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

Jelke Bloem
Amsterdam Center for Language and Communication
University of Amsterdam
Spuistraat 134, 1012 VB Amsterdam, Netherlands
j.bloem@uva.nl

Abstract

This work investigates the application of a measure of surprisal to modeling a grammatical variation phenomenon between near-synonymous constructions. We investigate a particular variation phenomenon, word order variation in Dutch two-verb clusters, where it has been established that word order choice is affected by processing cost. Several multifactorial corpus studies of Dutch verb clusters have used other measures of processing complexity to show that this factor affects word order choice. This previous work allows us to compare the surprisal measure, which is based on constraint satisfaction theories of language modeling, to those previously used measures, which are more directly linked to empirical observations of processing complexity. Our results show that surprisal does not predict the word order choice by itself, but is a significant predictor when used in a measure of uniform information density (UID). This lends support to the view that human language processing is facilitated not so much by predictable sequences of words but more by sequences of words in which information is spread evenly.

1 Introduction

According to functionalist theories of language, the way humans process language has shaped the grammars of 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 is other 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 the alternation between constructions where the particle 'up' is placed before or after the noun phrase:

(1) John picked *up* the book.

(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 multifactorial corpus studies are often related to language processing. Finding that construction (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. There are also more specific theories of language processing that are internally consistent and that have been used to account for a range of phenomena. While they may not cover all domains of linguistic complexity, they help to make the notion of processing effort more directly quantifiable. This means that they can be used as a single measure, that they can therefore be tested 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., in press). 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 section 2. Section 3 will address models of language processing and how language processing has been argued to affect grammatical variation in previous work. Section 4 describes our data, in section 5 we describe our language model, and in section 6 we present our results. The results are discussed in section 7.

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:

(3) Zij zei dat ze het **gelezen had**
She said that she it read had
'She said that she has read it.'

(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 to be 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., in press). Bloem et al. argue that the order in example (4), called the ascending order, is easier to process than the alternative order (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 (1) is easier to process. Additional evidence comes from the claim that the ascending order is also acquired earlier by children (Meyer and Weerman, 2016).

In Bloem et al.'s (in press) 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 used them.

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 and Tily, 2011). This human subjects definition of complexity has also been called 'relative complexity', in constrast 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.

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 and 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:

(5) The astronomer gazed [into the sky] [through his telescope].

(6) The astronomer gazed [through his telescope] [into the sky].

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 (Jurafsky, 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 and 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 and 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 to predict variation in other domains of language, such as discourse connective omission (Asr and Demberg, 2015). Therefore, this approach links probabilistic models of language that are typically used in natural language processing, to processing complexity.

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. (in press) 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 1. In this table, they are ranked by their information gain as measured in the stepwise regression procedure performed by Bloem et al. (in press). 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.

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

Table 1: List of factors in the Bloem et al. (in press) model of verb cluster order variation, ranked by information gain.

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 sentence (5th) is also discussed, where a longer middle field is argued to be more difficult to process due to longer dependencies. The factor EXTRAPOSITION (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 seperable verbs, such as *afwassen* '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. (in press).

All of the factors listed in Table 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.

4 Data

For reasons of comparison, we use the same corpus that was used by Bloem et al. (in press), 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. (in press), 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 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 and 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 more memory.

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

$$\begin{aligned}\hat{P}(w_n|w_{n-1}w_{n-2}) =& \lambda_1 P(w_n|w_{n-1}w_{n-2}) \\ &+ \lambda_2 P(w_n|w_{n-1}) \\ &+ \lambda_3 P(w_n) \qquad \text{where } P(w_n) = \frac{\delta + c(w_n)}{\delta|V| + c(N)}\end{aligned} \tag{1}$$

λ_1, λ_2 and λ_3 refer to the interpolation weights of trigram, bigram and (smoothed) unigram probabilities respectively. δ is the smoothing parameter. V refers to the vocabulary size (or, the number of types in the language model), N to all tokens, and $c()$ 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:

$$PP(W) = 2^{-log_2(P(V1)P(V2))} \tag{2}$$

$p(V1)$ and $p(V2)$ are the estimated probabilities of the two verbs in the cluster (as estimated in Equation 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:

$$PP(C) = 2^{-\frac{1}{2}log_2(P(V1)P(V2))} \tag{3}$$

The log probabilities of $P(V1)$ and $P(V2)$ 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.

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 (3) and (4), showing the two possible word orders. Example (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 (8) is in the **ascending** order, which is the opposite:

(7) Zij zei dat ze het **gelezen had**
She said that she it read had
'She said that she has read it.'

(8) Zij zei dat ze het **had gelezen**
She said that she it had read
'She said that she has read it.'

Word order	**Perplexity**
Ascending order clusters	1681.2
Descending order clusters	1675.8
Overall PPW	1679.8

Table 2: Perplexity per word (PPW) results for the two word orders, over all test-set clusters.

Table 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. (in press) — 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 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 increases by 0.00000000035 ($3.5\text{e}{-10}$). 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.

Condition	**Value**	**Perplexity**
Linear position	First verb	2264.9
	Second verb	1245.8
Verb type	Auxiliary verb in cluster	714.0
	Main verb in cluster	3952.0
Position and type	Auxiliary verb in descending cluster	178.2
	Main verb in descending cluster	15763.5
	