effect. In some simulations the proportion of 1-2 orders decreases until the marked point where the ‘contact’ happens, and increases again after
the language contact, converging to a 1-2 order language after all.

To improve upon this, instead of a neutral language model, one would want the ‘contact’ agents to have the language model of the contact language. But this is where the problems start. We do not have a hypothesis on what language Dutch might have had contact with to cause this word order change. It could also have been more than one language, or it could have been related to language standardization that happened around this time. There are many other problems with this approach, centered around the fact that language contact is usually not as simple as a sudden wave of speakers of another language appearing in the population. Even if there was a sudden wave of speakers, if they were adult speakers their own language would not change much, only capable of late language acquisition, and this would have to be reflected in the model. If they were child speakers, there would be early language acquisition, and their language would be able to change. Our model does not presently simulate this, as the agents are immortal and do not exhibit different behaviour as they age. Then there’s the possibility of other types of language contact besides sudden total immersion, which might involve subgroups of agents with different language models that are not completely connected to the other agents, or simulating the effect of ‘learned contact’, i.e. education in French. French uses the 1-2 order in its verb groups, and it was a prestige language in the Netherlands especially
between the 17th and the 19th century, as it was elsewhere in Europe. Furthermore, French was an important second language especially in the southern part of the Dutch-speaking area, which may have affected the language of the rest of the Dutch-speaking population. In a full simulation of the phenomenon, one would want to consider all of these factors.

All in all, further study and further development of the model is required to accurately study the change of direction that occurred in the development of Dutch verbal cluster order.

7.5 Discussion

With this study, we hope to have shown that an agent-based model with just a single learning bias can be used to gain insight into processes of language change, and to generate new hypotheses. Specifically, the model makes two predictions: that *to have* grammaticalized faster in English, and that subordinate clauses gained use more quickly in German. These predictions can be tested using historical corpora of these languages in future work.

In the model, the 2-1 order is supported by subordinate clauses. Due to the verb-second (V2) effect in these languages, the finite verb (the 1) precedes the other verb in main clauses (the 2). This 1-2 order differentiates main clauses from subordinate clauses, motivating the preservation of a 2-1 order in the subordinate clauses. Increased use of subordinate clauses may then have supported the 2-1 order as the default order. However, if *to have* grammaticalizes earlier in the model, the 1-2 order is supported. This new grammatical verb becomes associated with the most prevalent word order at the time, and pushes the language further in the direction of that word order. In the beginning this is the 1-2 order, more associated with main clauses in proto-West-Germanic due to V2 movement, but later on the 2-1 order is more prevalent, due to its association with subordinate clauses.

Overall, it can be concluded that the interaction of basic probabilistic choices of constructions with shifting input frequencies and shifting usage preferences of related constructions may be a key to understanding different word orders in the Germanic languages.

With our approach, in which we make minimal assumptions, only a single change can be modeled at a time. Furthermore, other factors that may affect the usage of a particular construction such as sociolinguistic factors or the processing complexity of the construction cannot be modeled. Cultural or demographic phenomena among agents such as language contact also cannot be modeled properly at present, though the model could be extended with such features in future work.

Nevertheless, we believe that agent-based modeling can be a useful tool for historical linguists, particularly those working with frequency-based explanations. The present work and the study of Pijpops & Beuls (2015) show that testing of different mechanisms and parameters in a simulation, informed by
historical data, can provide additional evidence for theories on what may or may not have been possible in a particular instance of language change, given the assumptions built into the model. We believe it is particularly interesting to test how few assumptions are necessary to explain the observed historical data, which is something that previous work has not focused on.

We would like to emphasize that this method is applicable to other cases of language change in which the use of structures changed over time. Any processes of historical change that can be captured in terms of frequencies and features may be used as factors to be investigated, and the fact that the model makes few assumptions also means that no particular social or cultural phenomena need to have happened for the model to be applicable. However, these simplifications also limit the extend of what can be modeled. In future work, contact phenomena could be simulated by including non-learning agents, or influxes of agents with different language models. Subsequent work on other cases of historical change may need to include such additional assumptions, if they are known to have been historically relevant.
Learned borrowing or contact-induced change: Verb cluster word order in Early-Modern Frisian

Chapter Highlights

Problem Statement

• Recently, word order variation in two-verb clusters has become more common in Frisian, while only one word order is considered normative. However, in older Frisian texts, this ‘ungrammatical’ auxiliary-first order has also been attested.

Research Questions

• Is the present-day use of this auxiliary-first word order in Frisian really a new development taken from Dutch, or something older?

• Is the use of this auxiliary-first word order in Early-Modern Frisian texts a contact effect, or a continuation of an older Germanic phenomenon?

8.1 Introduction

Verb clusters display much word order variation in the West Germanic languages. In Frisian,\(^1\) when there are two verbs in a cluster (an auxiliary verb and a main verb), the order of verbs is prescriptively considered to be fixed. However, in practice it appears that both logically possible verb cluster orders are now being used in Frisian:

\[(8.1)\]  
\begin{align*}
\text{Anne} & \text{ sei dat er my sjoen hie.} \\
\text{Anne said that he me seen had} \\
\text{‘Anne said that he had seen me’}
\end{align*}

\[(8.2)\]  
\begin{align*}
\text{Anne} & \text{ sei dat er my hie sjoen.} \\
\text{Anne said that he me had seen} \\
\text{‘Anne said that he had seen me’}
\end{align*}

---

\(^1\)In this article, we use the term Frisian to refer to the West-Frisian language (Westerslauwers Fries), the variety of Frisian spoken in the Dutch province of Fryslân, as opposed to Saterland Frisian or North Frisian, spoken in parts of northern Germany. The West-Frisian language should also not be confused with the West Frisian dialect of Dutch, which is a Hollandic Dutch dialect influenced by the West Frisian language.
Example 8.1 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). Example 8.2 shows the 1-2 order, where the head verb comes first. This terminology can also be extended to larger clusters — a 1-2-3-4 cluster is head-first and has the main verb last. A cluster that is completely head-first is called an ascending cluster (the numbers go up), a completely head-final cluster is descending (2-1, 3-2-1 etc.).

Prescriptively, only the order in example 8.1 is considered normative, but Frisian speakers now use two-verb clusters in the 1-2 order as well, as in example 8.2. This situation is similar to modern Dutch, where the main verb can be used in both positions. The change in Frisian verb cluster order appears to be recent, and influenced by language contact with Dutch (de Haan, 1996). It has even been found that Frisian bilingual children have similar word order preferences in their Frisian as in their Dutch, producing both orders (Meyer et al., 2015b). However, the non-normative 1-2 order also appears in older Frisian texts. For example, in the Oudfriese Oorkonden (late mediaeval charters), written between 1397 and 1545, this order is used 40% of the time, in a sample of 810 verb clusters (Larooij, 1991).

This raises some questions on the nature of this change in verb cluster order preferences. Is the modern use of this word order actually a new development taken from Dutch, a consequence of hundreds of years of language contact, or a continuation of an older language-internal development? And what kinds of language contact between Dutch and Frisian may have affected this change?

To study this, we have to consider an intermediate stage of the language, which is Early-Modern Frisian. Studying a 17th century essay from the time period in which Early-Modern Frisian was spoken, Hoekstra (2012) found that 10% of the verb clusters in this text are in the 1-2 order, taking this as a Dutch contact effect. But can it be a contact effect when the percentage of ‘Dutch-like’ orders is no different than in older texts? We would expect word order preferences between the languages to converge in the case of contact effects. The percentage of 1-2 orders should then be a quantity that is between the 20-40% of the Old Frisian charters and the 70% of modern day written Dutch. Alternatively, this finding of 10% 1-2 orders in the 17th century might be a continuation of verb cluster order variation as an older Germanic phenomenon, as found in the Oudfriese Oorkonden. And if it is a contact effect, why are the relative frequencies of 1-2 order clusters in these early-modern texts much lower than among young bilingual speakers of Frisian in the 21st century?

We present a corpus analysis of Early-Modern Frisian verb clusters as an investigation of the nature of the variation that occurs in the texts. To account for the 1-2 orders that are used in Early-Modern Frisian texts, we are particularly interested in the contexts in which this order is used in the Early-Modern Frisian
corpus, as this may help us to identify the source of this construction in Early-Modern Frisian.

We will proceed as follows. In section 8.2, we discuss verb cluster word order variation in Frisian more broadly, describing verb cluster order in earlier and later stages of Frisian to identify possible sources of variation. We end the section with three possible hypotheses on where the order variation found in Early-Modern Frisian texts might have come from. Next, in section 8.3, we discuss the nature of the variation in Dutch, which is necessary to understand any Dutch-Frisian contact effects that we might find. In section 8.4 we present the corpus and methodology of our study, in which we gather data on Early-Modern Frisian verb cluster orders. Results are presented in section 8.5, followed by a discussion in section 8.6 of the results in relation to the hypotheses on the source of the variation, and a conclusion in section 8.7.

8.2 Verb clusters in Frisian

In this section, we will identify possible sources for the verb cluster order variation in Early-Modern Frisian by examining verb clusters and language contact in Frisian more broadly. We first summarize what is known about verb cluster order in Modern Frisian, and then go back in time to verb clusters in historical varieties of Frisian in section 8.2.2. In section 8.2.3 we move to the topic of contact in historical Frisian texts by discussing the literature on texts containing mixed Middle Frisian. This literature proposes two possible contact situations that may have led to these mixed texts, which we analyse in section 8.2.4. Then, we move back to the topic of verb clusters and investigate how they are used in these Middle Frisian mixed texts in section 8.2.5. Lastly, all of this leads us to three possible sources for the use of 1-2 orders in Early-Modern Frisian texts, which are proposed in section 8.2.6.

8.2.1 Modern Frisian

Like other West-Germanic languages, Frisian expresses properties such as tense and aspect by means of auxiliary verbs. Frisian is a verb-final language (SOV), so these verbs end up clustered together at the end of the sentence in an embedded context. The auxiliary verb, which is the syntactic head of the cluster, is then typically in the final position, as in the example below (repeated from 8.1):

\[(8.3) \text{Anne sei dat er my sjoen hie.}
  \]
  \[\begin{array}{l}
  \text{Anne said that he me seen \ had} \\
  \text{‘Anne said that he had seen me’}
  \end{array} \]

If we were to base this section on verb cluster order in Frisian only on the rules listed in reference grammars, it would be a short section. Verb cluster order in Modern Frisian is typically described as fixed. According to most older speakers of Frisian, only this 2-1 verb cluster order should be used. In
fact, modern reference grammars still explicitly exclude verb cluster variation in Frisian. For example, in the Frisian grammar of Popkema (2006), it is mentioned that there is order variation in the Dutch verbal end group in some cases, but that this freedom is not present in Frisian (Popkema, 2006, p. 247). It is explained that only the 2-1 order should be used. In practice however, this is not the whole story. It has been shown that language contact between Frisian and Dutch has also affected the word order of verb clusters that speakers use in their Frisian: both word orders are used, as in Dutch.

The linguistic literature on modern Frisian usually describes it as a language that is affected by language contact. The variety of Frisian that has undergone contact-induced change has been called Interference Frisian (de Haan, 1997). According to de Haan (1997), Frisian is undergoing *interference under full bilingualism*, as opposed to situations where a first language affects a second language, or vice versa. This is because all speakers of Frisian also speak Dutch and are thus fully bilingual.

In Interference Frisian, both 2-1 and 1-2 verb orders are used and they are ‘accepted by an increasing number of Frisian speakers’ (de Haan, 1997, pp. 289-290). It has been noted that this is not just a case of copying the Dutch system of verb cluster ordering. Taking Frisian verb clusters of more than two verbs under consideration, Koeneman & Postma (2006) tested for the existence of hybrid word orders by means of a grammaticality judgement task. They find that hybrid orders are indeed accepted by Frisian secondary school children with two Frisian-speaking parents, yet less frequently than the orders that are typical in standard Dutch and standard Frisian. They argue that some of these hybrid orders can only exist by mixing grammatical parameters from the two standard languages. Alternatively, Hoekstra & Versloot (2016) argue that the acceptance of these hybrid orders can also be analysed as a consequence of frequency and similarity effects, without assuming grammatical parameters. In their study, the grammaticality judgement data is best modeled when Dutch and Frisian constructions are assumed to be equally frequent in the children’s language input. Even though the two analyses are based on different theories of syntax, both findings suggest that modern Frisian verb cluster order preferences are shifting due to a form of language contact in which constructions from two languages are being combined.

While grammaticality judgement evidence may not be reliable for a construction with such strong normative biases as Frisian verb cluster order, there is more direct evidence as well. Further evidence for this shift comes from Meyer et al. (2015b), who conducted a sentence repetition study with Dutch and Frisian two-verb clusters among Frisian bilingual children to investigate this ongoing language change. They compare the childrens’ performance on the two languages, and also compare their performance to that of monolingual Dutch children. They found a large amount of Dutch interference in the Frisian task, with the bilingual children having almost the same word order preferences as in the Dutch task. Sometimes, Dutch verbs were also used in the otherwise Frisian sentences. On the other hand, only marginal amounts of interference
were found in the Dutch task: the children had a slightly greater preference for 2-1 orders than monolingual Dutch children. These results indicate that the contact-induced change in verb cluster orders has progressed quite far.

### 8.2.2 Historical varieties of Frisian

It does not seem to be the case that the 1-2 order was always ungrammatical until recent times. In Middle Frisian texts, written before the order preferences in Dutch started shifting towards the 1-2 order, this non-normative 1-2 order also appears. For instance, in the Elder Skelteneriacht from 1485, the 1-2 order is used about 10% of the time (Bor, 1971). Furthermore, as mentioned in the introduction, in the Oudfriese Oorkonden, written between 1397 and 1545, this order is used 40% of the time according to Larooij (1991), with a sample size of 810 verb clusters. However, Larooij (1991) does not take the Middle Dutch interference in these texts into account, which may account for some of the 1-2 orders. This interference often takes the form of formulaic opening and closing sentences that show borrowed lexical elements: in the Old Frisian charters from 1397-1460, 52% of the closing sequences contain Dutch lexical elements, such as *ons* instead of *uws* for ‘our’, though in later years this percentage decreases. If such borrowing also took place within the verb clusters that occur in the formulaic opening sequences, the borrowed clusters would have been in the 1-2 order, as this order was prevalent in Middle Dutch charter opening sequences (Boonen, 2007).

In Old High German and Old English texts, verb cluster orders vary as well, from which Bloem et al. (2015) inferred that Proto-West-Germanic exhibited some degree of verb cluster order variation. Their reconstruction assumes particularly free word order options for verb clusters with ‘to have’ as an auxiliary verb, as this construction was just grammaticalizing at the time. The large degree of variation with this verb was also noted by Larooij’s (1991) study of the Old Frisian charters. These studies show that verb cluster order variation may be an older Germanic phenomenon that has disappeared in some varieties (English and 20th century Frisian) or in some cases only in the case of two-verb clusters (German).

Another reference point on the diachronic timeline of verb cluster variation is the study by Hoekstra (2012), who conducted a study of verb clusters in a 17th century essay by the Frisian author Gysbert Japicx written in Early-Modern Frisian: the largest Frisian prose text from the 17th and 18th century. Hoekstra found that 10% of the verb clusters in this text are clusters in the 1-2 order (or ascending, in the case of larger clusters), and that these clusters exhibit some properties that typical Frisian 2-1 ordered clusters do not show. Most notably, they show the Infinitivus-pro-Participio (IPP) effect, where an infinitival verb is used in middle positions of a cluster, when a participial verb might be expected. This phenomenon is normally found in Dutch and German. Therefore, Hoekstra (2012) takes these 1-2 clusters as a Dutch contact effect. Other factors noted by Hoekstra (2012) to be associated with ascending orders
are larger clusters, ‘te’-infinitive constructions, causative dwaan ‘to do’ as an auxiliary verb, and infinitival main verbs. Hoekstra (2012) argues that these factors are also associated with Dutch interference, particularly in the written modality. For example, larger clusters are more commonly ascending (i.e. an 1-2-3 order cluster), and larger clusters are more common in written Dutch than in spoken Frisian. Therefore, the use of large ascending clusters in written Frisian may be influenced by written Dutch. This link can be assumed because the author was bilingual, authored many pieces in Dutch, and had presumably learned to write in Dutch, as there was not much of a Frisian written tradition at this time.

If 1-2 order verb clusters are used in modern Frisian, in Early-Modern Frisian, and also in Middle Frisian, there is the possibility that the verb cluster order variation that is currently being observed in Interference Frisian was always there as a language-internal phenomenon. However, this idea that the 1-2 order has been in continuous use seems unlikely given the fact that it is now considered non-normative in reference grammars and according to many Frisians. There is indeed evidence to suggest that the 1-2 order was not used for some time after the Early-Modern Frisian texts were written and before approximately the 1970s or 1980s. Wolf (1996, p. 39) presents data from a series of studies in the early 90s, in which children in their final year of elementary school (around age 12) show 2-1 order preferences of about 40% to 60%, while their parents use the 2-1 order almost exclusively.

Some other relevant data can be found in Pauwels’s (1953) dialect survey, which presents the results of a questionnaire with 15 sentences containing verb clusters in subordinate clauses with various auxiliary verbs. This questionnaire was sent out for translation to speakers of local dialects throughout the Dutch language area. Pauwels concluded that the 2-1 order is nearly always used in Frisia, with exceptions mainly found in the southeast part of the province of Frisia, in which the Low Saxon Stellingwarfs dialect is spoken. Unfortunately, results were only presented by province, meaning that an exact percentage of 2-1 orders for only the Frisian-speaking part of Frisia is not available.

The Dynamische Syntactische Atlas van de Nederlandse Dialecten (Dynamic Syntactic Atlas of Dutch dialects, DynaSAND, Barbiers et al., 2006) is an atlas of syntactic constructions, based on transcriptions of dialect speech collected throughout the Netherlands and Flanders, including the Frisian speaking areas. The data was collected between 2000-2002 from older speakers, mostly born before 1940. The DynaSAND does not show any 1-2 orders in Frisian-speaking areas. For three-verb clusters, most examples are also fully descending, though two of the four test sentences show exceptions. This provides further evidence that the 1-2 order was unlikely to be used during most of these speakers’ life.

3In four test sentences focusing on two-verb clusters: verteld hat ‘has told’, roepen ha ‘have called’, stoutn is ‘has died’, zien meist ‘may see’.
4The test sentences zwemme kinne moat ‘swim can must’ and zwemme gongen is ‘swim went is’ show some exceptions where the auxiliary is placed first, the other test sentences are roepen kinne hie ‘call can had’ and make ha moat ‘made have must’.
Another source during this time period before Interference Frisian appeared is a short article by van der Meulen (1937), who claimed that both Stedfrysk (Town Frisian) and Lânfrysk (Frisian) have 2-1 as a fixed rule, i.e. no verb cluster order variation. Stedfrysk is a set of dialects that are spoken in certain larger Frisian towns, which have a vocabulary that is more similar to Hollandic dialects and a grammar that is more similar to Frisian. Therefore, this observation is surprising — one might expect to find a Hollandic phenomenon such as 1-2 orders in these dialects. However, van der Meulen (1937) also notes that both orders were still used in an older Stedfrysk text, A. Jeltema’s Het vermaak der Slagterij from 1768, with ascending orders appearing particularly in larger verb clusters. van der Meulen (1937) thus concludes from this that the fixed order in Stedfrysk is a fairly recent phenomenon. If 1-2 orders had disappeared even in this ‘mixed language’ of Stedfrysk, then it seems to us quite likely that they were also not used in Frisian at this time.

When discussing older Frisian texts, we must take the Frisian language’s history of contact into account. During the time period in which the available texts were written and throughout its history, Frisian was a lesser used language, interacting with various culturally dominant languages. During the Middle Frisian period from about the 14th to the 16th century these languages were Middle Low German and Middle Dutch, and later, modern Dutch. Potential evidence for early language contact is the existence of ‘mixed’ Frisian texts, which contain Middle Frisian mixed with word forms from superstrate languages that were spoken in the area. Several authors have discussed these texts and what kind of language contact might have given rise to them. In the next section 8.2.3 we summarize this literature, and in section 8.2.4 we present our view of the types of contact that these studies imply.

8.2.3 ‘Mixed Frisian’ texts and language contact

A particularly interesting Middle Frisian set of texts with regards to language contact are the Basle Wedding Speeches, initially annotated and edited by Buma (1957), who noted that the Middle Frisian in these texts is mixed with Middle Low German and Middle Dutch forms. Later, Blom (2008) also discussed some relevant sociolinguistic aspects of these texts, analysing deviant elements and comparing them to those in another mixed Frisian text from the fifteenth century, Thet Freske Riem. The Basle texts are interesting because they are wedding speeches, a register that is different from the legal documents that make up most of the extant Middle Frisian texts. Furthermore one of the three texts looks like it was written “in some haste” (Blom, 2008, p. 1). This is an indication of the use of a less formal register.

The language used in this manuscript appears to be a clear case of ‘contact’ Middle Frisian, influenced by Middle Dutch and Middle Low German. Several

Interestingly, this matches Hoekstra’s (2012) observation that larger clusters in Gysbert Japicx’s Early-Modern Frisian are also more likely to be in the ascending order.
authors have written about the possible nature of this contact. Bremmer (1997) argues that the writer, who calls himself Bernard of Roordahlzum (in West Frisia), is a bilingual with “a full command neither of Frisian nor Low German, certainly not in his writing, nor in all likelihood in his spoken usage” (Bremmer, 1997, p. 383). He comes to this conclusion after analysing another mixed text by the author of the Basle Wedding Speeches, the Life of St Hubert, which is a Low German text with Frisian elements, rather than a Frisian text with Low German elements. Bremmer claims that the Bernard’s mixed writing reflects a language change that is ongoing at the time:

Bernard’s use of Low German is an early witness of language mixing that has resulted in the Frisian substratum underlying the present-day Low German dialects of the former Frisian speaking areas of Groningen and Ostfriesland. In Bernard’s Frisian we see a process of language shift which, especially when large numbers of speakers are concerned, marks the beginning of language loss which will eventually lead to language death (Bremmer, 1997, p. 383).

This type of contact may also have resulted in the mixed language in the Basle Wedding Speeches, although it is unclear to us why a writer from West Frisia would be representative of language change in the more eastern areas mentioned in Bremmer’s quote above.

Blom (2008) proposes an alternative explanation, supported by an older study of Thet Freske Riiim (Campbell, 1950). Campbell finds a similar language mixture there, and suggests that it is comparable to the language of the Basle Wedding Speeches. This suggests that “such a mixture was normal in Frisian at the time”, and Campbell calls it “the curious mixed Frisian of the fifteenth century”, referring to written language (Campbell, 1950, p. 208). Blom (2008) finds that the two texts indeed have similar features, and takes this as evidence for the existence of a shared written register in which using borrowed forms was normal, a register that was used in late mediaeval Frisia. It is plausible that literate mediaeval Frisians were multilingual, because their works show familiarity with texts written in Middle Dutch and Middle Low German. This familiarity may have influenced their written Frisian (Blom, 2008, p. 21). After all, an author or scribe was most likely to be influenced by the language in which he had been trained to write (Blom, 2008, p. 14). Blom illustrates this with a quote from Bischoff (1985, p. 1263): ”im allgemeinen schreibt man nicht, wie man spricht, sondern wie man zu schreiben gelernt hat” [“Generally one does not write in the way one speaks, but in the way one has learned to write”].

8.2.4 Language contact, acquisition and change

The proposals by Bremmer and Blom, that of extensive mixing of the languages in general and that of a mixed written register among the literate Frisians, describe two possible contact situations that might have given rise to the mixed language in the texts. In this section, we will link these contact situations to
8.2. Verb clusters in Frisian

theories on language acquisition and change, in order to learn how these two types of language contact might be reflected in a text.

In the first proposal by Bremmer (1997), the contact has the form of widespread multilingualism, and in the second proposal by Blom (2008), the contact is a limited form of ‘learned contact’ affecting only the literate classes. These two proposals also roughly correspond to two kinds of language change that have been distinguished in the literature: change from below and change from above (Labov, 1965, 1994), where ‘below’ and ‘above’ refer to the level of awareness of the phenomenon. Furthermore, they correspond to two types of language acquisition: early acquisition and late acquisition (Weerman, 2011).

In a situation of widespread multilingualism, it is likely that children learn both languages when they are young, as is the case with Frisian and Dutch in modern-day Frisia — this is early acquisition. Learned contact takes place when a speaker is older and anything the speaker learns in this way can be considered late language acquisition, even if the speaker has already acquired the language early on. Late acquisition can take place in a speaker’s first language (adult L1 acquisition), as well as in acquisition of a second language (L2 acquisition) (Weerman, 2011). Early acquisition is characterized by (nearly) perfect acquisition of the input structures, while late acquisition may be imperfect, for example with regards to opaque aspects of paradigms or previously unknown phonemes.

In the situation described by Bremmer (1997), the Frisian spoken in the region in which the author lived would have changed ‘from below’: the contact phenomena might have become a part of the speakers’ language systems without being marked in some way, and would have appeared in the texts as a consequence of early acquisition by the author. Bremmer (1997) notes that a Low German text by the same author in the same manuscript also contains some Frisian forms, though less than Low German forms in the Frisian Basle Wedding Speeches, indicating a lack of awareness of the language mixing. In the situation described by Blom (2008), the language mixture would instead come ‘from above’ reflecting the writing education of the author, a form of late acquisition by the author.

There are a few factors that characterize these two different forms of language change. As noted by Weerman et al. (2013), changes from above will first become apparent in formal language, while changes from below will first become apparent in informal texts. However, this factor is not very relevant when studying Middle Frisian, as all extant texts from this time period are written in a formal register. Another possible difference is that language changes from below are likely to be more internalized in the language system, as they are acquired at a younger age. If a bilingual speaker of some form of ‘interference Frisian’ uses both verb cluster orders, but is not aware of this because it is normal to hear both orders in their input, it is simply a part of their grammar, just as it would be for a native speaker of the source language (i.e. Dutch or Low German). The use of the two orders will be subject to the same cognitive constraints that it is subject to in the source language, and is likely to follow typical usage patterns from the source language. On the other hand, if a Frisian writer who learned
to write in Dutch uses both verb cluster orders because he noticed that the 1-2 order is used in Dutch writing in addition to the 2-1 order, or uses some other construction which does not exist otherwise in the recipient language, this will not be internalized in the same way. Instead, the 1-2 order would be used as a stylistically marked option, used only in fixed constructions that are borrowed from Dutch, or used in different contexts than in the Dutch source language. For example, an author using learned borrowings, whose language has changed ‘from above’, might not have picked up on subtle patterns such as the fact that the 1-2 order is used much more frequently with infinitival main verbs than with participial main verbs, and therefore overuse it with participial main verbs or underuse it with infinitival main verbs.

These studies of the *Basle Wedding Speeches* and other mixed texts have not addressed verb cluster order specifically. As mentioned in the introduction, for verb clusters there is another option besides these two forms of language contact: that the use of 1-2 orders is a continuation of an older language-internal development. This is not unlikely, because Old High German and Old English have both orders as well, and therefore Proto-West-Germanic was likely to have both orders. Perhaps surprisingly, in this case we also expect to see similar usage patterns as in modern Dutch: after all, these Proto-West-Germanic verb clusters would be subject to the same cognitive constraints as their modern Dutch counterparts. However, in this case we would not observe any other contact effects, such as borrowed verb cluster constructions or an association between borrowed words with Dutch origins and the (potentially borrowed) 1-2 order.

Knowing about verb cluster order variation in Middle Frisian situation in a text with clear interference effects may help to understand the subsequent Early-Modern Frisian situation. In the case of language contact, it might be expected that the 1-2 order is used more frequently in clusters where Dutch lexical forms are used, which is something that can be observed in a mixed text. Do the 1-2 orders used in the Wedding Speeches show signs of contact, or could they be an option left over from Proto-West-Germanic? Before investigating the contact hypothesis, to investigate the plausibility of the language-internal hypothesis, we will now present a brief corpus analysis of verb clusters in the Middle-Frisian mixed texts of the *Basle Wedding Speeches*, and possible relationships between word order and the source languages of the words.

### 8.2.5 Verb clusters in ‘Mixed Frisian’

The *Basle Wedding Speeches* consist of three texts. In these texts we found 67 verbal clusters: 64 comprised of two verbs (being either two verbs in subordinate clauses or three verbs in main clauses, with one in the V2-position), 3 comprised of three verbs. For most of the verb clusters, the language could fairly easily be identified, using phonological and morphological criteria, e.g. *laten staan* ‘let stand’ is Dutch, *leta stan* ‘idem’ is Frisian, *heeft ghesproghen* ‘has spoken’ is Dutch, *spreka mey* ‘may speak’ is Frisian. For two clusters, the language
remained undefined.

Text I had 34% Dutch verb clusters \((n = 29)\), the other two 20% and 12% \((n = 10, 26)\). However, this contrast was not statistically significant \((\chi^2 = 4.13; df = 2; p = 0.13)\). This implies that there is no point in comparing the proportion between ascending and descending word order as a reflection of the overall language profile of the texts.

We also considered a potential relation between the language of the individual verbal clusters and the word order. Among the 62 two-word clusters with an identifiable language, there was no discernible contrast, with 38% ascending word orders in ‘Frisian’ clusters, against 33% in Dutch \((n = 47, 15; \text{ Fisher’s Exact Test, } df = 1; p = 0.73)\).

The hypothesis that there would be a relation between the linguistic origin of the language material from the perspective of phonology and morphology and the word order in verbal clusters is not supported by this data. Furthermore, the overall percentage of 38% ascending word order in two verb clusters \((n = 64)\) in this text with a lot of interferences from non-Frisian varieties is not particularly high, compared to the percentages in purely Frisian texts from the same period.\(^6\) This makes it less likely that the use of the ascending word order in late-Mediaeval and early-Modern Frisian should be ascribed to language contact with Dutch and Low German.

We also considered the relation between word order and construction type. We considered the following constructions: aspectual \(\text{gaet sitta goes sit = ‘to sit down’}\), modal auxiliary, verbal past participle with ‘to have’ \(\text{habba}\), past participle with ‘to be’ \(\text{wessa}\), passive with ‘to become’ \(\text{werda}\), present participle with ‘to be’ \(\text{libben sint living are = ‘are alive’}\). A \(\chi^2\) test over these six categories showed a statistically significant difference \((\chi^2 = 12.54; df = 5; p = 0.03)\). However, this result may not be reliable due to missing values: there are no passive constructions or ‘to be’ present participles in the ascending order. From other research (Versloot et al., 2012; Bloem et al., 2017b) it is known that these subcategories can be grouped into larger groups: modals and aspectual verbs tend to trigger ascending word order, past participles with ‘to have’ take an intermediate position, while past or present participles with copula may be considered of a rather adjectival nature and are more inclined towards descending word order in Dutch. The percentages of ascending word order for these three groups are: 58% \((n = 26)\), 33% \((n = 18)\), 15% \((n = 13)\). A \(\chi^2\) test over these three categories revealed a statistically significant bias in the distribution \((\chi^2 = 8.98; df = 2; p = 0.01)\).\(^7\)

None of the three three-verb clusters found in the texts were entirely de-

\(^{6}\)It is low in comparison with Frisian poetry and the Middle Frisian charters, but it is substantially higher than in various Middle and Early Modern prose texts. Most of the speeches consist of prose, with some scattered portions of local rhyming (without developing a real verse form).

\(^{7}\)When taking all perfect tense constructions with a past participle with either ‘to have’ or ‘to be’ together as one category, this also results in a significant skewed distribution with \(p = 0.002\).
Verb cluster word order in Early-Modern Frisian

ascending (*3-2-1) as in 20th century Modern West Frisian. All three clusters consisted of a finite form of a modal and werda ‘to become’ as the middle verb. The order was entirely ascending 1-2-3 (moete werde vergarret ‘has to be gathered’) one time, and in the 3-1-2 order two times. From these results, we can conclude the following:

1. The preference for ascending word order was not correlated with the use of Dutch word forms in these texts, neither on the level of the texts, nor on the level of the individual utterances. This makes it unlikely that ascending word orders found in Middle Frisian texts appear due to language contact with Dutch;

2. The author of the three speeches shows an inclination towards ascending word order of 38%, which is not uncommon for Frisian texts with fewer lexical and phonological Dutch interferences of that time;

3. The use of the word order is strongly regulated along the same cline as in Early Modern and Present Day Dutch, with a high level of application in combination with modal verbs and aspecltual verbs.

Point 1 discards specific Dutch influence in the choice for ascending word order as an expression of skilled bilingualism with different word orders for different languages (which would be a non-interference scenario). Point 2 makes it unlikely that the ascending word orders in these three texts are a specific feature of the author’s interference Frisian. Even if his language admixture was the result of an early bilingualism, his word order in verbal clusters did not differ substantially from the word order in other Frisian texts of those days, which, given their linguistic character, were written by people with much less linguistic interference. Point 3 suggests that the choice for ascending word order ran along similar lines as in Dutch, where we assume that it reflects aspects of sentence processing (as proposed and discussed by De Sutter 2009; Bloem et al. 2017b; Bloem 2016c). In case of a learned borrowing, one expects a random application or at least profoundly different rates and clines of application than in the donor language. Therefore, it is unlikely that borrowing was the decisive factor resulting in the verb cluster word orders found in this mixed text: neither full bilingualism with early acquisition, nor learned borrowing with late acquisition of the ascending verb cluster orders, but rather an older Germanic word order option that was still available at this time.

Now that we have discussed verb clusters before Early-Modern Frisian, and possible sources of the 1-2 order in Early-Modern Frisian, we can formulate several hypotheses about where the 1-2 order, and with that the variation, in Early-Modern Frisian might have come from.

8One might also consider an effect the other way around: from word order to phonological shape of the words; however, the priming effect may work two ways with the same conclusion.
8.2.6 Hypotheses

On the basis of the literature and data just discussed, we can sketch a timeline of verb cluster order variation in Frisian. To start, it is likely that Proto-West-Germanic had verb cluster order variation. Next, we just saw that that mediaeval Frisian texts show variation, but most likely not due to language contact at all. From Hoekstra (2012), we know that Early-Modern Frisian texts show variation, but not why. Lastly, from other work, we know that modern Frisian has verb cluster order variation that can be attributed to language contact with Dutch due to full bilingualism.

This leaves us with a knowledge gap at Early-Modern Frisian, for which there are three possibilities:

1. Variation in Early-Modern Frisian texts is not due to contact, but a continuation of the mediaeval situation, as in the mixed Frisian of the Basle Wedding Speeches.

2. Variation in Early-Modern Frisian texts is due to contact through bilingualism, with early acquisition of the optionality, similar to the modern Frisian situation and the arguments made by Bremmer (1997) on mixed Frisian.

3. Variation in Early-Modern Frisian texts is due to learned borrowing, with late acquisition of the optionality, along the lines of Blom’s (2008) account for the mixed Frisian texts.

To find out which of these sources can account for the 1-2 orders that are used in Early-Modern Frisian texts, we are particularly interested in the contexts in which the ‘Dutch’ 1-2 cluster order is used in the Early-Modern Frisian corpus. In our study, we test whether the Early-Modern Frisian 1-2 orders occur in the same contexts as modern Dutch 1-2 orders to see whether the use of this order is a contact effect, and what type of contact is responsible for them.

It has been argued that verb cluster order variation in Dutch has the function of facilitating sentence processing: the verb cluster order that is ‘easier’ or more economical in a particular context is used (De Sutter, 2009; Bloem et al., 2017b). By studying whether the variation in Early-Modern Frisian texts correlates with the same factors as the variation in modern Dutch, we can infer whether Early-Modern Frisian verb cluster order variation has the same functions as modern Dutch verb cluster order variation. This will allow us to find evidence for or against the hypothesis of language contact effects on verb cluster orders in older Frisian texts.

If Early-Modern Frisian 1-2 order clusters occur in similar contexts as modern Dutch clusters in the 1-2 order this would indicate that this order has the same function in both varieties, and is part of the grammar of the writer of the Early-Modern Frisian text. This can mean two things: Firstly, it could be the case that the order was not borrowed, but instead already existed in the
language from Proto-West-Germanic on, subject to the same cognitive constraints as its modern Dutch counterpart. If this is the case, and we find no other evidence of borrowing verb cluster constructions, an older stage of the language would be the source of the variation: this would support hypothesis 1. Secondly, it could be the case that ‘contact through bilingualism’ is the source of the variation: hypothesis 2.

If the usage contexts are not similar between Early-Modern Frisian and modern Dutch, this means it is likely that the 1-2 order has been borrowed in some way, but with a different function than the function it has in modern Dutch. In this case, learned borrowing would be the source of the variation: this would support hypothesis 3.

The reason why we compare Early-Modern Frisian to Modern Dutch here, and not Early-Modern Dutch or Modern Frisian, is that most existing research on verb cluster order variation was conducted on modern Dutch. Results of corpus studies on Modern Dutch verb clusters will give us the most detailed picture of the contexts in which 1-2 orders are preferred. As many of these preferences are related to ease of processing and cognitive constraints, which are unlikely to differ per language, we will work under the assumption that the same processing mechanisms that are involved in writing modern Dutch text also apply to Early-Modern Frisian writers. Furthermore, previous work has shown that word order preferences differed depending on the type of auxiliary verb both in Middle Dutch (Coussé, 2008) and in Early-Modern Dutch (Versloot et al., 2012), much like in modern Dutch. We summarize the findings for modern Dutch in the next section.

8.3 Verb clusters in Dutch

8.3.1 Factors of variation

Dutch verb cluster order variation has been studied more thoroughly than the variation in Frisian. As noted in the introduction, for two-verb clusters, both the 1-2 and 2-1 orders may be used in modern Dutch. Diachronically, there have been changes in the prevalence of the word orders, much as in Frisian. These changes have been studied and charted by Coussé (2008) and Coupé (2015). As for Frisian, both orders can be found in the oldest Dutch texts (De Schutter, 2012). Then, as in Frisian, the 2-1 order slowly became the dominant order in Dutch texts from around the years 1400 to 1600. But subsequently, unlike in Frisian, a development in the opposite direction began, and the use of the 1-2 order in Dutch texts started to increase (Coupé, 2015). This increase continued during the following centuries (Coussé, 2008) and appears to be ongoing — Olthof et al. (2017) showed that younger speakers in the Corpus Gesproken Nederlands (CGN, corpus of spoken Dutch) have a stronger preference for the 1-2 order than older speakers.

Synchronously, various corpus studies have addressed the choice that speak-
ers of modern Dutch have between the two verb orders. With most auxiliary verbs, the orders are in free variation, and speakers would not make a conscious decision to use one over the other. The difference is generally assumed not to correspond to a meaning difference. However, corpus studies show that the orders are not used randomly (De Sutter, 2005; Bloem et al., 2014). The use of particular orders has been shown to correlate with a range of factors, and various generalizations have been proposed to account for the relation between these factors and verb cluster word order:

**Minimizing processing complexity** To account for some observations from his corpus study, such as the correlation between main verb frequency and word order, De Sutter (2005) proposed the idea that speakers will use the verb cluster word order that they stylistically prefer when they have the spare cognitive capacity to do so, and will use the default order when they do not have spare cognitive capacity. Bloem et al. (2014) take this idea and suggest that it can be tested by studying whether factors indicating ease of processing correlate with one particular verb cluster word order, which would be the ‘default’ order. Subsequently, Bloem et al. (2017b) found various patterns that support this assumption, arguing that one of the word orders, the 1-2 order, might be easier to process. When both orders are grammatical, the 1-2 order appears to be preferred by speakers in contexts that are complex to process for the purpose of minimizing the processing complexity of the utterance.

**Structural priming** In a psycholinguistic experiment, Hartsuiker & Westenberg (2000) showed that verbal cluster orders can undergo structural priming — a 1-2 order was more likely to be produced when participants had been presented with this order in a priming sentence, even though different words were used. The effect was found for both the spoken and the written modality. Corpus studies also show that 1-2 orders are more likely to occur after other 1-2 orders (De Sutter, 2005; Bloem et al., 2017b).

**Semantics of the main verb** Collostructional analysis studies have found that certain main verbs are used statistically significantly more often in either the 2-1 or 1-2 order in written text (De Sutter, 2005; Bloem, 2016b) and that certain semantic classes of verbs also appear to have such usage preferences (Bloem, 2016b). For example, verbs with negative semantics are more likely to be used in the 2-1 order (Bloem, 2016b).

**Auxiliary verb class** De Schutter (1964) noted that the different auxiliary verbs have different frequencies of occurrence in the 1-2 and 2-1 order. This finding also holds when grouping instances of auxiliary verbs by their semantic class, such as modal verbs, auxiliaries of time (De Sutter, 2005).
Information weight  Some studies have discussed evenly distributing information weight throughout a sentence (De Sutter et al., 2007) and the verb cluster (Bloem, 2016c) as a factor, though this could also be considered a processing effect.

Rhythmic factor  Evidence has also been found that Dutch speakers move verbs in the verb cluster around to avoid stress clashes and to adhere to the rhythm of the Dutch language (De Schutter, 1996, 2012).

Along with these potential sources of intraspeaker variation, there is also interspeaker variation. Speakers from different regions have different word order preferences (Pauwels, 1953) and younger speakers have a stronger preference for the 1-2 order than older speakers (Olthof et al., 2017). Some other contextual factors play a role as well — the mode of communication (spoken or written), the interactivity of the discourse, the immediacy of the discourse, and stylistic preferences of the speaker or language community (De Sutter, 2005).

8.3.2 Processing verb cluster orders

It has been argued that, in more complex contexts, the ‘easier’ verb cluster order is used (De Sutter, 2009; Bloem et al., 2017b). Various arguments have been made regarding which order is ‘easier’ or ‘default’ in Dutch: De Sutter (2009) discusses the 2-1 order as easier to process, while Bloem et al. (2017b) conclude that it is likely to be the 1-2 order on the basis of a large corpus study. Other arguments for an 1-2 default are that it seems to be acquired earlier by children (Meyer et al., 2015b) and has a more uniform information density (Bloem, 2016c).

The arguments used for these claims could apply to Frisian as well, although they have not been tested. The 1-2 order is claimed to be acquired earlier than the 2-1 order by Dutch children because the 1-2 order seemingly violates verb-finality and is therefore a new construction to be learned. The most verb-like element, the finite verb, is not in the final position in a 1-2 order cluster (Meyer et al., 2015b). Verb-finality is learned quite early by Dutch children, and presumably by Frisian children as well, as Frisian is also a verb-final language. Therefore, this same argument should apply to Frisian. Furthermore, cognitive constraints on processing are unlikely to differ for speakers of different, closely related languages.

If both verb cluster orders are available in the mental grammars of the authors who wrote texts in Early-Modern Frisian, and this language change occurred ‘from below’, we would expect to find similar factors affecting Early-Modern Frisian verb cluster variation: the 1-2 order is used in more complex contexts. If the 1-2 order was a learned borrowing, stylistically marked, it would not be used in this way. Bloem et al. (2017b) used the following factors as indicators of processing complexity in their study of Dutch verb clusters:
### Syntactic priming

When a word or construction is perceived or produced it is activated, and some traces of this activation remain for a while, making it easier to re-activate them. In Dutch, verb cluster orders can syntactically prime each other Hartsuiker & Westenberg (2000) and priming or persistence effects can be found in written text too (De Sutter, 2005).

### Morphological structure of the main verb

Dutch has separable complex verbs, which contain a particle that may appear in a different place in the sentence in some contexts. Separable verbs have been argued to be more complex to process because, unlike regular verbs, syntactic processes can operate on them in addition to morphological processes (Bloem et al., 2017b). Separable complex verbs exist in Frisian too. If the 1-2 order is an available option in the writer’s grammar without being stylistically marked, we would expect it to be used in the more complex condition (i.e. when the verb is separable as opposed to inseparable).

### Length of the middle field

Verb clusters preceded by longer middle fields can be considered more difficult to process, because a longer middle field is more likely to lead to long dependency lengths and more dependencies between various parts of the middle field and the verb cluster at the end. According to resource-limitation models of language processing, this would lead to greater integration costs at the verb cluster.

### Structural depth

Verb clusters that are further down in the syntactic tree of a sentence may be considered more difficult to process, assuming that the apparent hierarchical syntactic structure of sentences is cognitively real.

### Information value of the last preverbal word

This factor concerns the word class of the word that is used directly before the verbal cluster. Psycholinguistic evidence suggests that closed-class words are processed differently than open-class words, however, it is not clear which word class is actually easier to process.

### Definiteness of the last preverbal word

When a definite NP occurs before the verb cluster, it has to be linked to a previously given discourse element, which may be more complex than processing an indefinite NP.

### Frequency of the main verb

Psycholinguistic theories generally assume that more frequent words are more easily activated, making them easier to process.

### Extraposition

In Dutch, it is possible to extrapose prepositional objects, positioning them after the verb cluster. This is called PP-over-V and is a
common construction (Willems & De Sutter, 2015) and it has been hypothesized to make long middle fields easier to process (van Haeringen, 1956). PP-over-V is grammatical in Frisian too.

**Multi-word units** A multi-word unit (MWU) is a lexical unit, consisting of multiple words that together carry a different meaning than that of its constituent parts. They are processed faster and remembered better than other sequences of words.

These factors could also be studied in Early-Modern Frisian. In this paper, we will limit ourselves to factors that can be extracted from a part-of-speech tagged historical corpus automatically. In section 8.4.4, we discuss which of these factors can be studied in the available Early-Modern Frisian corpus data without manually combing through the corpus.

### 8.4 Data and methodology

To study Early-Modern Frisian verb cluster orders and their usage contexts, we have extracted verb clusters from a corpus of Early-Modern Frisian texts, the Integrated Language Database\(^9\) (Taaldatabank, TDB) of the Fryske Akademy. This corpus contains all of the attested Frisian texts from the years 1550-1800. The texts have been tokenized, lemmatized and part-of-speech tagged manually. This section describes the corpus annotation, the selected texts and the process of verb cluster extraction and data pre-processing.

#### 8.4.1 Texts

From the corpus, we have selected several of the longer texts and text collections for analysis. We decided to focus on authors of whom the corpus contains more material, because this is more likely to give us large samples of verb clusters for a particular author, avoiding interspeaker variation. Larger samples are advantageous because it is easier to make inferences about the data when there is more of it.

We also make a distinction between rhyming texts and prose texts, selecting some of each, including a set of rhyming texts and prose texts by the same author. This factor of text type is relevant because in rhyming texts, metre and rhyme requirements restrict the constructions that an author can use, forcing authors to use other stylistic options that they might not use in prose, such as different word orders. A previously studied example of this can be found in Old English and Old High German poetry, where verb-first (V1) declarative clauses are more common in poetry than in prose written in these languages. This is a result of emphatic verb-fronting, a stylistic technique used by Old English and

---

8.4. Data and methodology

Old High German poets (Kaminska, 2007). Interestingly, this word order is also common in prose translated from Latin to these languages (Kaminska, 2007, p. 67), indicating that this use of the V1 word order may have been borrowed from Latin, becoming a stylistic option for Old English and Old High German, at least in poetry. It might be the case that such different uses of word orders occur in Frisian poetry as well. This section lists the texts and text collections that we have selected from the TDB, along with their year of writing or publication. We also include year as a factor, as word order preferences may have shifted over time. We have not used author as a factor in our main analysis, as it correlates strongly with the factor year and partially with the factor of text type.

Gysbert Japicx’s collected work (1639-1666) We included all Frisian texts by Gysbert Japicx that are in the Integrated Language Database (excluding shorter editions when longer editions of the same text were available). The texts were split by text type: the collected prose texts of Gysbert Japicx and the collected rhyming texts of Gysbert Japicx. Gysbert Japicx (1603-1666) was a highly influential poet and writer from Bolsward, a city in the west of Frisia. He wrote Dutch and Frisian poetry, as well as Frisian prose, some of which was translated from French. He also translated Latin psalms to Frisian through Dutch, and may have written poetry in Latin. Frisian-language publications were exceptional at this time, and Gysbert Japicx was responsible for the majority of published Frisian writings in the 17th century (Breuker, 1989). These text collections were chosen because of their importance and large size, as well as the fact that rhyme and prose texts by the same author are available.

Reyner Bogerman’s Frisian Sayings (1542 and 1551) These rhyming texts were selected because they are relatively old — it’s Early-Modern Frisian, but it’s still fairly close to the Frisian of the 15th century.

Van Hichtum’s wedding poetry (1609) These rhyming texts are said to have inspired Japicx to write in Frisian.

Burmania Sayings (1614) This text was selected because it is the oldest longer Early-Modern Frisian prose text in the corpus.

Waatze Gribberts Bruyloft (1701) This text was selected because it is a longer prose text, written in a register that appears to be relatively close to spoken language.

\[10\] Specifically, the following text numbers in the TDB were included: 2192 of text group 2, and 3195 of text group 3, all of text group 5 except 5001, 5149 and 5002, and texts 7154 to 7189 and 7191 of text group 7.

\[11\] Texts 1542a & 1551a in the TDB.

\[12\] Texts 1639b and 1639c in the TDB.

\[13\] Text 1614a in the TDB.

\[14\] Text 1701c in the TDB.
Aagtje IJsbrants (1779) by Eelke Meinerts  This text\textsuperscript{15} was also selected because it is a longer prose text, written in a register that appears to be relatively close to spoken language. It is of a slightly later date than Brugloft.

Eelke Meinerts’ rhyming texts (1777-1783)  These texts were selected to also include some rhyming texts from the author of Aagtje IJsbrants and from the later end of the year range.

8.4.2 Annotation

The texts in the TDB have been manually annotated by an expert on a word-by-word basis (Visser, 1996), mainly for lexical properties. This section summarizes what has been annotated and what has not been annotated, and how this annotation can provide information about verb clusters. We aim to use this information to identify verb clusters in the text automatically.

The annotation does not include information on the syntactic structure of sentences, such as constituent structure or dependency structure. Since verb clusters are syntactic phenomenon, and usually defined in terms of hierarchical structures, this means that the available annotation does not tell us which verbs form a verb cluster. For example, not every sequence of verbs in the linear order of the sentence is a verb cluster — some of the verbs may belong to a different verb phrase or clause. The annotation also does not include information on clause boundaries or sentence boundaries, further adding to the difficulty of identifying which verbs form a syntactic cluster. The annotations do not include information on the type of clause either. Verb clusters in subordinate clauses work differently from verb clusters in main clauses, because the syntactically highest verb is generally positioned in the second position in the main clause (V2-effect). Word order preferences in the two clause types are also known to be different in other Germanic languages, except for English and Icelandic (Askedal, 2006). For these reasons, the two clause types need to be distinguished.

There are various annotations at the word level however, including lemmatization. An uninflected, modern Frisian form of each word is annotated for each token. Numbers are used to distinguish lemmas that are spelled the same. This makes it possible to aggregate over verb forms or different spellings, i.e. to know that mot and mötte are both forms of moatte ‘must’. Information on the form of verbs is also included. This is useful because knowing whether it is an infinitive, participial or finite verb can help in automatically determining the verb’s syntactic position in the verb cluster. Furthermore, the annotation indicates whether a verb is a main verb, auxiliary verb or copula, though this property does not seem to take syntactic position into account — if ‘to have’ is the main verb of a clause, it is still annotated as being an auxiliary verb. The annotation of verb types is fairly extensive, also including information

\footnotetext{15}{Text 1779m in the TDB.}
on transitivity (for main verbs) and tense (for auxiliary verbs). Five types of auxiliary verbs are distinguished in the annotation: modal, passive, aspectual, present and future. Copular verbs are included as a different verb type, besides auxiliaries. This classification is also relevant to the study of verb clusters, as the different auxiliary verb types exhibit different word order preferences in other Germanic languages, i.e. Dutch (De Sutter, 2005). Furthermore, prefixes and suffixes are separated from lemmas by a dash. This includes verb particles, making it possible to distinguish separable verbs, which exhibit different word order preferences in Dutch.

8.4.3 Automatic extraction

From the chosen texts, we automatically extracted (potential) verb clusters and their properties. We wrote a Python script (van Rossum & Drake, 2012) that detects verb clusters in a rule-based manner, using the information available in the annotation. A verb cluster is defined roughly as follows:

- A sequence of verbs within the same ‘clause’ (sequence of text not interrupted by punctuation)
- The sequence may be interrupted by a verb particle or ‘te’ (to), which mark a gerundium (to-infinitive)
- The sequence must have at least one auxiliary verb and one main verb
- Auxiliary verbs are verbs that are annotated as modal verbs, auxiliaries of time, and passive verbs. They can also be aspectual verbs and copula verbs, but only when that verb is not an infinitive, to-infinitive, past participle or present participle (because then it would be a main verb).
- Main verbs are verbs that are annotated as infinitives, to-infinitives, past participles and present participles.

The word order of the verb cluster is then determined on the basis of the relative positions of its constituent verbs in the linear order of the sentence. It should be noted that this procedure is not 100% reliable, especially in clusters with infinitival auxiliary verbs, where auxiliary verbs and main verbs may have the same form.

8.4.4 Factors potentially affecting order variation

In section 8.3.2, we saw that some factors correlate with verb cluster word order variation in modern Dutch, and can account for this variation in terms of processing effects. While we would like to test these factors for Early-Modern Frisian too, it is not possible to extract all of them from the TDB corpus. The TDB annotation is more limited than the annotation of the Dutch corpus that was used by Bloem et al. (2017b): the Dutch Lassy Large corpus also contains
syntactic annotation. In this section, we will discuss whether each of the relevant factors can be extracted from the Early-Modern Frisian data in the TDB, and if so, how we operationalize them.

**Syntactic priming** Priming may have taken place when the order of two subsequent verb clusters is the same. This factor does not involve any syntactic dependency between the two verb clusters, so we can operationalize it in the same way as (Bloem et al., 2017b). We record whether the previous verb cluster was in the 1-2 or 2-1 order, if there was one.

**Morphological structure of the main verb** Just like Dutch, Frisian has separable complex verbs. In the TDB, such morphological properties are not annotated, but these verbs often appear to be written with a space or dash between the separable particle and the rest of the verb. Therefore, we take any verb containing a space or dash to be a separable complex verb. When a separable verb is actually separated, this is annotated both on the particle and on the main verb. These cases are also counted as separable.

**Length of the middle field** As the TDB lacks syntactic annotation, it is difficult to determine the length of the middle field of a clause before a cluster, or even of the clause. We measure the number of consecutive words without punctuation, or a line end (in the case of poetry) as an approximation of clause length, but this is likely to be inaccurate. For example, if the subordinate clause has an embedded clause, the length will only be counted from the end of the embedded clause. In section 8.5.1, we evaluate the accuracy of these estimates.

**Structural depth** This measure relates to the position of a verb cluster in the syntactic tree of a sentence. This information cannot be obtained from the TDB, and we have not included this factor in our study.

**Information value of the last preverbal word** This factor concerns the word class of the word that is used directly before the verbal cluster. As this depends only on linear order and word class annotation, it can be extracted from the TDB. We operationalize this factor in roughly the same way as De Sutter (2007) and Bloem et al. (2017b), in terms of three classes: highly informational (nouns, verbs, numerals, proper names), intermediate informational (adjectives and adverbs) and low informational (pronouns, conjunctions, prepositions, articles and interjections).

**Definiteness of the last preverbal word** This factor cannot be operationalized accurately for the TDB data. While some morphological properties of nouns are included in the annotation, such as gender, definiteness is not annotated. And because there is no syntactic annotation such as dependency links, we cannot easily retrieve whether a preverbal noun has a determiner that
indicates definiteness. Therefore, we have not included this factor in the present study.

**Frequency of the main verb** It is difficult to get an accurate estimation of general word frequencies in Early-Modern Frisian, especially since they may have varied over time. We estimated this factor by counting the frequencies of lemmas (not word forms) in all texts selected from the TDB. This estimate is likely biased due to the limited number of texts and the limited topics that these texts cover, but it is the best we can do.

**Extraposition** Extraposition of the prepositional object, also called PP-over-V, is another source of variation that can occur around verbal clusters. It cannot reliably be detected using the annotation available in the TDB, as it depends on syntactic links between the prepositional object and the clause before the verb cluster. Therefore, we have not included this factor in the present study.

**Multi-word units** No annotation on multi-word units, idiomatic expressions or similar information is available in the TDB, therefore we have not included this factor in the present study.

Table 8.1 summarizes the list of factors we could operationalize for the TDB, with frequency information for each possible value. Besides these processing-related factors, there are a few other relevant factors that are known to affect verb cluster order variation in Dutch that may be relevant here. As mentioned before, one is the type of auxiliary verb: the 1-2 order is more common with modal auxiliaries than with hebben ‘to have’ in Dutch, for example. Another is the use of to-infinitives, or te-infinitives in Dutch, where order preferences appear to be different. We have included these two factors in the study.

Another factor is clause type: word order preferences in main clause clusters are different than in subordinate clause clusters. In the present study, we only focus on subordinate clause clusters, as in Frisian main clauses the finite verb is placed in verb-second position, and due to the lack of syntactic annotation in the TDB we cannot determine which verbs go together. However, the annotation also does not distinguish between main clauses or subordinate clauses, so our data set is likely to include some main clause clusters that look like subordinate clause clusters (for example, if there is nothing between the finite verb in V2-position and the final verb). In section 8.5.1, we have evaluated the extent of this problem.

8.5 Results

In this section we present our results regarding verb cluster word order in Early-Modern Frisian. First, we present an evaluation of the reliability of our
Verb cluster word order in Early-Modern Frisian

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TEXT TYPE</strong></td>
<td>Rhyming text</td>
<td>377</td>
</tr>
<tr>
<td></td>
<td>Prose text</td>
<td>1,169</td>
</tr>
<tr>
<td><strong>TEXT YEAR</strong></td>
<td>Approximate year in which text was published</td>
<td>—</td>
</tr>
<tr>
<td><strong>TYPE OF AUXILIARY</strong></td>
<td>Modal</td>
<td>692</td>
</tr>
<tr>
<td></td>
<td>Copula</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Aspectual</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Passive</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Past tense</td>
<td>466</td>
</tr>
<tr>
<td></td>
<td>Future tense</td>
<td>94</td>
</tr>
<tr>
<td><strong>TO-INFINITIVE</strong></td>
<td>To-infinitive</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>Not a to-infinitive</td>
<td>1,467</td>
</tr>
<tr>
<td><strong>PRIMING</strong></td>
<td>1-2 order is primed</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td>2-1 order is primed</td>
<td>1,171</td>
</tr>
<tr>
<td><strong>SEPARABLE VERB</strong></td>
<td>Separable main verb</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Inseparable main verb</td>
<td>1,411</td>
</tr>
<tr>
<td><strong>INFORMATION VALUE</strong></td>
<td>High (nouns and verbs)</td>
<td>597</td>
</tr>
<tr>
<td></td>
<td>Medium (adverbs and adjectives)</td>
<td>536</td>
</tr>
<tr>
<td></td>
<td>Low (function words)</td>
<td>397</td>
</tr>
<tr>
<td><strong>FREQUENCY OF MAIN VERB</strong></td>
<td>Log frequency of the main verb in the overall corpus</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 8.1: Summary of factors and the frequency of occurrence of their levels
automatically extracted data by manually checking a part of it. Then, we present the results of this automatically extracted data from the Early-Modern Frisian corpus.

8.5.1 Evaluation

In the previous section, we have discussed how the data was automatically extracted from the corpus, and that the annotation of the TDB corpus was not always sufficient to extract verb clusters, their word order, and their associated information unambiguously. Specifically, the script can make the following errors:

- Categorize a part of a 3-verb cluster as a 2-verb cluster
- Categorize (a part of) a main clause verb cluster or group as a 2-verb subordinate clause cluster
- Categorize a sequence of verbs that spans clause boundaries as a verb cluster
- Incorrectly determine the order of a cluster, if verbs that can be both a main verb and an auxiliary verb are involved (i.e. ‘to have’)  
- Incorrectly estimate the length of the clause due to punctuation inside the clause, or lack of punctuation

Furthermore, there may be errors in the annotation. The first four error types affect the verb cluster order that the script reports, the last error types affects the clause length that the script reports. We have evaluated to what extent these errors influence the factors that we use in our study by having a Frisian linguist manually check the annotation of a random sample of 50 1-2 order clusters from Gysbert Japicx’s collected prose work and 50 2-1 order clusters from the same subcorpus. We chose to use prose text for this evaluation because sentence structure is more complex in these texts and the script is more likely to make mistakes. A pilot evaluation showed fewer errors in poetry. Furthermore, prose texts are expected to contain fewer 1-2 orders, so it is more important that the 1-2 orders that are identified, are correct. The sample is balanced for word order — it is not completely random, otherwise there would be more 2-1 orders in the sample, as the texts contain more 2-1 orders. We evaluate only for precision, not for recall (whether the script finds all the relevant clusters). That would involve reading a sample of the corpus and checking it for any verb clusters the script might have missed.

Word order errors  Of the 50 automatically extracted candidate 1-2 clusters, 34 were found to be actual two-verb clusters from subordinate clauses: a precision of 68%. Of the 50 2-1 clusters, all 50 met this requirement (100%
precision). Most of the erroneous candidate clusters were cases of a finite auxiliary verb in V2 position in a main clause, immediately followed by the main verb in final position, with no intervening objects. This looks exactly like a 1-2 order cluster consisting of a finite auxiliary verb and a main verb at the end of a subordinate clause. Main clause clusters cannot look like 2-1 order clusters, which explains the 100% precision for the 2-1 order. This evaluation shows that a statistical model of this data is likely to overestimate the probability of 1-2 orders somewhat, at least for prose text.

Clause length errors The Waatze Gribberts subcorpus contained a few verb clusters for which the length of the clause was estimated to be 484 words, due to a lack of punctuation in a section of the text. For these outliers, the length value was manually removed. All other estimates of clause lengths are 40 words or less.

For the 100 manually checked clusters described above, the correct clause lengths were manually determined. We used linear modeling to test whether there is a significant difference between the automatically determined clause lengths and the manually determined clause lengths.

The model’s estimate of the mean clause length (weighted equally over manual and automatic measurements) is 8.9 words (95% confidence interval 8.3 .. 9.5 words). The estimated mean of the automatically determined lengths does not differ significantly from the estimated mean of the manually corrected lengths ($t[168] = 1.1, p = 0.27$). We conclude that there is no evidence for an overall difference between automatic and manually corrected lengths (estimated difference = 0.7 words; 95% confidence interval -0.6 .. 1.9 words). The estimated mean of 1-2 order cluster clause lengths also does not differ significantly from the estimated mean of 2-1 order cluster clause lengths ($t[168] = -0.1, p = 0.91$). We conclude that there is no evidence for an overall difference between 1-2 and 2-1 order clause lengths (estimated difference = 0.06 words; 95% confidence interval -1.3 .. 1.2 words). However, both conclusions are qualified by a significant interaction between correction and word order ($p = 0.003$): the estimated correction-related length difference (corrected minus automatic) is 3.8 words greater for 1-2 orders than for 2-1 orders. In other words, there is evidence for a cross-over interaction where the effect of manual correction is opposite for 1-2 orders compared to 2-1 orders. The script significantly overestimates the length of clauses with 2-1 clusters and underestimates the length of clauses with 1-2 clusters. We must conclude that the automatically determined clause lengths are not reliable predictors of cluster order, even though they do not differ significantly from the manually corrected clause lengths overall. Therefore, we have excluded this factor from the model.

Annotation errors The expert annotator found a few errors that are likely caused by annotation errors in the TDB. There were three cases where the to-infinitive status of a main verb was not detected. Furthermore, the annotator
disagreed with the auxiliary type labels of two verbs. *Litte* ‘let’ is annotated as an aspectual verb, but the annotator considered it to be more like a modal verb, and *behearre* ‘should’ is annotated as a modal verb, but the annotator considered it to be more like an aspectual verb. Two of the 100 sample clusters contained *litte* and four contained *behearre*. We decided to follow the annotation on this point and not make modifications to the auxiliary verb classes in the TDB.

### 8.5.2 Multifactorial model of order variation

To be able to discuss the effect that different factors have on verbal cluster order variation in early-modern Frisian, we have created a multifactorial model using the factors described in section 8.4.4. We model verbal cluster order as a binary variable, in which the order can be 1-2 (ascending) or 2-1 (descending). This is the dependent variable. This variable is modeled in terms of the factors of the multivariate model, the independent variables, using logistic regression.

Verbal cluster order variation is one of many language variation phenomena where it has been established that multiple factors contribute significantly to the observed variation (Bloem et al., 2017b). Such phenomena are best studied using multifactorial models, rather than testing each factor one by one. Starting with Gries (2001), such multifactorial models have been successfully used in the study of language variation. Multifactorial models allow the researcher to examine how much a particular factor contributes to the choice of a construction or word order, while controlling for the other factors. Studying several factors in isolation may cause the same variation to be attributed to multiple factors. Testing each factor while applying the necessary corrections for running multiple tests also increases the chance of type II errors, a failure to reject a false null hypothesis, and the statistical power of a multifactorial model is greater.

We will start with some general observations regarding the data. The 2-1 order is more frequent overall: the data set contains 1546 clusters in total, of which 1175 clusters are in the 2-1 order and 371 clusters in the 1-2 order. Since we expect more word order variation in the rhyming text type, it is interesting to look at word orders per text type. Figures 8.1a and 8.2a shows the distribution of verb cluster orders by text type, revealing that the 1-2 order is relatively and absolutely more frequent in the rhyming texts, and that there is more word order variation there.

It is also interesting to look at word orders per type of auxiliary verb, as different auxiliary verbs exhibit clear word order preferences in Dutch. Figure 8.1b shows the absolute counts and figure 8.2b shows the relative percentages.

Table 8.2 lists all of the explanatory variables used in the model, along with their effect size. The model uses these variables to predict the binary dependent variable of verbal cluster word order. This table allows us to explore the effect of each factor on verbal cluster order. Most of the explanatory variables are categorical, and for these variables we coded orthogonal contrasts to be able to compare the estimated effect of categories to the estimated mean effect of
Verb cluster word order in Early-Modern Frisian

Figure 8.1: Absolute counts of verb clusters per order

(a) Verb cluster word order by text type
(b) Verb cluster word order by type of auxiliary verb
8.5. Results

(a) Verb cluster word order percentages by text type

(b) Verb cluster word order percentages by type of auxiliary verb

Figure 8.2: Relative percentages of verb clusters per order
multiple other categories. For example, for the variable TYPE OF AUXILIARY, the effect of the category of modal auxiliary verbs is compared to the mean effect of all other categories (shown in Table 8.2 as +Modal compared to -FutureAspectualCopulaPastPassive). The effect size listed for the +Modal auxiliary type then indicates the odds of a 1-2 order when there is a modal auxiliary verb, rather than another, all other conditions being the same.

One might wonder why we do not compare every type of auxiliary verb to the average of all other types. This is not possible in a single statistical test, as including more contrasts than the number of levels, minus one, would introduce multicollinearity — certain auxiliary types would be in the model more than once, and these ‘copies’ would then have a correlation of 1 with each other. An assumption of regression modeling is that the factors do not correlate with each other, so this cannot be done. Orthogonal contrasts are the next best option — another option would be to compare every type of auxiliary to one particular ‘baseline’ type, but there is no one type of auxiliary that is ‘default’ or ‘basic’ in verb cluster constructions.

We ordered the orthogonal contrasts such that the auxiliary verb types which deviate the most from the average in modern Dutch are tested first. Therefore, we start by comparing modal clusters to the average, as modal auxiliaries have a very strong 1-2 order preference in Dutch (the 2-1 order is sometimes even said to be ungrammatical). The same goes for future auxiliaries, and the less frequent aspectual auxiliaries. Next come the copular verbs, which deviate from the average in the other direction in Dutch, having a 2-1 order preference. Passive and past auxiliaries come last, as they have more or less average word order preferences in Dutch (Bloem et al., 2017b). In short, the order of comparison is: Modal, Future, Aspectual, Copula, Past, Passive.

The effect size of each variable is given as an odds ratio. An odds ratio further from 1 in either direction indicates a stronger effect. An odds ratio of 1 would mean that there is no difference in the odds of a 1-2 order when a modal is used, compared to another auxiliary verb. In line with previous work, we are reporting values for the 1-2 order. As a result, odds ratios > 1 indicate an effect of association with the 1-2 word order, odds ratios < 1 indicate the 2-1 order. In our results the odds ratio for the modal contrast is 1.19, showing that in our data, clusters with modal auxiliaries are estimated to be 1.19 times more likely to be in the 1-2 order than clusters with other auxiliary verbs. This difference is not significant however, so our model provides no evidence for an effect of modal auxiliaries ($t[1545] = 0.17$; 95% confidence interval = 0.76 .. 1.84 times; p from one = 0.44). In Table 8.2, asterisks indicate statistical significance: One asterisk (*) indicates statistical significance at the $p < 0.05$ level, two asterisks indicate significance at $p < 0.01$, and three asterisks indicate significance at $p < 0.001$.

The variables YEAR and FREQUENCY OF MAIN VERB are not categorical, but continuous. In these cases, the odds ratios show the increase or decrease in the odds of a 1-2 order per unit (i.e. per additional year).

It should be noted that the model is likely to over-estimate the number of
1-2 orders because, as we saw in section 8.5.1, the automatic extraction gives us false positives for 1-2 orders but not for 2-1 orders. Since the script achieved a precision of 68% on 1-2 orders, the true odds are likely to be around 0.68 times lower than the odds estimated by the model we present here, although this does not control for correlations between specific factors and the error rate.

The results in Table 8.2 show that the model estimates 1-2 orders to be 18.69 times more likely in rhyming text than in prose text (95% confidence interval 12.8 .. 27.8 times). 1-2 orders are significantly more likely in rhyming text than in prose text ($t[1545]=2.93; p = 5.51 \times 10^{-49}$). We conclude that Early-Modern Frisian rhyming texts are more likely to contain verb clusters in the 1-2 order than Early-Modern Frisian prose texts.

Regarding the type of auxiliary verb, we can observe several things. As mentioned, clusters with modal auxiliaries are estimated to be 1.19 times more likely to be in the 1-2 order than clusters with other auxiliary verbs, a difference that is not significant ($t[1545]=0.17; 95\% \text{ confidence interval} = 0.76 .. 1.84 \text{ times}; p \text{ from one} = 0.44$). Therefore, we have no evidence of an effect of the use of modal auxiliaries on verb cluster order in Early-Modern Frisian. In modern Dutch, verb clusters with modal auxiliaries have a strong 1-2 order preference (Bloem et al., 2017b), and diachronically, Dutch modal verb clusters, and other clusters taking infinitival main verbs, shifted towards a 1-2 order preference earlier than verb clusters with other auxiliaries (Coussé, 2008, p. 63).
We do observe a significant difference for the contrast of aspectual auxiliary verbs versus copular, passive and past ones: the model estimates aspectuals to be 7.2 times more likely to occur in the 1-2 order \((t[1545] = 1.97; 95\% \text{ confidence interval} = 2.7 \ldots 18.8 \times 10^{-45})\), even though this category only occurs 43 times in the data. We conclude that aspectual verbs are more likely to occur in the 1-2 order than copular, passive and past auxiliary verbs. In Dutch, aspectual verbs behave like modal verbs in that they are almost always used in the 1-2 order, though 2-1 orders can be judged acceptable. They are also more widely used in Dutch than in Frisian.

Another significant difference here is between copular auxiliaries, and past and passive ones. The model estimates copular auxiliaries to be 7.9 times more likely to occur in the 1-2 order \((t[1545] = 2.06; 95\% \text{ confidence interval} = 2 \ldots 36.5 \times 10^{-45})\), although this category is even more infrequent, occurring only 12 times in the data. Nevertheless, we can conclude that copular verbs are more likely to occur in the 1-2 order than passive and past auxiliary verbs. Conversely, in Dutch, copular auxiliaries have the strongest association with the 2-1 order out of all the auxiliary verb types.

The final significant difference among the auxiliary verb types is between passive auxiliaries and auxiliaries of past tense. Passive auxiliaries are estimated to be 2.5 times more likely to occur in the 1-2 order \((t[1545] = 1.62; 95\% \text{ confidence interval} = 2.5 \ldots 10.2 \times 10^{-45})\). We conclude that passive auxiliaries are more likely to occur in the 1-2 order than past auxiliary verbs. In Dutch, past tense auxiliaries (with participial main verbs) have a somewhat greater 1-2 order preference than passive auxiliaries (De Sutter, 2009; Bloem et al., 2017b).

Even though the data set does not include many verb clusters with a to-infinitive (79 out of 1546 clusters), we see that the model finds a significant effect. Clusters with to-infinitives are estimated to be 11.5 times more likely to be in the 1-2 order than other clusters \((t[1545] = 2.12; 95\% \text{ confidence interval} = 4.3 \ldots 16.0 \times 10^{-45})\). We conclude that to-infinitival clusters were more likely to be used in the 1-2 order in Early-Modern Frisian texts.

Lastly, we can observe a statistically significant effect of the information value of the preverbal word: when this word is medium-informative (an adverb or adjective) rather than low-informative (a function word), the cluster is estimated to be 0.24 times more likely to be in the 1-2 order, or in other words, 4.16 times more likely to be in the 2-1 order \((t[1545] = -1.44; 95\% \text{ confidence interval} = 0.16 \ldots 0.35 \times 10^{-12})\). We conclude that 2-1 order clusters were more likely following medium-informative words and 1-2 clusters were more likely following function words in Early-Modern Frisian texts. In modern Dutch, both function words and medium-informative words are more associated with the 1-2 order, while high-informative words before clusters are more associated with the 2-1 order (Bloem et al., 2017b).

We find no evidence for an effect of year. The direction of the model’s estimate is a slight decrease in 1-2 orders over time, but this effect is not statistically significant. For the remaining factors listed in Table 8.2, we also
found no evidence of an effect on verb cluster order in Early-Modern Frisian. In studies of modern written Dutch corpora, effects of all of these factors were found (De Sutter, 2005; Bloem et al., 2017b), though the fact that the Dutch corpora that have been studied are much bigger plays a major role in this: with a larger corpus, effects of smaller size can be proven, as there will be more evidence for statistical significance.

When using such a large multifactorial model, there is the risk that the different factors are not independent, and partially explain the same portion of the variation. For example, it might be the case that modal auxiliary verbs are generally also more likely to be used in rhyming text. Statistically, we can test for this by computing the variance inflation factor (VIF) of the variables in the model. This is a measure of multicollinearity — correlation between predictor variables. In our model, all of the VIFs are very low ($< 1.3$, less than 10 is considered acceptable). The factors therefore do not appear to correlate with each other, are independent, and do not statistically account for the same parts of the variation.

### 8.6 Discussion

Using the results above, we can now expand upon the three possible sources of verb cluster order variation in Early-Modern Frisian that we proposed in this article. These sources were the following:

1. Variation in Early-Modern Frisian texts is not due to contact, but a continuation of the mediaeval situation.

2. Variation in Early-Modern Frisian texts is due to contact through bilingualism, with early acquisition of the optionality, similar to the modern Frisian situation and the arguments made by Bremmer (1997).

3. Variation in Early-Modern Frisian texts is due to learned borrowing, with late acquisition of the optionality, along the lines of Blom’s (2008) account for the mixed Frisian texts.

To establish whether there was a language contact effect, and what type, we are interested in the similarities and differences to the factors that affect modern Dutch verb cluster order variation.

Firstly, our model of verb cluster order in the Early-Modern Frisian corpus showed us that there is a clear effect of text type on the use of 1-2 orders: the Dutch-like 1-2 order is used far more in rhyming text. While it can also be used in prose text, the stylistically demanding context of rhyme apparently prompted Early-Modern Frisian writers to make more use of the 1-2 option. The high usage percentages in the rhyming texts shows that the Early-Modern Frisian writers had no problem using this word order as a stylistic option, at least.

Furthermore, we observed an association between to-infinitives and the 1-2 order. This link was also noted by Hoekstra (2012), who found to-infinitives to
be relatively frequent in the 1-2 order in a Gysbert Japicx text. Hoekstra argues that this indicates borrowing from Dutch, where to-infinitives are said to be head-initial (i.e. auxiliary-first). This construction may have been convenient to borrow because there was no similar construction in Frisian, and it seems like it came as a package with the 1-2 cluster order from Dutch. A similar argument could be made for our finding that aspectual verbs are more associated with the 1-2 order — they are also more commonly used in Dutch than in Frisian (Hoekstra, 2012). These findings make the ‘no contact’ option a rather unlikely option.

Hoekstra (2012) provides the additional observation that larger verb clusters (with more than three verbs) are more likely to be in an ascending order (i.e. 1-2-3), but the sample size is small in that case, and we have not examined larger clusters in detail in the present study.

Generally, we did not observe much evidence for the processing effect that was observed for modern Dutch verb cluster order variation (Bloem et al., 2017b). While we were only able to make use of a limited range of factors, we found no evidence for typical processing effects such as an increased 1-2 order preference for separable verbs in our Early-Modern Frisian data. This indicates that the Early-Modern Frisian 1-2 order did not have the function of reducing processing loads in certain contexts that it appears to have in Dutch. There was also no evidence of a priming effect, although this is difficult to measure in text — in modern written Dutch, writers are more likely to use a 1-2 order after another 1-2 order, showing that these orders are elements that are integrated well enough into the language system to undergo structural priming. We cannot say this for the Early-Modern Frisian data.

It seems more plausible that the 1-2 order is mainly a stylistic device used by these authors in the written modality, rather than a construction with the function of decreasing language processing load as in Dutch. This points at the conclusion that the contact effects we do observe are a consequence of learned borrowing, not of widespread bilingualism as in modern Frisian.

Other examples of learned borrowings and similar phenomena can be found in historical Frisian texts. Brandsma (1936, p. 23) notes Gysbert Japicx’s overuse of -je-verbs. Frisian has two classes of weak verbs. Verbs in class I end in -e, which corresponds to standard Dutch verb endings, and verbs in class II end in -je. There is no Dutch or Low Saxon analogy for this second class. Brandsma’s (1936) study shows that Gysbert Japicx has mixed up the two classes in his writings. There were not much more than ten verb types of class I where he always used class I endings, while there were ‘many’ verbs that he used with different class endings — mainly class II endings on verbs that were originally class I, but some examples of the reverse were also found. It appears that Japicx overused the ending that is typical of Frisian. This might be a case of hypercorrection caused by knowledge of Dutch or Dutch writing, in order to avoid Dutch interference. It might also be a case of late acquisition (either adult L1 acquisition or L2 acquisition) in which not all the subtleties of the system were acquired by Japicx, resulting in this overgeneralization of class II endings.
8.7 Conclusion

Our data suggests that learned borrowing is the primary source of 1-2 verb cluster orders in Early-Modern Frisian texts. The lack of a function for the 1-2 order beyond stylistic usage indicates late acquisition, perhaps when the writers of these texts learned to write in another language than Frisian, as Frisian did not have much of a written tradition. In this case, the subtleties of the usage patterns of the 1-2 order in Dutch were either not fully acquired or not fully transferred into their Frisian. At the same time, there are clear signs of language contact, as some verb cluster constructions (such as the ones with *to*-infinitives) appear to have been borrowed from Dutch.

In the older, language-mixed Basle Wedding Speeches however, we did not find signs of borrowing of the 1-2 order, as there was no correlation between 1-2 verb cluster word order and the use of Dutch word forms. This indicated to us that Middle Frisian retained the 1-2 order option from Proto-West-Germanic. We can therefore also not exclude that this language-internal source still played a role in Early-Modern Frisian, given the fact that the authors from the early-Modern period apparently did not consider the use of 1-2 orders to be ungrammatical in their writing. However, in early 20th century Frisian the 1-2 order appears to be completely ungrammatical, so the present day usage of the 1-2 order cannot be a remnant of Proto-West-Germanic, but is rather a consequence of widespread bilingualism among the Frisian-speaking population, as evidenced by the fact that it has been observed in children as young as 4 to 6 years old (Meyer et al., 2015b).

We conclude that these 1-2 verb cluster word orders used in Early-Modern Frisian texts likely came from Dutch, but do not have the same underlying cause as the modern Dutch word orders. The contact-induced change that led to the present modern Frisian situation appears to be recent.
Chapter Highlights

Problem Statement

- Word orders in Dutch verb clusters have been shifting over time from the participle – auxiliary (2-1) to the auxiliary – participle (1-2) order. Synchronically, both orders are used. Diachronic change typically leads to synchronic variation, but it is not clear whether these two observations are linked, and whether the diachronic change is still ongoing.

Research Questions

- Is the present-day synchronic variation related to the diachronic development observed in historical text?
- Is the diachronic shift towards the 1-2-order still ongoing?
- How can a diachronic word order change be studied synchronically?

Research Contributions

- Our study of synchronic data from a spoken Dutch corpus show that younger speakers indeed use the 1-2 order more frequently.

*This chapter was adapted from Olthof, Marieke, Westendorp, Maud, Bloem, Jelke, & Weerman, Fred (2017). Synchronous variation and diachronic change in Dutch two-verb clusters. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 133(1), 34-60.
9.1 Introduction

A well-known issue in Dutch linguistics concerns the order variation found in verb clusters. A verb cluster can be defined as a combination of two or more verbs in the right periphery of a subordinate clause, which typically cannot be interrupted and which have a shared argument structure (Coupé, 2015, p. 1-2). One common type of two-verb clusters consists of a perfective or passive auxiliary and a past participle. In Dutch, these two-verb clusters can both show the order perfective/passive auxiliary – participle, as in examples 9.1 and 9.3, and the order participle – perfective/passive auxiliary, as shown in 9.2 and 9.4.

(9.1) Jan zeg-t dat hij heeft
Jan say-3SG.PRS COMP 3SG.M.NOM have.3SG.PRS
PFV
ge-fiets-t.
PTCP-cycle-PTCP
PTCP

(9.2) Jan zeg-t dat hij ge-fiets-t
Jan say-3SG.PRS COMP 3SG.M.NOM PTCP-cycle-PTCP
PTCP
heeft.
have.3SG.PRS
PFV
‘Jan says that he has cycled.’

(9.3) Jan zie-t een deur die word-t
Jan see-3SG.PRS INDF door REL become.3SG.PRS
PASS
ge-schilder-d.
PTCP-paint-PTCP
PTCP

Verb clusters also occur in main clauses. However, most studies focus on subordinate clause clusters, where there is no V2 effect. As our study only involves subordinate clause clusters as well, we provide this definition.

Synchronic variation and diachronic change in Dutch two-verb clusters

(9.4) Jan zie-t een deur die ge-schilder-d
      Jan see-3SG.PRS INDF door REL PTCP-paint-PTCP
      PTCP

word-t.
become.3SG.PRS
PASS

'Jan sees a door that is being painted.'

These auxiliary-participle clusters can occur with three different auxiliaries: the perfective auxiliaries hebben ‘to have’ and zijn ‘to be’ or the passive auxiliaries worden ‘to become’ and zijn ‘to be’. Since the auxiliary is higher in the hierarchical structure of the cluster than the participle, the auxiliary is called verb 1 and the participle is called verb 2 (Coupé, 2015, p. 19-20; Wurmbrand, 2006, 229; Zwart, 2011, p. 296). Consequently, the order auxiliary – participle, shown in 9.1 and 9.3 is known as the 1-2 order, whereas the order participle – auxiliary, exemplified in 9.2 and 9.4 is described as the 2-1 order. Longer clusters are also possible, such as an 1-2-3 order, in which case verb number 3 is the participle. The 1-2 order is also known as the ‘red order’, while the 2-1 order is sometimes called the ‘green order’, referring to the colours used for the two orders on dialect maps in an early study on verb clusters by Pauwels (1953).

Two-verb clusters are typically divided into two types (Wurmbrand, 2004, p. 44): clusters with a perfective/passive auxiliary and a past participle, exemplified in 9.1 to 9.4, and clusters with a modal auxiliary and an infinitive, such as the one shown in 9.5 and 9.6:

(9.5) Jan zeg-t dat hij wil slap-en.
      Jan say-3SG.PRS COMP 3SG.M.NOM want.3SG.PRS sleep-INF
      MOD INF

(9.6) Jan zeg-t dat hij slap-en wil.
      Jan say-3SG.PRS COMP 3SG.M.NOM sleep-INF want.3SG.PRS
      INF MOD

'Jan says that he wants to sleep.'

Clusters of the latter type may also have an aspectual auxiliary, a causative auxiliary or a perception verb, instead of a modal auxiliary (Barbiers et al., 2008, p. 12). Both clusters with a perfective/passive auxiliary and a participle and clusters combining a modal auxiliary and an infinitive allow two order possibilities: 1-2 and 2-1 (Wurmbrand, 2004, p. 45). However, the auxiliary-participle clusters exhibit a lot more variation than the modal-infinitive clusters, which are used almost exclusively in the 1-2 order nowadays.

Dutch verb clusters and their order variation have been the topic of a large number of linguistic studies. Most of these studies focus on subordinate clauses, as these are the only clauses in which all verbs are grouped together (Coupé, 2015, 10; Swerts, 1998, p. 300; Zwart, 1996, p. 232). Although Dutch is generally classified as a verb-final language, in its main clauses the so-called ‘verb second’
rule places the finite verb in the second position of the clause (Koster, 1975, p. 111). In main clauses only non-finite verbs occur clause-finally. If a main clause contains two verbs, they do not cluster together.

Interestingly, while order variation is usually associated with variation in meaning or information structure, the different orders in verb clusters are generally thought to not express systematically distinct semantics or pragmatics (Coupé, 2015, p. 27; Coussé, 2008, p. 2; De Schutter, 2012, p. 6-7), although some possible semantic associations have been noted (Pardoen, 1991; Bloem, 2016b). Consequently, an important question in the research on Dutch verb clusters is how we can account for this order variation. Many researchers have focused on possible explanations for the variation that has been observed and on the different factors and circumstances that may affect the choice for one or the other order. Different linguistic and sociolinguistic variables have indeed been shown to influence the choice between the 1-2 and the 2-1 order in particular contexts. This variation can take the form of interspeaker variation, between different speaker groups (such as people from different regions) or intraspeaker variation, where a single person may use both word orders, depending on various factors.

In addition, it has been observed that the variation between the two orders has not been stable over time. Therefore, the historical perspective is important when studying verb cluster variation. Whereas Dutch around 1400 almost exclusively showed 2-1 orders, over the centuries the 1-2 order has become more and more frequent (Coupé, 2015; Coussé, 2008). Crucially, Coussé (2008, p 190) argues that the change towards more 1-2 orders is still in progress.

In previous work, these observations on synchronic and diachronic variation have not been linked empirically. But as ongoing language change typically causes synchronic variation between old and new forms (Labov, 1965), and as the availability of the two verb cluster orders has been hypothesized to be caused by related diachronic changes (van den Berg, 1949; Coussé, 2008), we may hypothesize that the order variation in two-verb clusters in present-day Dutch also depends on the changing preferences with respect to these orders. The present paper investigates whether the synchronic variation in the orders of Dutch two-verb clusters can indeed be explained on the basis of the diachronic development towards more 1-2 orders, by means of an apparent-time study on spoken Dutch corpus data. We also test whether there is intraspeaker variation, rather than just different people using different word orders.

The outline of the paper is as follows. Section 9.2 addresses the factors that have been shown to affect the choice for a particular verb cluster order. In addition, studies on the diachronic changes in the use of verb cluster orders are reviewed. Section 9.3 then provides some important theoretical and methodological notions concerning the relation between synchronic variation and diachronic change. In addition, it is shown how these notions lead to predictions about the synchronic order variation in Dutch verb clusters. The method of the study is presented in section 9.4, which also addresses issues in the data analysis. Subsequently, section 9.5 contains the results of the study. Finally, section 9.6
Synchronic variation and diachronic change in Dutch two-verb clusters

summarizes the outcomes of the study and concludes whether they support the hypothesis that the synchronic order variation in Dutch verb clusters is related to a diachronic change.

9.2 Background: variation and change in the order of Dutch verb clusters

9.2.1 Factors influencing the order variation in verb clusters in present-day Dutch

Since two-verb clusters allow for order variation without obvious semantic or pragmatic correlations, many studies have investigated what may determine the choice for one or the other order. This section summarizes the generalizations that have been made in the literature, starting with more contextual factors that vary between speakers (interspeaker variation) and then moving to more language-internal factors, which may also affect variation within the same speaker (intraspeaker variation).

To start, the regional background of the language user clearly plays a role in the choice between the orders. Different dialects of Dutch show different order preferences (Pauwels, 1953; Stroop, 1970; Barbiers et al., 2008). For instance, important differences are found between Dutch spoken in Belgium, i.e. Flemish, and Dutch spoken in the Netherlands (De Sutter et al., 2005, 109-111). In Dutch spoken in the Netherlands the 1-2 order is more frequent than in Flemish.

In addition, the difference between spoken and written language use seems to play a role in the order variation in Dutch verb clusters (De Sutter et al., 2005, p. 114; Stroop, 2009, p. 461. For instance, (Stroop, 2009, p. 461) shows that in spoken Dutch, the 2-1 order is preferred over the 1-2 order in perfective/passive auxiliary – participle clusters. According to his data, 2-1 occurs in 63% of the spoken two-verb clusters of this type. By contrast, (Arfs, 2007a, p. 229) focuses on written Dutch and finds that the 1-2 order is used in 72% of the perfective/passive auxiliary – participle clusters and thus appears more often than the 2-1 order. Thus, the difference between spoken and written language also plays a role in order variation. In relation to these findings it has been noted that many people believe that the 2-1 order is not correct as it is the order used in German (Haeseryn, 1990, 40; Stroop, 2009; Swerts, 1998, p. 300.

For this reason, speakers may try to avoid this order, and they may especially do so in written language, as people tend to follow such normative ideas more strictly in written than in spoken language (Haeseryn, 1990, 37). In addition, people may think that the order that they use in their dialect, which is often the 2-1 order, is not correct in standard, written language (De Sutter et al., 2005, p. 102). A somewhat related factor affecting the choice between the 1-2 and the 2-1 order involves the register or context of language use. De Sutter et al. (2005, p. 122-123) show for instance that in dialogues the 2-1 order occurs...
more often than in monologues. In addition, clusters in texts that have been edited show more 1-2 orders than clusters in non-edited texts (De Sutter et al., 2005, p. 122).

The factors just discussed all indicate variation between different contexts, domains and speakers, but there are also factors suggesting intraspeaker grammatical variation, i.e. a single speaker choosing between two possible word orders that are both grammatical and suitable in a particular context.

One such factor that has been shown to influence the choice for either the 1-2 or the 2-1 order concerns the type of syntagm, i.e. which kinds of verbs a cluster consists of. Clusters consisting of a modal auxiliary and an infinitive almost exclusively occur in the 1-2 order, whereas clusters with a perfective or passive auxiliary combined with a participle use both the 1-2 and the 2-1 order quite frequently (Bloem et al., 2017b, p. 11; Coupé, 2015, p. 27-28; Haeseryn, 1990, p. 44-46; Wurmbrand, 2004, p. 43; Zwart, 2011, p. 44). In addition, clusters with the perfective auxiliaries hebben ‘have’ and zijn ‘be’ seem to show more 1-2 orders than clusters with passive auxiliaries zijn ‘be’ and worden ‘become’ (De Schutter, 2012, p. 7-8; Haeseryn, 1990, p. 44-45). A possible explanation for this finding involves the observation that zijn and worden, but not hebben, also function as copulas in clauses with adjectival predicates. In subordinate clauses the copula and adjectival predicate occur in the right periphery, just like verb clusters, but, in contrast to verb clusters, they always show the order adjective – zijn/worden, as in example 9.7 below:

(9.7) Jan zeg-t dat de winkel open is.
(9.8) * Jan zeg-t dat de winkel is open.
(9.9) Jan zeg-t dat de winkel ge-slot-en
(9.10) Jan zeg-t dat de winkel is ge-slot-en.

‘Jan says that the store is open.’
‘Jan says that the store has been closed.’
With an adjective, the order in example 9.8 is not grammatical, as this sentence is not verb-final. The very similar example 9.10 is grammatical, as long as *gesloten* is interpreted as a participle, not as an adjective. However, it may be dispreferred by analogy with 9.8. Thus, the relatively high frequency of the order participle – *zijn/worden* may be due to analogy with the order of the construction adjective – *zijn/worden* (De Schutter, 2012, p. 7-8; Haeseryn, 1990, p. 44-45). The main verb or the type of main verb has also been studied as a factor (Pardoen, 1991; Bloem, 2016b).

Furthermore, a few studies focus on processing complexity (Bloem et al., 2017b; De Sutter, 2007). For instance, Bloem et al. hypothesize that the 1-2 order is easier to process, since it occurs most often in complex contexts, which themselves already require a lot of processing. In easier contexts, by contrast, 2-1 orders occur frequently. Processing complexity thus also seems to play a role in the choice between the 1-2 and the 2-1 order.

Researchers such as Arfs (2007a) and Swerts (1998) have also focused on the effect of rhythm and prosody on the choice between the 1-2 and the 2-1 order. For instance, Swerts (1998, p. 304) finds that when the prosodic accent is on the participle, the 2-1 order is preferred. By contrast, when the element directly preceding the verb clusters is stressed, the 1-2 order occurs more frequently.

Finally, priming may play a role in the choice between the 1-2 and the 2-1 order. Hartsuiker & Westenberg (2000, p. 36) find that when a clause contains a cluster with a 1-2 order, speakers tend to use the 1-2 order also in a cluster in the next clause. By contrast, a 2-1 order may lead to the use of another 2-1 order in a following clause. This effect occurs both in their spoken and written experiment.

9.2.2 Diachronic developments in the order variation in Dutch two-verb clusters

Various studies have demonstrated how a variety of linguistic and sociolinguistic factors influence the order variation in verb clusters found in present-day standard Dutch. It may be noted however, that these studies generally only take into account the synchronic order variation, while the issue of order variation in verb clusters also has an important diachronic dimension.

Diachronic studies have shown that, during the past centuries, the Dutch verb clusters have undergone a noteworthy development with respect to their orders (Coupé, 2015; Coussé, 2008). Already in the earliest documents used in these studies, from the second part of the 13th century, both two-verb clusters with the 1-2 and with the 2-1 order occurred. The variation between the two orders is thus remarkably old (De Schutter, 2012, p. 6-7). However, soon a preference for the 2-1 order arose, leading to the establishment of the 2-1 order as the dominant order around the year 1400. The 15th and 16th century both showed a preference for the 2-1 order. However, already during the 16th century, a development in the opposite direction began and the use of the 1-2 order started to increase. This change first seems to have affected two-verb clusters
with a modal auxiliary, such as zullen ‘shall’, and an infinitive, while clusters with
perfective auxiliaries, such as hebben, and passive auxiliaries followed slightly
later (Couvé, 2015, p. 104). The increase in the frequency of the 1-2 order in
both modal auxiliary – infinitive and passive/perfective auxiliary – participle
clusters continued in the following centuries (Coussé, 2008).

Looking at present-day standard Dutch, the change towards more 1-2 orders
in modal auxiliary – infinitival clusters appears to have proceeded so far that, as
discussed in section 9.1, these clusters now almost exclusively show the 1-2 order.
By contrast, the perfective/passive auxiliary – participle clusters still show a
high degree of variation between the two orders. Importantly, according to
Coussé (2008, p. 190) in these perfective/passive auxiliary – participle clusters,
the change in order preference that started four centuries ago is still ongoing
today, and the use of the 1-2 order is thus still increasing. Consequently, this
diachronic development may play a role in the variation that can be found in
standard Dutch today.

An agent-based simulation by Bloem et al. (2015) was able to account for the
change towards the 1-2 order up until the 16th century on the basis of frequencies
of occurrence of various verb cluster constructions, using an exemplar-based
model of interacting language agents. They propose that this change was driven
by the grammaticalization of ‘have’ as an auxiliary verb, and by the increasing
use of subordinate clauses. However, this model was not able to account for the
subsequent change towards 1-2 orders.

We should also consider that acquisition plays a role in the historical change
from the descending to the ascending word order. Meyer & Weerman (2016)
argue that in first language acquisition of verb clusters, the 1-2 order is learned
as the default order. Subsequently, children have to discover how and when to
deviate from this default. However, as the distribution of 1-2 and 2-1 orders
is quite complex, children may acquire a slightly different distribution pattern
than that of their parents and other adults. Since the 1-2 order is the default,
changes leading to an increase of use of the ‘default’ 1-2 order can be expected
here. This way, continuous development in the direction of ascending clusters
in combination with a preference for a certain order in acquisition may lead
to differences between generations. Interestingly, the acquisitional pathway as
proposed by Meyer & Weerman (2016) mimics the pattern of diachronic change
in Dutch two-verb clusters. Both processes start with modal-infinitive clusters
and subsequently generalize the ascending order over all cluster types.

Leaving aside regional variation and other contextual factors, the next section
discusses the ways in which diachronic changes may be visible in synchronic
variation, and formulates a hypothesis about how the diachronic dimension may
affect the use of the 1-2 and 2-1 orders in present-day Dutch.
9.3 The relation between diachronic change and synchronic variation

9.3.1 Synchronic variation and diachronic change connected

Sociolinguistic research and studies with a focus on language change have shown that the synchronic and the diachronic dimensions of a linguistic phenomenon are directly related (Meyerhoff, 2006, p. 133-134). Synchronic variation is often the source of a linguistic change (Kay, 1975; Weinreich et al., 1968). At the same time, diachronic changes often result in synchronic variation (Bickerton, 1973; Labov, 1965). These observations may also hold for the order variation found in verb clusters.

One of the first and most famous studies showing a direct relation between synchronic variation and diachronic change is Labov’s (1963) study of Martha’s Vineyard, an island on the east coast of the United States. Labov compared his own findings on diphthong pronunciation to results from a number of interviews held thirty years earlier, which reported on the pronunciation of the same diphthongs by speakers from the same island. The data from these interviews showed very little or even no centralization in the diphthongs under investigation, while Labov’s own data did show it, and more central pronunciations were produced by younger speakers. Consequently, Labov argued that the centralisation had to be a recent development, and concluded that his data revealed a change in progress.

Labov’s study shows an example of how synchronic variation between different age groups can be explained as a result of a diachronic change. More specifically, with respect to a particular linguistic phenomenon, younger speakers may represent a newer stage of the language than older speakers. One explanation for this pattern is that each generation acquires the language during the critical period, i.e. early childhood, and that speakers generally do not change their grammars substantially later in life (Boberg, 2004, p. 256; Meyerhoff, 2006, p. 133). Speakers thus keep using the language as it was when they acquired it. Following this line of reasoning, it is in the acquisition process that changes may arise, and indeed, children do often not acquire exactly the same grammar as their parents. Moreover, since each generation in this way shows slightly different language use, children’s input may also be somewhat different from that of their parents. In this way, language changes may proceed. As different generations live together and communicate, the differences between them appear as synchronic variation.

9.3.2 The apparent-time method

A real diachronic study requires longitudinal data, which may take several decades to gather. This procedure clearly has practical disadvantages, such as
9.3. The relation between diachronic change and synchronic variation

the possibility that participants die and the consequence that the researcher
must wait many years before results are obtained (McMahon, 1994, p. 240).
Researchers who want to investigate a diachronic change may therefore instead
make use of the so-called apparent-time method (Labov, 1965), which is based
on the observation that synchronic variation between different age groups may
be due to an ongoing language change (Boberg, 2004, p. 250-251; McMahon,
1994, p. 240. In this method, one compares the synchronic linguistic behaviour
of different generations and assumes that the generations reflect different stages
of the language. Differences between these age groups may then show that a
particular phenomenon gradually increases or decreases in frequency over time.

The apparent-time method thus provides a relatively easy procedure to
investigate ongoing language changes on the basis of synchronic data. However,
data obtained through the apparent-time method that show differences between
the various age groups do not always reflect a language change. Differences
between speakers of different ages may also be caused by age-grading, i.e. a
stable situation with variation between generations (Labov, 1965, 1994; Sankoff,
2006). Some language features may be associated with a particular phase in life,
in that speakers may use a certain feature when they have a particular age but
abandon this feature again when they grow older. If this pattern of acquiring
and dropping features remains constant through the generations, differences
found between age groups are due to age-grading, and no diachronic change is
involved.

A possible way to decide whether one’s data reflect an ongoing change or
an age-grading effect is to combine the apparent-time data with real-time data
(Boberg, 2004, p. 25; McMahon, 1994, p. 240; Sankoff, 2006). Thus, when real
diachronic data have already indicated that a change is going on or has been
going on, it is very likely that differences between the age groups that suggest
a similar change indeed are due to this diachronic change. This strategy was
also applied by Labov (1963, 1965), who combined data from earlier interviews
with his apparent-time data.

Although the apparent-time method has sometimes been criticized for being
a theoretical construct rather than a valid practical method, a large number of
studies have shown that apparent-time data indeed are able to show ongoing
linguistic change, as long as the method is used in the right way (Sankoff,
2006). For instance, an important assumption in the method is that language
acquisition is limited to the early years of life and that only few changes occur
in adulthood. This claim seems to hold mainly for ‘the more abstract levels of
grammar, such as phonology and syntax’, which make these domains suitable for
the apparent-time method (Boberg, 2004, p. 265). By contrast, lexical knowledge
typically develops during life and differences between age groups in lexical
knowledge thus do not have to reflect a diachronic change (Meyerhoff, 2006, p.
152). In addition, it has been shown that the apparent-time method is not always
able to show the rate of change, but the direction of a change can usually be well
attested (Sankoff, 2006). If such limitations are kept in mind, the apparent-time
method generally seems quite robust (Meyerhoff, 2006, p. 152; Sankoff, 2006).
In the present study, which focuses on a syntactic phenomenon and moreover investigates the presence rather than the rate of a change, we therefore adopt the apparent-time method as well.

9.3.3 Hypothesis and predictions

Based on previous work on synchronic variation and diachronic change in the order of the verbs in Dutch two-verb clusters of the type perfective/passive auxiliary – participle, we may now hypothesize that the order variation that is found in these clusters is at least partly due to an ongoing language change. We hypothesize that interspeaker variation may be a result of different generations having different word order preferences, and that intraspeaker variation may be enabled by an ongoing language change, with two competing forms available in the grammar.

Diachronic data have shown that while 600 years ago the 2-1 order was very dominant, an increase in the frequency of the 1-2 order has taken place since the 16th century, which still may be ongoing. At the same time, synchronic data have shown that there is indeed still variation between the 1-2 and the 2-1 order. The present study therefore investigates whether the change is indeed still proceeding and how this change can be linked to the variation in the order of the verbs today.

On the basis of the hypothesis, the prediction is that younger speakers use the 1-2 order more frequently in perfective/passive auxiliary – participle clusters than older speakers, as the younger speakers are expected to reflect a later stage in the diachronic change than the older ones. An apparent-time study focusing on the orders that speakers of different ages use in their perfective/passive auxiliary – participle clusters may show whether this is indeed the case. We compare speakers of different age groups on their proportions of 1-2 and 2-1 orders. Importantly, if differences between age groups are found, it is very likely that these indeed reflect language change and not age-grading, since the diachronic data gathered by Coupé (2015) and Coussé (2008) have already shown that a change in order preferences in Dutch clusters has been initiated in the past.

If the prediction that younger speakers use more 1-2 orders than older speakers is borne out, we may conclude that Dutch is indeed still undergoing a change in order preferences. Moreover, such a finding may add to our understanding of the synchronic variation in the clusters. By contrast, if the prediction is not borne out, the diachronic change that Coupé (2015) and Coussé (2008) describe may no longer be ongoing. We must then assume that this diachronic change cannot have a direct influence on the synchronic variation in terms of differences between age groups.
9.4 Method and analysis

9.4.1 Corpus Gesproken Nederlands (CGN)

The data for the apparent-time study on the order variation in perfective/passive auxiliary – participle clusters were taken from the Corpus Gesproken Nederlands (CGN, corpus of spoken Dutch, Oostdijk et al., 2002), a 10-million word corpus of spoken language from Dutch and Flemish speakers, recorded between 1998 and 2003. The corpus comprises different types of speech, such as conversations with varying degrees of spontaneity, formal and informal speech situations, broadcasted speech and monologues (Oostdijk & Broeder, 2003). Table 9.1 summarizes the text types that are included.

Metadata are included for each speaker in the corpus, which make it possible to classify the speakers on the basis of their birth years. In this way, we formed six different age groups, which are shown in Table 9.2:

Each age group covers a period of 10 years, except for the group of oldest speakers, which comprises 15 years, as the data from older speakers is scarcest. Since the data were recorded between 1998-2003 and the youngest speakers were born in 1984, the corpus only includes speakers that were 14 years or older. It may therefore be assumed that all data in our study are from speakers that have already fully acquired the language. Previous research has shown that speakers as young as 8 years old may already provide useful data in apparent-time studies (Labov, 1994, p. 49).

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tta</td>
<td>Spontaneous conversations (face-to-face)</td>
</tr>
<tr>
<td>ttb</td>
<td>Interviews with teachers of Dutch</td>
</tr>
<tr>
<td>ttc</td>
<td>Spontaneous telephone dialogues (recorded via a switchboard)</td>
</tr>
<tr>
<td>ttd</td>
<td>Spontaneous telephone dialogues (recorded on MD with local interface)</td>
</tr>
<tr>
<td>tte</td>
<td>Simulated business negotiations</td>
</tr>
<tr>
<td>ttf</td>
<td>Interviews/discussions/debates (broadcast)</td>
</tr>
<tr>
<td>ttg</td>
<td>(political) discussions/debates/meetings (non-broadcast)</td>
</tr>
<tr>
<td>tth</td>
<td>Lessons recorded in a classroom</td>
</tr>
<tr>
<td>tti</td>
<td>Live (e.g. sport) commentaries (broadcast)</td>
</tr>
<tr>
<td>ttj</td>
<td>Newsreports/reportages (broadcast)</td>
</tr>
<tr>
<td>ttk</td>
<td>News (broadcast)</td>
</tr>
<tr>
<td>ttl</td>
<td>Commentaries/columns/reviews (broadcast)</td>
</tr>
<tr>
<td>ttm</td>
<td>Ceremonious speeches/sermons</td>
</tr>
<tr>
<td>ttn</td>
<td>Lectures/seminars</td>
</tr>
<tr>
<td>tto</td>
<td>Read speech</td>
</tr>
</tbody>
</table>

Table 9.1: Text types included in CGN (Weijers, 2004, p. 37)
The CGN also includes different types of annotations, which are useful when searching the corpus for particular linguistic phenomena. Two types of annotations are relevant for the present study. Firstly, the whole corpus is annotated with parts-of-speech tags (Van Eynde, 2004), which make it possible to search for different parts of speech such as verbs and subordinating conjunctions. Secondly, about 10% of the corpus is syntactically annotated (Hoekstra et al., 2003). With the help of the syntactic annotations, particular syntactic structures such as subordinate clauses can be found.

A potential concern regarding the use of the CGN is that it contains spoken language, while the historical sources that have been used in previous diachronic studies of verb cluster variation contain written language, and no spoken language corpora are available for these time periods. Since mode is a factor in verb cluster variation (De Sutter, 2005), it might be the case that verb clusters were also used differently in spoken language than in written language historically. However, the same argument can be made against comparing diachronic written data to synchronic written data – the circumstances under which the texts have been composed are radically different. Other relevant factors that can affect verb cluster variation may also have changed, such as prescriptive norms, formality of the texts, and the level of education and social class of the writers of the available texts. These factors have to be considered in any analysis referring to historical data, whether there is a modality difference or not. Our interest is mainly in the underlying language system, where modality is merely one factor of many affecting the observed variation. Furthermore, we will not perform a direct comparison with historical corpora – the only evidence we need is that there is an ongoing diachronic change that might have enabled the synchronic variation in spoken and written language. A change is observed by Coupé (2015) and Coussé (2008) for the written modality, and for the spoken modality there is no way to find out.

In order to find the perfective/passive auxiliary – participle clusters produced in the CGN, we used two different search tools, Parse & Query (PaQu) and the CGN Corpus Exploitation Software (COREX). PaQu is an application that...
was recently developed to enable querying of various Dutch language corpora via a web interface, including the CGN. For a more complete introduction and a demonstration study using PaQu, we refer to Odijk (2015). PaQu is easier to use than most existing querying systems, which often require knowledge of particular query languages, but its basic interface can only search for word pairs. Another easy-to-use web-based corpus search engine for Dutch is GrETEL (Greedy Extraction of Trees for Empirical Linguistics), which searches the CGN (Augustinus et al., 2013) and uses example sentences instead of queries. However, this application does not provide the speaker metadata, such as age, that we require. The two tools can be combined, though, and this is what we did in the present study. The PaQu tool allowed us to search for verb clusters on the basis of the syntactic annotations, while also providing information on speaker age. We composed the queries using GrETEL’s example-based querying method, and then ran the resulting XPath-expressions using PaQu’s XPath query function. However, because of these tools’ reliance on syntactic annotations, they can only search the syntactically annotated part of the CGN. In addition, we therefore used the Corpus Exploitation Software (COREX) of the CGN (Kilpatrick & Hellwig, 2002), to search the rest of the corpus for clusters on the basis of the parts-of-speech tags.

### 9.4.2 PaQu syntactic search

Since our study only concerns subordinate clause verb clusters, in the PaQu tool we searched for subordinate clauses. We furthermore specified that the subordinate clauses should contain a form of the verb *hebben*, *zijn* or *worden* with a participle as its verbal complement. We retrieved these verbs both in the 1-2 and the 2-1 order. The queries that we used to search the corpus can be found in the appendix, or reproduced by clicking the numbers in Table 3 and 4 below in the electronic version of this dissertation.

We then manually went through these results to remove some exceptional cases that do not correspond to our definition of two-verb clusters — they were erroneously retrieved by our search. This was due to the limitations of the standard GrETEL interface, which we used to compose our search queries. While it would have been possible to manually edit the XPath queries output by GrETEL, we chose not to do this, since it is difficult to properly formulate queries manually and exceptional cases are not that common. The first case is verb cluster interruptions, where a non-verbal element occurs inside something that was annotated as a cluster. These could have been excluded from a search by specifying adjacency requirements (i.e. requiring that one verb should be next to the other), but this is not possible in the standard GrETEL interface. Secondly, we manually removed the clusters of more than two verbs from the data, which are included due to GrETEL’s greedy extraction method, which automatically matches larger syntactic structures if the specified structure is a part of it. Furthermore, in a few cases the auxiliary and the participle were adjacent to each other, but one of them belonged to the main clause that followed...
Our search resulted in 2320 two-verb clusters of the type perfective/passive auxiliary – participle. The distribution of the 1-2 and 2-1 orders in these clusters, represented in Table 9.3, shows that, as argued in previous literature, both orders are used relatively frequently and there is no strong preference for one particular order.

By contrast, a quick check showed that the clusters of the type modal auxiliary – infinitive do not show this variation. When searching for subordinate clauses with a form of the verb *kunnen*, *moeten* or *zullen* and an infinitive as verbal complement, only very few examples of the 2-1 order appeared.

The data in Table 9.4 confirm the claim that the change towards more 1-2 orders in modal auxiliary – infinitive clusters seems to be almost completed. In the present study, the focus will therefore only be on perfective/passive auxiliary – participle clusters, for which there is still a large degree of order variation.

3The numbers reported by PaQu will be slightly higher than the ones listed in the table, because we have manually checked the results for false positives and removed them.

4The 1-2 clusters reported in this table have not been checked for three-verb clusters, interruptions and cases where one of the verbs is part of the main clause instead of the subordinate clause.
Method and analysis

Figure 9.1: Number of 1-2 and 2-1 orders in perfective/passive auxiliary – participle clusters in the syntactically annotated part of the CGN, by speakers from the Netherlands (NL) and Flanders, Belgium (BE).

A Pearson’s chi-square test shows that the distributions of 1-2 and 2-1 orders are estimated to be significantly different for the two countries ($\chi^2 = 55.6429; df = 1; p < .0001$). This observation may have very interesting implications, for instance, that the change towards more 1-2 orders may have started earlier or proceeded more quickly in the Netherlands than in Flanders. However, in the present study our focus is not on regional differences in order preferences. A detailed description of these differences in the CGN data can be found in Stroop (2009). Rather, we want to limit the influence from variables such as regional background as much as possible, in order to be able to zoom in on differences between age groups. Hence, in the search for clusters in our second search tool COREX and in the final results only clusters produced by speakers from the Netherlands will be included.

Research has also shown that the context or register of speech may have an influence on the order distributions. Figure 9.2 therefore shows the 2320 perfective/passive auxiliary – participle clusters split by text type, demonstrating that the distribution of 1-2 and 2-1 orders is quite different among text types. In general, the text types shown on the left side of the figure occur relatively frequently in the 2-1 order. These are the text types with the most spontaneous and informal speech. By contrast, text types containing more formal and
Synchronic variation and diachronic change in Dutch two-verb clusters

Figure 9.2: Number of 1-2 and 2-1 orders in perfective/passive auxiliary – participle clusters in the syntactically annotated part of the CGN, split by text type. The text types are described in table 9.1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tta</td>
<td>Spontaneous conversations (face-to-face)</td>
</tr>
<tr>
<td>ttb</td>
<td>Interviews with teachers of Dutch</td>
</tr>
<tr>
<td>ttc</td>
<td>Spontaneous telephone dialogues (recorded via a switchboard)</td>
</tr>
<tr>
<td>ttd</td>
<td>Spontaneous telephone dialogues (recorded on MD with local interface)</td>
</tr>
</tbody>
</table>

Table 9.5: Text types of the CGN included in the study.

Prepared language show a preference for the 1-2 order. We may thus conclude from these data that the context and register indeed also influence the use of the orders in our data. However, in the present study we are most interested in spontaneous spoken language. Register has been shown to be a factor affecting verb word order preferences, both in terms of the interactivity of a discourse type as well as in the amount of editorial control over prepared spoken texts (De Sutter, 2005, p. 89). In particular, prepared speech may have been edited by an editor from a different age group. We can control for this by limiting the data set to spontaneous spoken language. In our second search tool COREX and in the final results, we therefore only included the text types with the most spontaneous speech, i.e. text types a-d, as shown in Table 9.5. Using only these text types leaves us with 557 clusters from the PaQu dataset for the final results. In Stroop’s (2009) study on the use of verb clusters in spontaneous spoken Dutch on the basis of the CGN, the research is also limited to these four text types. It thus seems reasonable to focus on exactly these types.
9.4.3 COREX

In the COREX tool we used the parts-of-speech tags to find the examples that matched our definition of verb clusters. As noted in section 9.4.1, in this tool we excluded the Flemish speakers and limited our search to the text types a-d. Moreover, we excluded all speakers for which no birth year is included in the metadata of the CGN. Finally, as we had already searched the syntactically annotated part of the corpus in PaQu, and the process of checking the COREX results is more time consuming than the process of checking the PaQu results, we excluded this part from our search in the COREX tool.

In order to find the perfective/passive auxiliary – participle clusters in subordinate clauses, we looked for clauses with a subordinating conjunction followed by a form of the verb hebben, zijn or worden, which in turn had to be followed or preceded directly by a past participle. Since COREX allows for specification of such adjacency, all interrupted cluster-like verb groups were automatically excluded. To find subordinating conjunctions, we queried for the VG (voegwoord, ‘conjunction’) part-of-speech tag. Between the subordinating conjunction and the auxiliary verb, between 0 and 100 words were allowed to occur. We then manually removed all examples in which more than one subordinating conjunction occurred or in which the subordinating conjunction did not introduce a subordinate clause with a verb cluster, for instance because the cluster was part of a subordinate clause that itself was embedded in another subordinate clause. Finally, the examples in which either the auxiliary or the participle was part of the main clause instead of the subordinate clauses were omitted. In this way, the COREX search resulted in 3418 clusters.

A limitation of this search method is that we only selected subordinate clauses on the basis of subordinating conjunctions. Without having access to any syntactic annotation, that is the best remaining way of identifying subordinate clauses. However, subordinate clauses may also be introduced by a relative or interrogative pronoun instead of a conjunction. We did not expect an effect of this factor, but nevertheless, we compared the word order preferences for subordinate clauses introduced by different pronoun types. Figure 9.3 shows the distribution of 1-2 and 2-1 orders in subordinate clauses with a subordinating conjunction in text type a and in subordinate clauses with a relative or interrogative pronoun in the same text type. A Pearson’s chi-squared test shows that there is no evidence for a significant difference between the order distributions in the two types of subordinate clauses ($\chi^2 = 2.9523; df = 1; p = .08576$). Since there is little apparent effect, we will limit ourselves to clusters in subordinate clauses with a subordinating conjunction.

9.4.4 Analysis methods

From the clusters we gathered from the CGN, we took three properties as variables in our subsequent analysis: order, speaker age and cluster type. Speaker age was taken as a categorical variable with the categories shown in table...
Synchronic variation and diachronic change in Dutch two-verb clusters

Figure 9.3: Number of 1-2 and 2-1 orders in perfective/passive auxiliary – participle clusters in text type a of the CGN, split by type of subordinate clause.

2, and there are three cluster types: clusters with hebben, zijn and worden. Therefore, all of the variables are categorical. For our analysis, we first performed some descriptive statistics by creating histograms, and then used inferential statistics to test our predictions. To test for associations between word order and speaker age, we employ the Pearson’s chi-squared test for independence. This test is appropriate because all of the variables are categorical, the groups are independent, we are dealing with frequency data, and it does not make any distributional assumptions. Corpus data generally follows a Zipfian distribution, rather than a normal distribution. For a more detailed analysis, we also separated the clusters by cluster types and performed tests on all clusters of a particular type. Section 9.5 discusses the results of these analyses.

9.4.5 Summary of the selected data

All in all, the data selected to test our hypothesis are the following: all uninterrupted perfective/passive auxiliary – participle clusters in subordinate clauses produced by Dutch speakers from the Netherlands, born between 1920 and 1984, in the spontaneous conversations, interviews with Dutch teachers and telephone dialogues (text types a-d) in the CGN. From the syntactically annotated part of the corpus we took clusters from all types of subordinate clauses, while from the rest of the corpus only clusters from subordinate clauses introduced by a subordinating conjunction were extracted. There is no overlap between the data sets from COREX and PaQu, because the syntactically annotated data searched by PaQu was excluded from the COREX query. This combined set of data, a number of 3975 clusters, will be used in the next section, which presents the results of the research.
9.5 Results

The data gathered can now be used to test the hypothesis that the synchronic order variation in perfective/passive auxiliary – participle clusters is related to the diachronic change towards more 1-2 orders. This hypothesis predicts that the variation between 1-2 and 2-1 orders is related to different order preferences in our age groups. More specifically, younger speakers would use more 1-2 orders than the older speakers.

9.5.1 Intraspeaker variation

We will first discuss whether intraspeaker variation is actually present, using only the syntactically annotated section of the corpus (i.e. the section searched with PaQu) as an easily searchable sample of the data. Out of the 147 speakers in this sample, 84 produced more than one verb cluster. 62% of them used both cluster orders, and the highest number of clusters produced by one particular speaker was 11 (five 1-2 orders and six 2-1 orders). Most speakers produced only two or three clusters. Given that despite these low numbers, most speakers produce both orders, we can conclude that there is indeed intraspeaker variation in verb cluster ordering. For this analysis, we used the spontaneous speech text types (a-d) only, and only speakers from the Netherlands, and only from the syntactically annotated part of the corpus.

9.5.2 Word order preferences across age groups

As for the order frequency distributions, Figure 9.4 demonstrates first of all that all age groups show quite a high number of 2-1 orders. Apparently, the 2-1 order is still very frequent in spontaneous spoken Dutch, more so than in written language (cf. Bloem et al., 2017b). Importantly, however, a Pearson’s chi-squared test shows that there is a significant association between age group and distribution of 1-2 and 2-1 orders ($\chi^2 = 142.8204; df = 5; p < .0001$) in our data. This indicates that the variable age or birth year indeed plays a role in the synchronic variation in the perfective/passive auxiliaries, as predicted by the hypothesis. Moreover, as becomes clear in the figure, it is indeed the youngest group that shows the highest proportion of 1-2 orders, and there appears to be a trend of increasing 1-2 order preferences. By computing the Pearson’s correlation coefficient, we can see that the birth year groups positively correlate with the percentage of 1-2 orders used ($r = 0.913; n = 6; p = 0.011$), confirming this trend.

Only in this youngest group, there is a preference for the 1-2 order over the 2-1 order, and a post-hoc chi-square goodness of fit test shows that this preference is estimated to be significant compared to a baseline of 50% preference for each order ($\chi^2 = 14.7827; df = 1; p < .0001$). By contrast, all other, older groups,
still show a significant preference for the 2-1 order in post-hoc testing.\footnote{Age group 1920-1934: $\chi^2 = 60.1485; df = 1; p < .0001$, age group 1935-1944: $\chi^2 = 43.5567; df = 1; p < .0001$, age group 1945-1954: $\chi^2 = 55.9645; df = 1; p < .0001$, age group 1955-1964: $\chi^2 = 10.8407; df = 1; p < .001$, age group 1965-1974: $\chi^2 = 36.0502; df = 1; p < .0001$.}

Nevertheless, in general there seems to be an increase in the use of the 1-2 orders with age group, as shown in Figure 9.5: the younger the age group, the higher the relative frequency of the 1-2 order.

We may thus summarize that the variation between 1-2 and 2-1 orders is indeed associated with age group, and that the younger speakers indeed show more 1-2 orders than the older speakers. Furthermore, the youngest group of speakers suddenly shows a much stronger increase than the older groups. The youngest group also clearly deviates from the older groups in that this group shows a preference for the 1-2 order, while the 2-1 order is still dominant in the older groups.

9.5.3 Word order preferences across age groups per auxiliary verb

The factor ‘type of auxiliary’ significantly affects word order preferences in our data ($\chi^2 = 53.5147, p < 0.001; df = 2$), which is to be expected on the basis of previous studies that found the same kind of effect. Splitting the results by type of auxiliary yields some additional interesting results. Figure 9.6 to 9.8 present the numbers of 1-2 and the 2-1 orders in clusters with the auxiliary hebben, zijn and worden respectively. Note the different scales of the y-axes: the figures show that clusters with hebben occur more frequently than clusters with...
9.5. Results

Figure 9.5: Percentage of 1-2 orders in perfective/passive auxiliary – participle clusters in subordinate clauses in text types a – d in the CGN, split by age group.

A total number of 2086 clusters of the type hebben – participle was found, while the numbers of clusters with zijn and worden are 1352 and 537 respectively.

A first important result of the data presented in figure 9.6 to 9.8 is that for each type of cluster, i.e. those with hebben, zijn and worden as auxiliary verbs, there is a significant association between age group and 1-2 vs. 2-1 distribution, as estimated by Pearson’s chi-squared tests over each set of clusters. Thus, for all three cluster types, the distribution of 1-2 and 2-1 orders depends on the age group. Secondly, for each type of cluster, it is the youngest group that shows the most 1-2 orders. In the case of hebben, a post-hoc chi-square goodness of fit test shows that this group has a significant preference for the 1-2 order ($\chi^2 = 33.4848; df = 1; p < .0001$), and all other groups show more 2-1 than 1-2 orders. By contrast, in the zijn and worden clusters more 2-1 than 1-2 orders are found even in this youngest group. However, the preference for the 2-1 order in this group in clusters with zijn and worden is not significant. Finally, for each type of cluster, one can see a general trend showing an increase in the frequency of use of the 1-2 order with birth year.

Figure 9.9 shows that the younger groups generally use more 1-2 orders in each type of clusters than the older groups, with again a sudden increase in the use of the 1-2 order in the youngest group. It also demonstrates that the proportion of 1-2 orders is generally larger in the clusters with hebben than in those with zijn or worden as auxiliary. This finding matches earlier research by

---

6 For the hebben clusters: $\chi^2 = 96.3459; df = 5; p = value < .0001$, for the zijn clusters: $\chi^2 = 22.253; df = 5, p < .001$, for the worden clusters: $\chi^2 = 24.4464; df = 5; p < .001$.

7 For the zijn clusters: $\chi^2 = 0.198; df = 1; p = 0.6563$, for the worden clusters: $\chi^2 = 0.8462; df = 1; p = 0.3576$. 
Figure 9.6: Number of 1-2 and 2-1 orders in hebben – participle clusters in subordinate clauses in text types a – d in the CGN, split by age group.

Figure 9.7: Number of 1-2 and 2-1 orders in zijn – participle clusters in subordinate clauses in text types a – d in the CGN, split by age group.

Figure 9.8: Number of 1-2 and 2-1 orders in worden – participle clusters in subordinate clauses in text types a – d in the CGN, split by age group.
9.6 Discussion and conclusion

The present chapter has investigated the hypothesis that the synchronic variation in the order of two-verb clusters of the type perfective/passive auxiliary – participle is partly due to the diachronic change in the use of the different orders. Based on this hypothesis, it was predicted that younger speakers would use more 1-2 orders than older speakers. The apparent-time study has indeed shown that the distribution of 1-2 vs. 2-1 orders is associated with the age group factor. Moreover, the prediction that the youngest group of speakers uses the most 1-2 orders is confirmed, which corresponds to the older diachronic development reported by Coupé (2015) and Coussé (2008). In addition, the overall trend suggests that the frequency of use of the 1-2 order increases with the investigated age groups, i.e. the younger the age group, the more 1-2 orders are found. These results seem to indicate that the different age groups indeed reflect different stages of an ongoing change towards the more frequent use of the 1-2 order, causing synchronic interspeaker variation. Consequently, our hypothesis that the synchronic order variation in verb clusters is related to a diachronic change is borne out. It appears that a diachronic change is ongoing, enabling synchronic intraspeaker variation. We also showed that such intraspeaker variation is present in the corpus.

However, the data show a few issues that require further exploration. In the first place, a striking finding is the very sudden increase in the frequency

Haeseryn (1990) and De Schutter (2012) who suggest that this effect may be due to influence from the order adjective – copula.
of the 1-2 order in the youngest group, i.e. the speakers born between 1975 and 1984. This group is the only group with an overall preference for the 1-2 order, and the difference between the 1965-1974 and the 1975-1984 group is remarkably large. If we assume that the variation between the groups is due to a diachronic change, which has already been proceeding for several centuries, such an abrupt increase is quite notable. If this is an accurate observation, it may have consequences for the interpretation of language acquisition studies, whose younger subjects can be expected to have a greater 1-2 order preference than the average adult speaker.

Several explanations for this finding may be considered. Firstly, the notable distribution of 1-2 and 2-1 orders in the youngest group may be due to age-grading. However, as stated in section 9.3, age-grading is unlikely when real-time data also indicate a diachronic change, such as Coupé’s (2015) and Coussé’s (2008) data do in the case of the verb clusters.

Secondly, it is possible that, in contrast to our assumption, the youngest speakers have not fully completed their acquisition of the verb clusters and their order variation yet and may thus still be developing their order preferences. Following research on the acquisition of verb clusters, this scenario seems possible. Meyer & Weerman (2016) argue that the 1-2 order is the default order for clusters and that after an initial stage in which only 2-1 orders are used, due to analogical pressure from the object-verb word order, children tend to prefer the 1-2 order. In this development, the modal auxiliary – infinitive clusters seem to precede the perfective/passive auxiliary clusters. According to Meyer & Weerman, extra-linguistic factors may in the end lead to an adult-like, more balanced distribution of the 1-2 and 2-1 orders. One possibility is then that it takes children relatively long to abandon their 1-2 preference in favour of a more balanced distribution. However, in this scenario, the young speakers would be moving away from written and formal word order preferences (the 1-2 order is more prevalent in these domains) during their adolescence, which seems unlikely.

Thirdly, we may note that the 2-1 order is used more frequently in dialects. It is often assumed that the use of dialects steadily declines, especially among younger speakers (Barbiers, 2005, p. 8-9; Barbiers et al., 2006. One possibility is thus that the youngest group is those with fewest dialect speakers, which may have led to a reinforcement of the increase in the frequency of the 1-2 order in this group.

A second question that emerges on the basis of the patterns found in the research is how and why this change towards more 1-2 orders is going on and has been going on for so many years. Here, previous research may provide some suggestions. On the one hand, many of the factors that have been proposed to account for the synchronic variation in Dutch verb clusters probably do not have a stable effect over time. For instance, as discussed in section 9.4.1, the influence of regional background or style and register probably changes during 600 years. On the other hand, factors such as processing complexity may be able to stimulate a linguistic change over time: if the 1-2 order is indeed easier
to process than the 2-1 order, this difference between the orders may play a role at all times. Furthermore, the development towards the 1-2 order fits the observation that 2-1 orders are vulnerable in language acquisition and contact situations. As discussed by Meyer et al. (2015a), a comparison between English, Dutch, and German shows that language contact tends to lead to more 1-2 orders. English is the language that has experienced most language contact of these three languages and is also the language that only allows 1-2 orders. German, by contrast, has had relatively little language contact and allows only 2-1 orders. Dutch has had more language contact than German but less than English, and allows both the 1-2 and the 2-1 order. The availability of 1-2 and 2-1 orders shows a typical ‘Germanic sandwich’ pattern, with Dutch taking a typological position between German and English. This pattern has been observed for various other phenomena (van Haeringen, 1956), and has been linked, among other factors, to the extent to which these languages have been influenced by language contact (Weerman, 2006). Over time, this moderate degree of contact may have led to a slow rise in the frequency of the 1-2 order. In their agent-based model study, Bloem et al. (2015) also observe that model-internal factors cannot account for the Dutch change towards 1-2 orders, while the model does account for the current state of English and German. They speculate that language contact or standardization would be a logical explanation for the change in Dutch. This would also mean that the apparent sudden increase in use of the 1-2 order among the youngest group could be explained as a consequence of increased language contact, for example, with the English language, which only uses 1-2-like ordering in verb groups. It would be interesting to study a population of bilingual speakers to see whether such a word order preference transfer could take place. Furthermore, language contact has already been observed to affect verb cluster order in Frisian (Koeneman & Postma, 2006; Hoekstra & Versloot, 2016), showing that verb cluster orders can be sensitive to language contact.

However, another possibility is that the apparent recent increase in 1-2 order use is simply the S-shaped curve that one often finds in processes of language change – change is most rapid around the 50% threshold point, where the alternative form takes over from the previous form as the most dominant form. Indeed, our observations show 55% 1-2 orders in the youngest group, and 38.9% 1-2 orders in the second youngest group, which is around the 50% threshold. Therefore, an increase in the rate of change may not require a separate explanation at all.

Written historical sources could be used to provide additional evidence regarding the diachronic change. They, too, could be classified into more formal and less formal text types, and a study of verb clusters in the less formal text types could then provide more evidence indicating when the change might have started. It may be the case that the change started earlier in more formal registers, as there is evidence that the 1-2 order is stylistically preferred (De Sutter, 2005). This difference could also be compared to the synchronic differences in word order preferences per register.

Finally, it may be interesting to consider that the 2-1 order is still quite
frequent in spontaneous spoken Dutch, more so than in written language. The data by Coupé (2015) and Coussé (2008) show that the development towards more 1-2 orders started at least five centuries ago, in written language. And as was discussed earlier, present-day written Dutch shows a stronger preference for the 1-2 order. These observations suggest that the change towards more 1-2 orders may have progressed further in the written than in the spoken language, or it may even have started sooner in the written modality. This is an interesting possibility, as language changes typically seem to take place earlier and faster in spoken than in written form. For future work, it would be interesting to perform an apparent time study similar to the present study on a written corpus for which writer age metadata is available.

All in all, the present study has shown that in investigating the synchronic order variation in Dutch verb clusters, the diachronic dimension should also be taken into account. The research has provided empirical evidence showing that the change towards more 1-2 orders is still ongoing and that it leads to differences between speakers of different ages and thus to synchronic variation. Furthermore, the data presented in this study adds to our understanding of the most recent stage of this ongoing language change. In this way, the paper has tried to connect the synchronic and diachronic dimension of Dutch verb clusters and has hopefully contributed to the understanding of both the synchronic variation and the diachronic changes in the orders of Dutch two-verb clusters.
I started this dissertation by asking: why are there two ways to order clusters of two Dutch verbs, when the two orders do not have a clear meaning difference? While much has been written on this topic already, the aim of this dissertation was to use new quantitative and computational methods to paint a more complete picture of the many factors that affect this variation.

The previous eight chapters each discussed research questions aimed at revealing more about the variation. This concluding chapter will first summarize these questions and the studies that have addressed them in section 10.1, and then this dissertation is concluded in section 10.2 with a discussion of a core theme that turned out to be relevant for many aspects of verb cluster order variation: the role of processing.

10.1 Answers to the research questions

10.1.1 Automatically annotated corpora

In several chapters of this dissertation, I have made use of large, automatically annotated corpora, instead of taking the more typical corpus linguistics approach of using corpora that have been manually or semi-automatically annotated by experts. When using a new method, it is always good to evaluate it on a well-studied problem. This allows for a comparison of the new method to existing approaches by comparing the results of both methods on this well-studied problem. For the use of an automatically annotated corpus, this means comparing the results to those obtained from a manually annotated corpus.
A quantitative corpus study of two-verb clusters had already been performed by De Sutter (2009), on a manually annotated corpus of Flemish newspaper texts. As this study was described in sufficient detail, and because the kind of probabilistic word order preferences found in this study could be estimated more accurately on a larger dataset, I set out to replicate this study on an automatically annotated corpus in chapter 2, addressing the following question:

Can we replicate previous quantitative work on verb cluster order variation, which made use of manually annotated data, using only automatically annotated data?

Limiting the investigation to automatically annotated data also limited the extent to which we could replicate De Sutter’s (2009) study: only the information available in the annotation format of the corpus, the Lassy Large corpus (van Noord, 2009), could be used. Still, only one of the factors needed to be excluded entirely due to these limitations: distance between stress accents. The corpus did not contain word stress information. Some other factors had to be operationalized differently, which may have affected the results. For example, auxiliary verbs had to be grouped by lemma instead of by function, as their function (i.e. auxiliary of time) was not annotated. Nevertheless, besides the factor that was excluded, all of the factors that were shown to have effects on verb cluster order by De Sutter (2009) also significantly affected verb cluster order variation in an automatically annotated part of the Lassy Large corpus. This study therefore confirmed the effects that De Sutter found in his manual study.

Taking advantage of the larger dataset and automated procedure, I extended the study to some other types of two-verb clusters that De Sutter (2009) did not cover, such as clusters with modal verbs, as well as to two distinct domains of text: Wikipedia text and European parliament proceedings. Using a manual approach, this would entail extracting, checking and possibly annotating a large amount of additional data, while with an automatic approach, this is a simple matter. Studying additional constructions only requires modifying the corpus search query, and studying different domains of text only requires using the same query to search a different corpus or subcorpus that has been automatically annotated in the same format. By doing this, I found that De Sutter’s (2009) factors also hold for verb clusters with modal verbs, although variation is more rare there. The same was found for verb clusters in main clauses, whereas De Sutter only studied those in subordinate clauses. Modeling verb clusters in the two domains of texts showed that, while the 1-2 order is used more in the European parliament texts, the same factors account for the variation in both domains.

While this study showed that useful results could be obtained from an automatically annotated corpus, this finding is still a rather indirect form of evaluation, and requires conducting an entire study. Generally, it would be more useful to know about the quality of the automatic annotation in advance, before
starting a study. One evaluation measure that could be used for this purpose is parsing accuracy, which measures the percentage of parsing steps or syntactic dependencies that the parser annotates correctly. However, parsing accuracy is computed for an entire corpus, while a linguist wanting to use the corpus is typically only interested in a particular (type of) construction, such as verb clusters. A parser might not be as good at parsing verb clusters as it is at connecting adjectives to nouns, for example. But adjective-noun dependencies are more frequent than verb clusters, and also easier, so they will have a greater impact on the overall accuracy score of the parser. Therefore, this score is unlikely to tell us much about how well the parser has annotated verb clusters in the corpus. In chapter 3, I therefore address the following question:

Are there ways in which linguists can estimate the reliability of an automatically annotated corpus for the study of a particular construction, such as verb clusters?

The comparison in this chapter showed that different aspects of construction-specific annotation accuracy, such as precision and recall, can be estimated using four different approaches. The most difficult measure to evaluate is recall — in a large automatically annotated corpus, it is difficult to know whether any instances of the construction have been missed. Searching for particular instances of a construction using string-based search can be used to calculate recall over a portion of the data, determining how many examples might have been missed. As for precision, manual evaluation of the results of a corpus query for a particular construction can be used to determine the precision of the query’s results. Another method I explored is falling back to simpler annotation, such as part-of-speech tags in a treebank, as these are likely to be more accurate. This method can be used as a verification of the syntactic annotation of the corpus, even over larger amounts of data, and provide a rough estimate of recall. When combined, these methods allow linguists to get a clearer impression of the quality of the annotation of the particular construction they are investigating in the corpus, while still preserving the advantage of being able to obtain many exemplars with relatively little manual effort.

The large number of verb clusters available in automatically annotated corpora allowed me to further explore an issue that has not received much attention in the literature so far: main verbs and word order in verb clusters. The problem with main verbs is that there are many of them. Any Dutch verb can be the main verb of a verb cluster. If a particular main verb in a verb cluster only occurs once in every 2000 verb clusters, a large corpus is necessary to obtain enough instances of this type of verb cluster. In his manual corpus study, De Sutter (2005, 242-248) devotes a subsection to this issue, but does not mention how many instances he has found of each main verb in a verb cluster. As his dataset consisted of 2390 verb clusters, there were not likely to be many main verbs occurring in more than five clusters. Using the automatically annotated corpus, we were able to extract 411,623 verb clusters in chapter 2. This dataset allowed
me to address this issue in more detail, so in chapter 6 I studied the question:

Is there any evidence that main verbs, or the semantics thereof, affect verb cluster word order choice?

In the study, I found significant associations between main verbs and particular verb cluster word orders, indicating that the two verb cluster word orders have distinct lexical and semantic preferences. A majority of the main verbs are significantly associated with either the 1-2 or 2-1 word order in the corpus, even after controlling for various confounding factors. In De Sutter’s (2005) results with a smaller corpus, this was only a small percentage (less than 5%), though the results are not comparable because De Sutter did not use a frequency cutoff. In his study, all verbs occurring in a verb cluster were included, while we only included the ones that occurred in a verb cluster more than fifty times.

Furthermore, the two word orders had distinct semantic preferences: the 2-1 order took more main verbs with negative semantics and cognition verbs. No difference in dynamic verbs was found, even though this was hypothesized by Pardoen (1991). These findings suggest that two word orders are distinct constructions, for which these preferences are somehow stored in the mental lexicon.

With the large amount of data that is available through automatically annotated corpora, it would be interesting to study more fine-grained semantic distinctions than the verb classes mentioned above, in order to discover the semantics of the 1-2 or 2-1 order construction, if there is any. Using distributional semantic models, which model the meaning of words on the basis of the contexts in which they are used, would make this possible. It would also be possible to quantify the semantic difference between the two orders. This would be an interesting direction for future work, as the question of the meaning or interpretation difference between the two orders still does not have a clear answer.

The corpus-based research conducted in this dissertation may also contribute to bridging the gap that has been identified to exist between corpus linguistics on the one hand, and theoretical cognitive linguistics and psycholinguistics on the other hand. Gries (2010) notes that the two sides share a lot of similar theoretical concepts, but use different terms for them. In this dissertation, I have aimed to use terminology from both fields and addressed several of the commonalities between the fields that Gries (2010) mentions by using corpus methods to address cognitive theoretical concepts. Specifically: the relation between token frequencies and entrenchment came up in chapter 4, the connection between n-grams and psycholinguistic measures of processing was central in chapter 5, the connection between syntax and lexis was relevant in chapter 6, and I talked about about words, patterns and linguistic unit status (in chapters 4 and 6).
10.1.2 Processing

During the course of writing this dissertation, the factor of processing emerged as both an understudied aspect of verb cluster order variation and an interesting way to account for factors that have been associated with verb cluster word order variation in the literature. Minimizing processing effort has been proposed as an explanation for other cases of grammatical variation in the literature, and De Sutter (2005) suggests it as a potential explanation for verb cluster order variation, requiring further study.

The general idea is that speakers prefer to minimize their use of cognitive resources, formulating sentences in a way that minimizes processing complexity when multiple ways to express something are available, such as the two orders in the case of two-verb clusters. Processing complexity refers to the amount of cognitive effort required to produce or comprehend an utterance. To decide whether this is the case for verb cluster orders, we need to know what kinds of contexts or features of verb clusters are more difficult to process, and see whether one of the two orders occurs significantly more often in these complex types of context. For this, we need a theory that tells us what is difficult to process and what is easier to process — in other words, a theory of human language processing. I have addressed these issues of processing in chapters 4 and 5, and they are summarized by the following research question:

Do either of the orders occur in contexts that are generally more difficult to process, and which order is more difficult to process?

In chapter 4, I show that minimizing processing complexity is indeed a plausible explanation: many factors that correlate with a particular verb cluster order can be explained in terms of processing, on the basis of psycholinguistic literature. Furthermore, this association is largely consistent: for most factors, it is the case that the more difficult condition is associated with the 1-2 order. I therefore propose that the 1-2 order is the default, easier order.

However, the notion of complexity used in this chapter is not based on a particular theory. Rather, it is based on various empirical measures of complexity from psycholinguistic experiments. To establish the complexity of a particular construction, I surveyed the psycholinguistic literature for studies, such as reaction time experiments, that discuss the processing difficulty of that construction or its relevant features. As also noted by De Sutter (2009, p. 226–227), there is no single theory of human language processing with the aim of explaining variation within constituents, which is a problem for verb clusters, as they are situated entirely within the verb phrase. Nevertheless, it would be good if we could base an account of processing in verb clusters on a single theory of processing, as it would provide a consistent and well-defined basis for this factor of complexity. This is the issue I addressed in chapter 5.

In chapter 5, I aimed to implement a particular theory of human language processing to serve as a measure of processing complexity: surprisal theory.
This theory is based around the idea that more surprising elements are more difficult to process. I took the surprisal (perplexity per word) of a simple trigram language model on verb clusters from the Lassy Large corpus to represent the processing complexity of those verb clusters, though any type of probabilistic language model could have been used here. The results showed that perplexity-per-word as a measure of surprisal does not predict word order variation in two-verb clusters, but the difference in surprisal between verbs within the verb cluster construction does. This measure of uniform information density accounts for a large portion of verb cluster order variation when added to the model from chapter 4. Part of the variation that the empirical measures of processing complexity from chapter 4 account for, is also accounted for by the uniform information density measure, but not all of it.

According to the measure, verb clusters in the 1-2 order generally have a more uniform information density than 2-1 order clusters. Under the assumption made by Levy & Jaeger (2007) that uniform information density facilitates processing, this supports the processing hypothesis for Dutch verb cluster order and the hypothesis that the 1-2 order is the default order, from chapter 4.

However, this finding raises many further questions. There are indications that processing also plays a role in other word order phenomena around the second verbal position. As an example, the case of prepositional phrase extraposition has been linked to various processing-related factors as well, such as the length of the extraposed element and morphological complexity (Jansen, 1978; Canegem-Ardijns, 2006), and a multifactorial corpus study confirming such results has been conducted (Willems & De Sutter, 2015). If uniform information density is processed more efficiently, perhaps these variation phenomena interact to produce word orders in which information density is more uniformly distributed. In chapter 4 I already noted an effect of the position of extraposable PPs on verb cluster order. The literature often assumes that such balancing takes place, but that has not been investigated empirically.

In chapter 5 I chose to measure surprisal using a very standard type of language model. However, it would be interesting to explore more cognitively plausible language models in future work. It might be the case that a measure of surprisal that takes more structure or syntax into account would be more predictive of verb cluster order variation or other alternations. A delexicalized n-gram model could be used to make the measures we used more sensitive to structure. A language model that is more similar to human language processing would allow us to measure complexity better, though of course there is the problem that we do not know exactly how human language processing works. For this reason, it would also be good to compare the predictions of language models to the results of psycholinguistic experiments on a particular construction.

The hypothesis from chapter 4, that the 1-2 order is the default order, is testable psycholinguistically or otherwise experimentally. A reaction time study such as a self-paced reading task could reveal whether the 1-2 order is processed faster than the 2-1 order by comparing reading times of both orders with the same verbs. Such an experiment could also be used to evaluate different language
models as measures of complexity: language models whose surprisal more closely correlate with slower reading times on various verb clusters in various contexts are better predictors of complexity. The default 1-2 order hypothesis could even be tested without measuring reaction times: one could elicit verb clusters with a sentence completion task in more complex and less complex contexts, testing whether 1-2 orders are produced relatively more often in more complex contexts.

10.1.3 Variation affected by change

Two chapters addressed the link between variation and change in verbal clusters. While much has been written about verb clusters in Middle Dutch and Early-Modern Dutch, there are still many open questions at the extreme ends of the timeline — the change from Proto-West-Germanic to what we find in mediaeval corpora, and present-day Dutch. On the first topic, I addressed the following question in chapter 7:

In English, verbal word order converged on the 1-2 order, while in German, the 2-1 order took over in all contexts. In Proto-West-Germanic, there appears to have been variation. Why did the orders diverge?

In this chapter, I aimed to create the simplest possible agent-based model that could still account for this divergence, on the basis of historical frequency data. The agents in the model are equipped only with a bias for decomposing an utterance into individual features and storing it that way, producing utterances on the basis of feature frequencies. The model produces the ‘English’ outcome (1-2 orders only) when the feature for *to have* as an auxiliary verb is introduced faster, representing the historical process of the grammaticalization of this verb. It produces the ‘German’ outcome (2-1 orders only) when the usage of subordinate clauses is increased faster, which was also a historical process, due to the use of 2-1 orders in Proto-Germanic subordinate clauses. The model therefore gives us two hypotheses: that *to have* grammaticalized faster in English, and that subordinate clauses gained use more quickly in German. These predictions can now be tested using historical corpora of these languages in future work.

The model discussed in this chapter cannot yet account for the current state of the Dutch language, which first moved towards mainly 2-1 orders like German, and then shifted towards 1-2 orders again (Coussé, 2008), a change that is still in progress, according to the findings in chapter 9. As discussed in the previous section, there is evidence that the 1-2 order has become the default order, and this second change was likely caused by a factor outside the scope of the model, such as language contact.

On the second topic, present-day Dutch, the following question was addressed in chapter 9:

Do younger speakers use more 1-2 orders than older speakers in their spoken Dutch?
This question was investigated by means of an apparent time study in the Corpus of Spoken Dutch: by comparing the verb cluster word orders of younger and older speakers in this corpus, under the assumption that speakers’ preferences have not changed drastically throughout their lifetime, we can draw conclusions about present-day language change. Previous work already established that there has been a shift from the 2-1 order to the 1-2 order in Early-Modern Dutch (Coussé, 2008; Coupé, 2015), and Coussé (2008, p 190) hypothesized that this change is ongoing. Our results show that this is indeed the case: younger speakers in the corpus use more 1-2 orders than older speakers. In the youngest group it is even the case that the majority of verb clusters in the study is in the 1-2 order, although this is still a lower proportion of 1-2 orders than what is typically found in written Dutch.

Comparing these observations from spoken data to written data, as has been done before by Stroop (2009, p. 461) a.o., and considering this conclusion, does leave us with an interesting question. Given that there is an ongoing change in the direction of the 1-2 order, why has this change progressed further in written language than in spoken language? Typically, written language is more conservative than spoken language. What is special about verb clusters for them to exhibit this phenomenon? Perhaps this generalization does not apply to certain kinds of change, such as reordering within a constituent? If so, why? A study of similar phenomena of language change might reveal some insights on the nature of language change here.

The findings from these two chapters can also be interpreted in terms of the processing hypothesis discussed in the previous section. If the 1-2 order is easier to process, one would indeed expect an ongoing language change in the direction of the 1-2 order, as observed in chapter 9. After all, if it is so much easier, why would one still need the 2-1 order in any context? However, the model from chapter 7 also suggests a possible mechanism for a change in the other direction: changes in frequency of other, related constructions that are associated with the 2-1 order.

10.1.4 Language contact

Many instances of language change and variation have been driven by language contact, but for verb cluster constructions, not much has been written on this topic. Yet there is a very interesting recent case study available: West Frisian, which has an extensive history as a lesser used language, having been in contact with various culturally dominant languages, most recently Dutch. Furthermore, verb cluster word orders have started changing quite recently in this language. Hence, I have addressed the following question in chapter 8:

Is the modern use of the 1-2 word order in Frisian actually a new development taken from Dutch, a consequence of hundreds of years of language contact, or a continuation of an older language-internal development? And what kinds
of language contact between Dutch and Frisian may have affected this change?

In this chapter, I looked for verb cluster order variation in a corpus of Early-Modern Frisian texts, identifying three possible sources for this variation: the variation that already existed in Proto-West-Germanic, variation introduced by contact through bilingualism, and variation introduced through learned borrowing. By studying the contexts in which the Early-Modern Frisian verb clusters in the 1-2 order occur, it became clear that the most plausible source is learned borrowing. This contrasts with present-day Frisian, where contact through bilingualism appears to be driving the change. There was little evidence of a processing effect similar to the one we found for modern Dutch in earlier chapters, in these Early-Modern Frisian texts.

10.2 Processing verb clusters

Throughout this dissertation, processing has emerged as a factor that can provide an account for a variety of facts about the usage of the 1-2 and 2-1 verb cluster orders. In chapter 4, it can account for usage patterns in present-day written Dutch, namely the association between particular orders and particular types of contexts that can be found in written corpus data, with the 1-2 order mostly being associated with more complex contexts. In chapter 5, observed differences in information density between the two verb cluster orders lead to an explanation in terms of surprisal theory, a theory of human language processing. In chapter 8, the usage differences between 1-2 order clusters in present-day Frisian and Early-Modern Frisian could be explained based on processing factors, helping to provide an account of why these non-normative 1-2 orders were affected by two different types of language contact. In chapter 9, the difference in processing complexity can be viewed as a motivation for the ongoing shift towards the 1-2 order in present-day Dutch, as evidenced by data from speakers of different ages from the Corpus of Spoken Dutch.

The idea of the 1-2 order being easier to process than the 2-1 order has some other implications that were not covered in this dissertation. If the 1-2 order is more economical than the 2-1 order, one would expect it to spread more easily in language contact situations. As we saw in chapter 9, this is indeed the case for present-day Frisian. English, which has historically undergone extensive contact that shaped Middle English, has the 1-2 order. But perhaps there are counterexamples and it would be interesting to perform a full survey of relevant contact situations. One interesting case to investigate would be *Cité Duits*, a hybrid German-Dutch-Limburgian contact variety in a Belgian Limburg coal miners’ district which appears to allow more types of PP extraposition and other elements in the postfield than both Dutch and German (Pecht, 2015). Yet, this variety was mainly influenced by languages and varieties in which the 2-1 order is more frequent, so it would be an interesting test of this contact hypothesis. It would also be possible to test the effect of language contact in
an agent-based model such as that from chapter 7, by introducing some ‘1-2-speaking’ agents into the agent population suddenly, and seeing if the 1-2 order can spread. The agent-based model in chapter 7 could not account for the shift from the 2-1 order to the 1-2 order in Early-Modern Dutch. This shift has been explained as change for the purpose of focus marking (Coussé, 2008) or as a consequence of semantic broadening and syntactic extension (Coupé, 2015), but perhaps it could also be analysed as the result of a contact situation between different language varieties, in which ease of processing caused the 1-2 order to have an advantage, as in present day Frisian.

A default 1-2 order also has implications for language acquisition, and has been used to explain the acquisition pathway of verb clusters in Dutch children (Meyer & Weerman, 2016). After acquiring the verb cluster construction, older children in Meyer & Weerman’s (2016) experiment have a stronger 1-2 order preference than adults in their spoken language. This could reflect the ongoing language change discussed in chapter 9, which is also one of the suggestions of Meyer & Weerman (2016). Alternatively, it could be explained in terms of processing: the more limited cognitive abilities of children lead to a stronger preference for the easier 1-2 order, which might decrease as the children get older. This is something that could be tested in an acquisition study of older L1 learners of Dutch than those that were tested by Meyer & Weerman (2016).

10.2.1 The limits of the processing account

Does this mean that verb cluster order variation is all about processing? Certainly not. There are other factors that contribute, such as regional variation, mode of communication, prosody, semantics, and others, as established in previous literature discussed in chapter 4 and elsewhere in this dissertation. And there need to be other factors, because if the ease of processing of the 1-2 order was the only factor that mattered, we would expect to see only 1-2 orders used everywhere throughout the Germanic language areas. This is not the case, for example in German, where the 2-1 order is used in two-verb clusters. How could that be accounted for?

Besides looking for explanations in terms of the factors listed above, chapter 7 suggested another possibility: the effect of other, related constructions. In the model from chapter 7, the German 2-1 order became dominant when subordinate clauses became more frequent, as subordinate clause clusters were already associated with the 2-1 order. Another related construction in German that could have an effect is verb-finality. German appears to be more strictly verb-final than Dutch: for example, prepositional objects are typically not placed after the verb in subordinate clauses, unlike in Dutch. Perhaps this stricter verb finality in German also discourages placement of the participle after the head verb of the cluster (i.e. the 1-2 order). Verb finality is acquired earlier than verb clusters by L1 children, and thus usage of the 1-2 order might be inhibited by this earlier verb-final construction.

Violating verb-finality is not completely ruled out, though: in larger German
Conclusion

verb clusters of more than two verbs, non-descending orders (containing verbs after the head verb) are judged as acceptable (Wurmbrand, 2004). It would be interesting to study whether processing complexity plays a role in these non-verb-final cases in German — this variation has been claimed to be free variation (Bader, 2015). The same can be said for Dutch verb clusters of more than two verbs. In this dissertation I have limited myself to two-verb clusters, but it is quite plausible that many of the arguments for processing effects in two-verb cluster order variation, also apply to order variation in larger verb clusters. The approach using large, automatically annotated corpora demonstrated in chapter 2 should yield enough examples even of these less frequent types of verb clusters.

In this dissertation, my investigation of the processing factor has been limited only to constructions that are actually used, i.e. verb cluster orders that are grammatical and appear in corpora. I discussed whether processing plays a role when both orders are available in the speakers’ grammar. However, a significant portion of the literature on verb clusters is mainly concerned with explaining why some orders are grammatical and others are not, typically in larger verb clusters. This is another domain in which language processing might play a role: According to functionalist theories of language, the way humans process language has shaped the grammars of natural languages. An example of this is the Performance-Grammar correspondence hypothesis of Hawkins (2014): “Grammars have conventionalized syntactic structures in proportion to their degrees of preference in performance, as evidenced by patterns of selection in corpora and by ease of processing in psycholinguistic experiments”. So, if some three-verb cluster orders are grammatical in a particular language variety and others are not, this could be a consequence of processing factors, according to this statement.

Comparisons of typological data on the grammaticality of various verb cluster orders in various language varieties to theories of human language processing, or even psycholinguistic evidence, could be used to study this topic. However, it is not at all clear whether this type of conventionalization that Hawkins (2014) mentions can take place for word order variants within a constituent, such as the VP of a verb cluster. Furthermore, many theories on processing do not make claims on the relative difficulties of processing a particular word order compared to another within a single constituent. In the literature on models of language processing, most attention goes out to larger structures and longer dependencies, not the processing of a single constituent. In her dissertation, Hendriks (2018) has compared various models of processing and found that none of them clearly predict differences in acceptability between the different logically possible three-verb cluster word orders, suggesting that processing cannot have shaped this aspect of grammar, unless combined with syntactic constraints. However, this comparison is based on acceptability judgement data which does not necessarily only reflect processing complexity (Wasow & Arnold, 2005; Pullum, 2007). Perhaps these results can be accounted for if other factors, such as prescriptive biases, are included in the analysis.
10.2. Processing verb clusters

Another possibility to account for the grammaticality and ungrammaticality of three-verb cluster orders is to search for an explanation in terms of effects of related constructions. Agent-based models of language change, such as the one employed in Chapter 7, could be used to model conditions that could lead to certain outcomes of grammatical and ungrammatical orders. Familiarity with common verb cluster bigrams (groups of two verbs) might affect the acceptability of three-verb clusters that contain these bigrams, for example.

The processing factor also cannot explain why the 1-2 order is more frequent in written language than in spoken language. As noted in the previous section, it is strange that the order that the language is changing towards, the 1-2 order, is used more in the written modality, a modality that is typically more conservative. Furthermore, De Sutter (2005) noted that spoken language is more immediate and therefore might be more difficult to process, so that is another reason why we might expect to see more of the ‘easier’ 1-2 order in spoken Dutch than in written Dutch, but the opposite is true. There must be a different factor involved here, such as stylistic aspects or the prescriptive tendencies in favour of the 1-2 order mentioned by De Sutter (2005), which are likely to be stronger in written and more formal language.

As this all shows, a hypothetical complete model of synchronic verb cluster word order variation will not only need to include language-internal factors such as processing complexity and related constructions, but also language-external factors such as the ones just mentioned. Such a model would have to be based on a variety of language data including spoken and written language of various genres and from various regions. When the right data can be found, it is certainly possible to construct this type of model: an example would be Levshina’s (2016) model of permissive constructions in English. In this study, internal factors (such as verb frequencies and distance between the verbs, representing complexity), external factors (such as mode and domain), and collostructional measures (as in chapter 6 of this dissertation, representing effects of related constructions) are included in a single Bayesian multinomial model.

10.2.2 Processing word order variation

Dutch verb clusters are certainly not unique in having relatively free word order without a clear meaning difference, nor is is the only word order alternation in which processing has been argued to play a role. Even if we just stay close to the Dutch verb cluster, Dutch has relatively free word order variation around its second verbal position or pole in multiple respects. Various word order variation phenomena occur near or in this position, such as verb cluster word order variation, verb cluster interruption, the position of separable verb particles and prepositional phrase extrapolation. While these instances of variation have each been studied, little is known about the way in which these options interact. It is sometimes assumed that variation phenomena interact, but I am not aware of a larger empirical study on this, involving multiple variation phenomena.

When speakers produce a verb cluster that is difficult to process, are they
more likely to use extraposition to balance this out and make the clause easier to process? Do speakers use the word order options to make the information density of their utterance more uniform, i.e. to distribute information evenly throughout the clause? Cognitively plausible language models are an ideally suited method to study such questions, as they model language in information-theoretic terms such as entropy and surprisal. Together with psycholinguistic evidence of processing ease or difficulty of the same constructions, more advanced language models could be developed that represent human language processing more accurately, i.e. they have problems with the same constructions that humans do, and make use of the same structures. While our language model from chapter 5 only used words, it is likely that other levels of structure should also be included in a cognitively plausible language model. Because of its well-studied word order optionalities around the verb cluster, Dutch is a particularly suitable test case on which such language models can be developed further. Furthermore, the language resources and infrastructure that are necessary for large-scale modeling are well-developed for the Dutch language, and corpora are easily available.
APPENDIX A

Queries

XPath queries used with PaQu in chapter 9, generated by GrETEL. For reasons of space, we have only printed four representative queries out of the twelve that were used. Results for the other verbs can be found by replacing the verbs in the queries. In the electronic version of this dissertation, the titles of the queries can be clicked to view the results in PaQu.

In the other chapters of this dissertation where the Lassy Large corpus was used, no full manual correction took place and therefore a different set of queries was used. Those queries are more accurate, but also more complex and too long to print here. Examples of such queries can be found in the online supplementary materials of Bloem (2020): https://github.com/bloemj/5verbclusters.

1. Verb clusters with hebben and a participle in the 1-2 order

```
//node[
  @cat="ssub" and
node[
  @rel="hd" and
  @pt="ww" and
  @lemma="hebben" and
  number(@begin) < ../node[
  @rel="vc" and
  @cat="ppart"
]/node[
  @rel="hd" and
  @pt="ww"
```
2. Verb clusters with hebben and a participle in the 2-1 order

```
//node[
    @cat="ssub" and
    node[
        @rel="vc" and
        @cat="ppart" and
        node[
            @rel="hd" and
            @pt="ww" and
            number(@begin) < ../../node[
                @rel="hd" and
                @pt="ww" and
                @lemma="hebben"
            ]/number(@begin)
        ]
    ]
] and
node[
    @rel="hd" and
    @pt="ww" and
    @lemma="hebben"
]
```

3. Verb clusters with kunnen and an infinitival main verb in the 1-2 order

```
//node[
    @cat="ssub" and
    node[
        @rel="hd" and
        @pt="ww" and
```
Verb clusters with *kunnen* and an infinitival main verb in the 2-1 order

```xml
_queries
@lemma="kunnen" and
number(@begin) < ../node[
    @rel="vc" and
    @cat="inf"
] /node[
    @rel="hd" and
    @pt="ww"
] /number(@begin)
] and
node[
    @rel="vc" and
    @cat="inf" and
    node[
        @rel="hd" and
        @pt="ww"
    ]
]
]

4. Verb clusters with *kunnen* and an infinitival main verb in the 2-1 order

```xml
//node[
    @cat="ssub" and
    node[
        @rel="vc" and
        @cat="inf" and
        node[
            @rel="hd" and
            @pt="ww" and
            number(@begin) < ../../node[
                @rel="hd" and
                @pt="ww" and
                @lemma="kunnen"
            ]/number(@begin)
        ]
    ] and
    node[
        @rel="hd" and
        @pt="ww" and
        @lemma="kunnen"
    ]
]
```


Bibliography


Bor, Arie (1971). Word-groups in the language of the Skeltana Riucht: a syntactic analysis with occasional lexicological observations; followed by an inquiry into its punctuation and the possibility of the influence on the text of spoken language. H. Veenman & Zonen.


Bibliography


Jurafsky, Dan (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production.


Kilpatrick, Paul & Hellwig, Birgit (2002). *Corpus Gesproken Nederlands (COREX)*. CGN-CD.


Bibliography


Weerman, Fred, Olson, Mike, & Cloutier, Robert (2013). Synchronic variation and loss of case: Formal and informal language in a Dutch


Processing verb clusters

This dissertation describes the word order variation that exists in Dutch two-verb clusters, such as *gezien hebben* ‘seen have’ or *hebben gezien* ‘have seen’, and shows that the processing complexity of the utterance plays an important role in the choice between these two word orders. We can only process a limited amount of information at any point in time, and by varying the word order, we can spread this information throughout a sentence, making it easier to process an utterance. This may cause a person to say *gezien hebben* one time, and *hebben gezien* the next, while she means the same thing and did not suddenly start speaking a different dialect. In the title of this dissertation, ‘processing’ has a second meaning: that of language processing by computer systems. This is often called natural language processing, as a synonym of computational linguistics. Natural language processing allows for partly automatic analysis of larger amounts of linguistic data, making it possible to collect more evidence for linguistic hypotheses, and provides the means to study new research questions.

What are verb clusters?

In Dutch, as in other Germanic languages, we often express grammatical properties such as tense, aspect and modality using auxiliary verbs. These are positioned together with the main verb in a position that is called the second pole. This position is somewhere near the end of the sentence or clause, though sometimes followed by, for instance, a prepositional or indirect object. The only verb that does not join the others in the second pole is the finite verb in the main clause, which is placed on the first pole (usually the second position in the sentence), as in example A.1:
This system of auxiliary verbs permits the use of large clusters of verbs that form a chain of syntactic dependency, particularly in subordinate clauses. They are called verb clusters, shown in bold in examples A.2 and A.3 (two sentences from a collection of Wikipedia texts):

(A.2)  
Alle talen beschikken in zowel de gesproken als de geschreven vorm over een eigen woordenschat ofwel lexicon, en een regelsysteem dat grammatica of syntaxis genoemd wordt om alle elementen uit de woordenschat tot welgevormde zinnen te verenigen. All languages possess in both the spoken and the written form a own vocabulary or lexicon, and a rule system that grammar or syntax called is to all elements from the vocabulary to well-formed sentences to combine.

[A.D.1]  
Ik heb het gezien.  
I have it seen.  
I have seen it.

These clusters can sometimes consist of quite a few verbs:

(A.4)  
Vandaar dat er al langer een debat bezig is over de Therefore that there already longer a debate ongoing is about the
That’s why there has been an ongoing debate on whether it should be possible to occasionally violate that sovereignty through ‘humanitarian interventions’.

Dutch verb clusters show a striking amount of word order variation. In two-verb clusters, both of the logically possible orders can be used: the auxiliary verb can come after the main verb (as in example A.2) or before (as in example A.3), especially when the main verb is a participle. When the auxiliary verb comes first, this is called the 1-2 order, because the syntactically highest verb (denoted with 1) comes first. The term red order is also used in the Dutch literature. Conversely, example A.2 is the 2-1 order, or the green order. The red and green colour terminology came into use after Pauwels (1953) used these two colours to represent the two orders on the dialect maps she made, based on a large survey of their use in the Dutch language area.1 This study showed extensive regional variation: in the north of the Netherlands, there was a strong 2-1 order preference, while in the west, the 1-2 order was more frequently used in the responses.

Apart from this regional variation, many more factors have turned out to affect the occurrence of these two word orders. In a review article, Coussé et al. (2008) note the role of the manner of communication (spoken or written, monologue or dialogue) as a factor, as well as the sentential rhythm, the semantics of the auxiliary verb, as well as priming, which can cause the order of a previously perceived or produced verb cluster to influence that of the next produced one. Despite these findings, it is not yet clear why such a wide variety of factors influences the word order of verbs in the Dutch verbal end group.

This dissertation consists of eight articles on verb cluster order variation, with a focus on new, computational research methods that can be applied to the problem. Chapters two, three and six mainly concern the use of these new digital tools for linguistic research, while the chapters four, five, seven, eight and nine are mainly on explaining verb cluster variation.

Linguistic research using computational methods

In theory, linguists’ lives are becoming easier due to the increasing availability of digital texts that can serve as data for linguistic research, at least for widely used languages. The internet is full of language data and its volume is increasing at a faster rate than any linguist could keep up with. Yet this data is only used

1Available online at the Digital Library of Dutch Literature (DBNL): https://www.dbnl.org/tekst/pauw022plaa01_01/pauw022plaa01_01_0034.php
Summary to a limited extent, because it is not always clear how we can find, among this
sea of data, the utterances and constructions that are of particular interest to us linguists. When we do try to do this, it results in such a large amount of
data that any further processing will have to be automatic, using models that
can generalize over large volumes of data.

Automatic annotation software (e.g. parsers) can be used to make this data
searchable in terms of its linguistic properties, but the use of such software can also raise doubts because automatic annotation inevitably contains errors, more
so than manual expert annotation by a linguist. My research, however, shows
that large, automatically annotated text collections can also make it possible
to conduct studies that could not have been carried out in any other way. A
Dutch-language example of such a text collection is the Lassy Large-corpus
(van Noord et al., 2013), which contains about 700 million words and has been
syntactically annotated using the Alpino-parser (van Noord et al., 2006). The
example sentences at the start of this summary were taken from this corpus.

In chapter two, we start by showing that we can use the data from this
corpus to replicate a previous study on a larger scale, and that we can draw
the same conclusions as from that study, which was based on manual analysis
by a linguist instead of automatic annotation and large-scale searching. This
automatic approach also allowed us to relatively easily extend the study to
different text types, and more accurately estimate the effect of various factors
on verb cluster order. Based on this result, I have also used the Lassy Large
corpus, and in particular its Wikipedia part, as a data source in all further
chapters of this dissertation that discuss present-day Dutch.

Chapter three discusses how errors made by automatic annotation software
can be checked: it is quite possible for a linguist to estimate the quality of the
automatic annotation by applying three evaluation methods that I discuss. By
manually checking a sample of the result of a query for the target construction,
and also searching for the same hits without making use of the automatic
annotation and checking the results of that, the linguist can compute an estimate
of the proportions of false positive and false negative results.

In chapter six, I discuss that these large text corpora open up the possibility
of addressing new research questions, such as the effect of the main verb on the
order of the verb cluster it is in. Every imaginable verb can be a main verb in
a verb cluster, so to find out whether there is an association between a specific
verb and a specific word order, huge amounts of text are required to find enough
cases. It is also difficult to exclude the effect of other factors besides the identity
of the verb, such as whether it is an abstract or concrete verb. Therefore, I
try out several different statistical methods that can be used to examine the
association between words and word orders, while excluding the effects of other
factors as much as possible. I do find evidence for such associations, and this
implies that linguistic theories should take into account the possibility that our
language system encodes specific associations between words and the structures
in which they are used.
Accounting for variation

This dissertation aims to further account for two-verb cluster order variation. It mainly focuses on factors that have been less thoroughly studied in previous work, but also on the question on why both of these orders are available in Dutch. When we know why we have these two word order options, despite the apparent lack of a semantic distinction between them, we may gain further insight into the language system that produced this variation, and into processes of language change.

Previous work has shown that there are many regional differences in the use of these two word orders (Pauwels, 1953), and a possible effect of the sentential rhythm on word order has been studied (De Schutter, 1996). There have been some dissertations that describe extensive research into the development of verb clusters and their historic word order variation in Dutch (Coussé, 2008; Arfs, 2007b), and whether the factor of style plays a role (De Sutter, 2005). An association between verb cluster order and sentence complexity has also been proposed (De Sutter, 2005).

In studies of similar cases of word order variation, in which information can be conveyed in different orders, the complexity of the sentence turned out to be a major factor in the choice between the possible variants. For instance, this has been studied for the choice between “I passed her the ball” and “I passed the ball to her” (Bresnan et al., 2007), or the choice between “I promise I will leave” and “I promise that I will leave” (Jaeger, 2010). It is interesting to hypothesize that this might also play a role in verb cluster order variation, which prompted me to research this further. Therefore, in chapter four we ask whether complexity might be the factor underlying the observation that such a variety of factors are associated with verb cluster order variation. In this study, we observe that there is a preference for placing the main verb last in contexts that are more complex. This indicates that complexity can lead to variation, and suggests that it is this 1-2-order that is the less complex one, and easier to process for our language system.

In chapter five, I investigate this further by discussing whether there is a specific theory of language processing that can account for the observed variation. To this end, I have applied surprisal theory to verb clusters in Dutch text. The study in this chapter showed that the two verb cluster order differ not so much in surprisal, but rather in the uniformity of their information density. This means that in the 1-2 order, the information conveyed is better spread across the two words. First, you get the relatively predictable auxiliary verb, such as to have, though it is in a position where there could also have been other constituents besides a verb. Next, you get the less predictable main verb, which could be any possible verb. However, because it is not the first verb of the sequence of verbs, you already know that it has to be a verb. In the reverse 2-1 order, almost all information is at the start of the cluster: both the fact that it is time for a verb cluster, as well as which of the many possible main verbs is used. This factor of information density turns out to statistically account for much of the order
Another variational aspect of verb clusters reveals itself when we take a broader view of the West-Germanic languages. In German, the 2-1 order is used exclusively in many regions, while in English, verb groups are only in the 1-2 order. In this, Dutch appears to be between German and English. In chapter 7, we go back to Proto-West Germanic to look for explanations for the observation that German, English and Dutch appear to have diverged in terms of verb cluster order. With the help of an agent-based model, we propose the hypothesis that the development of word order preferences depended on how related constructions developed in each language. Our model predicts that the German word order many have developed through increasing usage of subordinate clauses, while the English word order is predicted by an earlier grammaticalization of ‘to have’ as an auxiliary verb. The Dutch situation is more complex — as previously described by Coussé (2008) and Arfs (2007b), Dutch initially developed in the same way as German, only to later move towards the English order. Our model does not yet account for this, although we propose language contact as a possible factor that may have triggered the divergence from German.

The variation in Frisian verb clusters is another interesting case in West-Germanic, as this language has been significantly influenced by neighbouring languages throughout its history. While Frisian language courses still teach that the auxiliary verb should come last (the 2-1 order, as in German), it has become clear that younger speakers of Frisian use both orders, presumably because everyone is by now a Dutch-Frisian bilingual. Interestingly, the ‘Dutch’ 1-2 order can also be regularly found in older Frisian texts. In chapter eight, we study whether verb cluster word order variation in older Frisian texts might be caused by language contact, and what kind of language contact this might have been. We show that the 1-2 orders in these texts are used differently than those in Dutch, yet sometimes do occur in constructions that appear to be borrowed from Dutch. This indicates that these orders have been influenced by language contact, but not in the same way as the 1-2 orders found in present-day Frisian — the contact probably occurred when the writer was older, because the use of the ‘Dutch’ orders in the early-modern Frisian texts has properties of late acquisition. An example of such late acquisition of this verb cluster order may have been in learning how to write, a skill which may not have been taught in Frisian.

Lastly, there are some open questions concerning the variation found in present-day Dutch. We know about the historical changes, but are the word order preferences between the two orders still shifting? If so, we would also expect variation between present-day speakers — the word order preference of older speakers should reflect an earlier stage of the language change than that of younger speakers, who would have learned the latest verb cluster order preferences. If we can find this type of variation, we can make claims on whether this shift is still ongoing. The study in chapter nine shows that younger speakers in the Corpus Gesproken Nederlands (Corpus of Spoken Dutch) more frequently use the 1-2 order. This matches up with the historical trend towards the 1-2
order, which has been observed in data going back to the 16th century.

**Processing verb clusters**

Different chapters in this dissertation show that the factor of processing complexity can account for the use of different word orders in the Dutch verbal end group. In chapter four, usage patterns in present-day Dutch are explained in terms of processing complexity: the 1-2 order is used in contexts that are more complex. In chapter five, differences in information density between the two orders lead to an account of this variation based on surprisal theory, a specific theory of human language processing. In chapter eight, differences between verb clusters in the 1-2 order in present-day Frisian and early-modern Frisian are accounted for in terms of processing complexity, allowing us to distinguish two types of language contact that might have affected the use of these 1-2 orders in Frisian. In chapter nine, the difference in processing complexity between the two orders can be viewed as a possible reason for a shift towards the use of more 1-2 orders in present-day Dutch, as observed among speakers of different ages in the Corpus of Spoken Dutch.

Of course, these findings do not mean that Dutch verb cluster variation is exclusively due to processing complexity. Other, previously studied factors, such as regional differences, discourse type, prosody, semantics and others, contribute as well. After all, if this were not the case, there would be no reason for the ‘more complex’ 2-1 order to continue being used. The research described in chapter seven has shown that related constructions may also affect the use of a word order variant. In chapter ten, I therefore conclude that a complete account of verb cluster order variation will have to include all of these factors that enable, yet also constrain, the observed variation.

Because processing complexity also affects other types of variation, I additionally ask whether this line of research can be broadened. Are other types of word order variation, such as the position of prepositional objects, involved in distributing information throughout a clause? This possibility is discussed in chapter ten, and I conclude that Dutch is actually a good language in which to study this. In Dutch, there are various way to change up the word order, without affecting meaning too much, yet limited to specific constructions or constituents (it is not a free word order language). In addition, a large amount of Dutch digital text is available to use as data for this type of research. In future work, more theories of language processing can be expected to be tested on Dutch verb clusters.
Samenvatting

De verwerking van werkwoordsclusters

Dit proefschrift beschrijft de variatie in woordvolgorde binnen tweeledige werkwoordsclusters, zoals gezien hebben of hebben gezien, en laat zien dat de verwerkingscomplexiteit van de uiting een belangrijke rol speelt in de keuze tussen deze twee volgorden. We kunnen maar een beperkte hoeveelheid informatie tegelijk verwerken, en door met de woordvolgorde te variëren, kunnen we die informatie door de zin spreiden, waardoor het makkelijker wordt om een taaluiting te verwerken. Hierdoor kan het gebeuren dat dezelfde persoon het ene moment “gezien hebben” zegt, en het volgende moment “hebben gezien”, terwijl ze hetzelfde bedoelt en niet ineens een ander dialect is gaan spreken. In de titel van dit proefschrift heeft ‘verwerking’ nog een tweede betekenis: die van taalverwerking door de computer, wat vaak ‘natuurlijke taalverwerking’ (natural language processing) genoemd wordt, als synoniem van computerlinguïstiek. Natuurlijke taalverwerking maakt het mogelijk om grotere hoeveelheden taaldata deels automatisch te analyseren, waardoor we meer bewijs voor taalkundige hypothesen kunnen verzamelen, en in kunnen gaan op nieuwe onderzoeksvragen.

Wat zijn werkwoordsclusters?

In het Nederlands, net als in andere Germaanse talen, drukken wij grammaticale eigenschappen als tijd, aspect en modus vaak uit met behulp van hulpwerkwoorden, die dan in het Nederlands samen met het hoofdwoord geplaatst worden op een positie die we de tweede pool noemen. Dit is ergens aan het einde van de zin, bijzijn of ander zinsdeel, maar soms nog gevolgd door bijvoorbeeld een voorzetselvoorwerp of meewerkend voorwerp. Het enige werkwoord dat daar niet bij komt staan is het finiete werkwoord (de persoonsvorm) in de hoofdzin, die op de eerste pool (meestal de tweede positie in de zin) geplaatst wordt, zoals in voorbeeld A.1:
(A.1) Ik heb het gezien.

Dankzij dit systeem van hulpwerkwoorden kunnen, vooral in de bijzin, grote groepen werkwoorden gebruikt worden, die ook syntactisch van elkaar afhanke-
lijk zijn. Dit worden werkwoordsclusters genoemd, dikgedrukt te zien in voor-
beelden A.2 en A.3, twee zinnen uit een verzameling van Wikipedia-teksten:

(A.2) Alle talen beschikken in zowel de gesproken als de geschreven vorm
over een eigen woordenschat ofwel lexicon, en een regelsysteem dat
grammatica of syntax genoemd wordt om alle elementen uit de
woordenschat tot welgevormde zinnen te verenigen. [wik_part0008/1678-
17-1]

(A.3) De Chomskyaanse benadering t.a.v. syntaxis, die vaak transformationeel
generatieve grammatica (TGG) wordt genoemd, is een dominante
theorie onder taalkundigen maar een veelgehoord punt van kritiek betreft
de bijna exclusieve nadruk die bij het onderzoek gelegd wordt op het
Engels en enige andere Europese talen. [wik_part0032/15107-22-1]

Deze clusters kunnen een behoorlijk aantal werkwoorden bevatten:

(A.4) Vandaar dat er al langer een debat bezig is over de vraag of die sou-
vereiniteit soms niet gebroken zou moeten kunnen worden door
‘humanitaire interventies’. [WR-P-P-G_part00231/
WR-P-P-G-0000111732.p.17.s.7]

Opvallend is dat in het Nederlands verschillende woordvolgorden gebruikt
worden in deze werkwoordsclusters. Bij gebruik van twee werkwoorden kunnen,
vooral wanneer er sprake is van een participium als hoofdwerkwoord, beide
logische volgorden gebruikt worden: het hulpwerkwoord erna (zoals in voor-
beeld A.2) of het hulpwerkwoord ervoor (zoals in voorbeeld A.3). Wanneer het
hulpwerkwoord voorop geplaatst wordt, wordt dit de 1-2-volgorde genoemd,
omdat het syntactisch hoogste werkwoord, het hulpwerkwoord (aangeduid met
1) eerst komt. Ook wordt de term rode volgorde gebruikt. Voorbeeld A.2 is dan
de 2-1 volgorde, of de groene volgorde. De kleuren rood en groen zijn in gebruik
geraakt nadat Pauwels (1953) deze kleuren gebruikte om de volgorden aan te
duiden op dialectkaarten die ze maakte op basis van een groot onderzoek naar
het gebruik van deze volgorden in het Nederlandse taalgebied.1 In dit onderzoek
op basis van een enquête was veel regionale variatie te zien: in het noorden van
Nederland had de 2-1 volgorde sterk de voorkeur, terwijl respondenten uit de
Randstad vaker de 1-2 volgorde gebruikten in hun antwoorden.

Naast deze regionale variatie is gebleken dat nog veel meer factoren een
rol spelen bij de keuze tussen deze twee volgorden. In een overzicht noemen
Coussé et al. (2008) bijvoorbeeld de manier van communiceren (gesproken of
geschreven taal, een dialoog of een monoloog), het zinsritme, de betekenis van het

1Hier online te zien bij de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL):
https://www.dbnl.org/tekst/pauw022plaa01_01/pauw022plaa01_01_0034.php
hulpwerkwoord, en *priming*, waarbij de volgorde van een eerder waargenomen of geproduceerde werkwoordsgroep invloed kan hebben op de volgende. Waarom zoveel verschillende soorten factoren van invloed zijn op de woordvolgorde van werkwoorden in de werkwoordelijke eindgroep, is echter minder duidelijk.

 Dit proefschrift bestaat uit acht artikelen over variatie in werkwoordclusters, met daarbij aandacht voor nieuwe, digitale onderzoeksmethoden die daarbij ingezet kunnen worden. De hoofdstukken twee, drie en zes gaan voornamelijk over het gebruik van deze nieuwe digitale hulpmiddelen voor taalonderzoek, en de hoofdstukken vier, vijf, zeven, acht en negen gaan vooral over het verklaren van variatie in werkwoordclusters.

**Taalonderzoek met digitale hulpmiddelen**

In theorie wordt het leven van de taalkundige steeds makkelijker, omdat er voor de grotere talen steeds meer tekst digitaal beschikbaar is die als data voor taalkundig onderzoek gebruikt kan worden. Het internet staat vol met taalmateriaal en daar komt elke dag meer bij dan wat een taalkundige ooit zou kunnen doorspitten. Toch wordt hier maar beperkt gebruik van gemaakt, omdat niet altijd duidelijk is hoe we in deze zee van data de uitingen en constructies kunnen vinden waar we als taalkundigen geïnteresseerd in zijn. En als we dat wel doen, levert dit zo’n grote hoeveelheid data op dat verdere analyse ook automatisch moet gebeuren, met behulp van modellen die over data kunnen generaliseren.

 Automatische annotatiesoftware kan worden ingezet om die data doorzoekbaar te maken op haar talige eigenschappen, maar het gebruik hiervan roept ook wel eens twijfels op omdat automatische annotatie onvermijdelijk fouten bevat, meer dan bij handmatige annotatie door een ervaren taalkundige. Uit mijn onderzoek blijkt echter dat grote, automatisch geannoteerde tekstverzamelingen wel degelijk onderzoek mogelijk maken dat anders niet uitgevoerd had kunnen worden. Een Nederlands voorbeeld van zo’n tekstverzameling is het Lassy Groot-corpus (van Noord et al., 2013), dat zo’n 700 miljoen woorden bevat en syntactisch geannoteerd is met behulp van de Alpino-parser (van Noord et al., 2006). De voorbeeldzinnen aan het begin van deze samenvatting komen uit dit corpus.

 We beginnen in hoofdstuk twee door te laten zien dat we met deze data een eerder onderzoek op grotere schaal kunnen repliceren, en daaruit dezelfde conclusies kunnen trekken als uit deze studie, die over hetzelfde onderwerp gaat, maar dit op basis van handmatige analyse door een taalkundige onderzocht. Ook kunnen we met automatische analyse relatief eenvoudig het onderzoek uitbreiden naar verschillende soorten teksten, en het effect van bepaalde factoren op werkwoordclustersinvloed nauwkeuriger inschatten. Op basis van dit resultaat heb ik in alle verdere hoofdstukken van dit proefschrift die over het hedendaags Nederlands gaan, gebruik gemaakt van het Wikipedia-geeënte van het Lassy Groot-corpus als databron.
In hoofdstuk drie laat ik zien dat fouten die door automatische annotatiesoftware gemaakt worden, controleerbaar zijn: het is voor een taalkundige goed mogelijk om de kwaliteit van de annotatie in te schatten door drie evaluatiemethoden toe te passen. Door een deel van de zoekresultaten voor de doelconstructie handmatig te controleren, en daarnaast ook buiten de automatische annotatie om naar dezelfde resultaten te zoeken en het resultaat daarvan te beoordelen, kan een schatting van zowel het percentage foutpositieve als het percentage foutnegatieve resultaten berekend worden.

Ook blijkt in hoofdstuk zes dat deze grote dataverzamelingen het mogelijk maken om nieuwe onderzoeksvragen te beantwoorden, bijvoorbeeld over de invloed van het hoofdwerkwoord op de volgorde van het werkwoordscluster. Elk denkbare werkwoord kan als hoofdwerkwoord in een werkwoordscluster gebruikt worden, dus om uit te zoeken of er een verband is tussen een specifiek werkwoord en een specifieke volgorde en daar genoeg gevallen van te vinden, is heel veel tekst nodig. Ook is het lastig om de invloed van andere factoren dan het woord zelf, bijvoorbeeld of het om een abstract of concreet woord gaat, uit te sluiten. Ik test daarom verschillende statistische methoden waarmee het verband tussen woorden en woordvolgorden bekeken kan worden en waarbij de invloed van andere factoren zo veel mogelijk uitgesloten kan worden, en deze verbanden lijken er wel degelijk te zijn. Dit betekent ook dat in taalkundige theorieën rekening gehouden moet worden met de mogelijkheid dat ons taalsysteem specifieke verbanden bevat tussen woorden en de structuren waar ze in gebruikt worden.

Variatie verklaren
Het onderzoek in dit proefschrift heeft tot doel om deze werkwoordschustervariatie te verklaren, en gaat daarbij vooral in op factoren die in eerder onderzoek onderbelicht zijn gebleven, maar ook op de vraag waarom deze beide mogelijkheden beschikbaar zijn in het Nederlands. Als we weten waarom we deze twee volgordemogelijkheden hebben, ondanks dat het ogenschijnlijk qua betekenis weinig toevoegt, kan dit ons meer inzicht geven in ons taalsysteem waar deze variatie uit voortkomt, en in processen van taalverandering.

Uit voorgaand werk was al duidelijk dat er veel regionale verschillen zijn in het gebruik van deze twee volgorden (Pauwels, 1953), en ook was al onderzocht of er een verband is met het zinsritme (De Schutter, 1996). Daarnaast is in eerdere proefschriften al uitgebreid onderzoek gedaan naar de ontwikkeling van werkwoordsclusters en hun historische volgordevariatie in het Nederlands (Cousse, 2008; Arfs, 2007b), en of er een verband is met stijl (De Sutter, 2005). Ook is een mogelijk verband tussen werkwoordschustervolgorde en zinscomplexiteit voorgesteld (De Sutter, 2005).

Bij onderzoek naar vergelijkbare gevallen van variatie, waarbij informatie ook in meerdere volgorden gegeven kan worden, bleek de complexiteit van de zin een belangrijke rol te spelen in de keuze tussen de mogelijke varianten. Dit is bijvoorbeeld onderzocht voor de keuze tussen “ik gaf haar de bal” en “ik gaf de
Samenvatting

bal aan haar” (Bresnan et al., 2007), of de keuze tussen “ik beloof om weg te gaan” en “ik beloof weg te gaan” (Bouma, 2017b). Het is een interessante hypothese om te veronderstellen dat dit ook voor de werkwoordsculstervolgorden een rol zou kunnen spelen, hetgeen uitnodigde tot verder onderzoek. Daarom stellen wij ons in hoofdstuk vier de vraag of complexiteit misschien de achterliggende factor is voor de observatie dat er zoveel verschillende factoren verband houden met werkwoordsculstervolgorde. In dit onderzoek zien we dat het hoofdwerkwoord achteraan wordt geplaatst in complexere contexten. Dit geeft niet alleen aan dat complexiteit aanleiding kan geven tot variatie, maar het suggereert ook dat deze 1-2-volgorde degene is die minder complex is, en makkelijker te verwerken is voor ons taalsysteem.

In hoofdstuk vijf ga ik hier verder op in door te onderzoeken of er ook een specifieke theorie over taalverwerking is die de gevonden variatie kan verklaren. Ik heb hiervoor surprisal theory toegepast op werkwoordsculsters in Nederlandse tekst, en daaruit bleek dat er tussen de twee werkwoordsvolgorden niet zozeer een verschil was qua surprisal (onverwachtbeiheid), maar eerder qua uniformiteit van informatie. Dat betekent dat in de 1-2-volgorde, de informatie beter gespreid wordt over de twee woorden. Eerst komt het redelijk voorspelbare hulpwerkwoord, bijvoorbeeld hebben, maar wel op een positie waar ook iets anders dan een werkwoord had kunnen staan. Daarna komt het minder voorspelbare hoofdwerkwoord, wat elk mogelijk werkwoord kan zijn, maar omdat het niet het eerste werkwoord van de groep is, weten we al dat het een werkwoord moet zijn. Bij de omgekeerde volgorde komt alle informatie aan het begin: dat we aan de werkwoordsgroep toe zijn, en ook welk van de vele mogelijke hoofdwerkwoorden gebruikt wordt. Deze factor van informatiespreiding blijkt veel van de volgordevariatie die in teksten te vinden is, te verklaren.


De variatie in Friese werkwoordsculsters neemt ook een unieke positie in in dit plaatje, doordat deze taal door de tijd heen veel beïnvloed is door omliggende
Samenvatting

talen. Terwijl in de leerboeken van het Fries nog staat dat het hulpwerkwoord als laatste moet komen (de 2-1-volgorde, als in het Duits) is het duidelijk dat jongere sprekers van het Fries beide volgorden gebruiken, waarschijnlijk doordat iedereen inmiddels tweetalig Nederlands-Fries is. Interessant is echter dat de ‘Nederlandse’ 1-2 volgorde ook in oudere Friese teksten regelmatig te vinden is.

In hoofdstuk acht onderzoeken wij in hoeverre er sprake is van volgordevariatie veroorzaakt door taalcontact in historische Friese teksten, en wat voor soort taalcontact dit heeft kunnen zijn. Uit ons onderzoek blijkt dat de 1-2 volgorden in deze teksten op andere manieren worden gebruikt als in het Nederlands, maar soms wel voor komen in constructies die duidelijk uit het Nederlands geleend zijn. Dit duidt erop dat ze wel beïnvloed zijn door taalcontact, maar niet op dezelfde manier als nu — waarschijnlijk heeft het contact op latere leeftijd plaatsgevonden, want het gebruik van de Nederlandse volgorden in de vroeg-modern Friese teksten heeft eigenschappen van late verwerving. Dit kan bijvoorbeeld gebeurd zijn bij het leren schrijven, wat misschien niet in het Fries geleerd werd.

Ten slotte waren er ook nog vraagtekens rond de variatie in ons huidige Nederlandse taalgebruik. Is het gebruik van de twee volgorden nog steeds onderhevig aan verandering? Als dat zo is, dan verwachten we ook variatie tussen sprekers nu — de volgordevoorkeuren van oudere sprekers zullen een eerder stadium van de taalverandering weergeven, terwijl jongere sprekers de nieuwste volgordevoorkeuren geleerd zullen hebben. Als hier ook variatie te zien is, kunnen we iets zeggen over de vraag of dit aspect van de taal nog steeds verandert. Dit is wat het onderzoek in hoofdstuk negen beoogt, en daaruit blijkt dat jongere sprekers in het Corpus Gesproken Nederlands vaker de 1-2-volgorde gebruiken. Dit volgt de historische trend richting die 1-2-volgorde, die sinds de 16e eeuw te zien is.

Werkwoordclusters verwerken

Uit verschillende hoofdstukken in dit proefschrift blijkt dat verwerkingscomplexiteit een factor is die verklaringen kan bieden voor het gebruik van de verschillende woordvolgorden in de werkwoordelijke eindgroep. In hoofdstuk vier kunnen gebruikspatronen in het hedendaags Nederlands in termen van verwerkingscomplexiteit verklaard worden: de 1-2-volgorde wordt gebruikt in complexere contexten. In hoofdstuk vijf leiden verschillen in informatiespreiding tussen de twee volgorden tot een verklaring van deze volgordevariatie op basis van surprisal theory, een specifieke theorie over menselijke taalverwerking. In hoofdstuk acht worden verschillen tussen werkwoordclusters in de 1-2-volgorde in het hedendaags Fries en het vroegmoderne Fries verklaard met behulp van de factor verwerkingscomplexiteit, waardoor we twee soorten taalcontact kunnen onderscheiden die invloed gehad kunnen hebben op het gebruik van deze 1-2-volgorden in het Fries. In hoofdstuk negen kan het verschil in verwerkingscomplexiteit tussen de twee volgorden gezien worden als een reden voor een verschuiving richting meer gebruik van de 1-2-volgorde in het hedendaags Neder-
Samenvatting

lands, zoals geobserveerd bij sprekers van verschillende leeftijden in het Corpus Gesproken Nederlands.

De bevindingen uit dit proefschrift betekenen natuurlijk niet dat we onze variatie in werkwoordsclustervolgorde alleen maar te danken hebben aan verwerkingsfactoren. Andere, eerder onderzochte factoren, zoals regio, communicatiesoort, prosodie, betekenis, en andere, dragen ook bij. Als dit niet zo zou zijn, zou de ‘complexere’ 2-1-volgorde immers al lang in onbruik zijn geraakt. Het onderzoek in hoofdstuk zeven heeft laten zien dat de invloed van gerelateerde constructies ook bij kan dragen aan het gebruik van een variant. In hoofdstuk tien concludeer ik dan ook dat een volledige verklaring van werkwoordsclustervariantie rekening zal moeten houden met al deze factoren die de variatie mogelijk maken, maar ook beperken.

Omdat verwerkingscomplexiteit ook een rol speelt bij andere soorten variatie, rijst de vraag of dit onderzoek breder getrokken kan worden. Worden ook andere mogelijkheden tot volgordevariatie, zoals de positie van voorzetselvoorwerpen, ingezet om informatie over de zin te spreiden? Deze mogelijkheid wordt in hoofdstuk tien besproken, waar ik concluder dat het Nederlands bij uitstek geschikt is om deze vraag te onderzoeken. Er zijn in het Nederlands allerlei manieren om met de woordvolgorde te varieren, zonder dat hier al te veel betekenis van af hangt, maar wel beperkt tot bepaalde constructies (er is geen volledig vrije woordvolgorde). Bovendien zijn er grote hoeveelheden Nederlandstalige digitale tekst beschikbaar om dit soort onderzoek op uit te voeren. Het is dus te verwachten dat er in toekomstig werk meer theorieën over taalverwerking op werkwoordsclusters losgelaten zouden kunnen gaan worden.
Jelke Bloem was born on November 9, 1988 in Groningen. After graduating from the CS Vincent van Gogh in Assen in 2006, he started his bachelor in Information Science (Alfa-informatica) at the University of Groningen, with a minor in Neurolinguistics. After graduating in 2009, he successively began the Information Science masters programme, and a year later with the research master in Linguistics. During 2010, Jelke was a research assistant to George Welling, and in 2011 to John Nerbonne. He took courses at the University of Saarland in Germany in the 2011-2012 winter semester, and did an internship there with Manfred Pinkal. He graduated from the Information Science MA in 2013.

In May 2013, Jelke started the PhD project Learner modelling and language contact in syntax at the University of Amsterdam, which was jointly funded by the Amsterdam Center for Language and Communication and the Fryske Akademy. This project was supervised by Arjen Versloot and Fred Weerman and resulted in this dissertation. A research stay at the Quantitative Lexicology and Variational Linguistics group of the KU Leuven (2016) was part of this project. In addition, Jelke taught courses in the bachelor programmes of Dutch and Cognition, Language & Communication. Besides the written results bundled in this dissertation, he presented the results of this project in 21 cities in 14 countries.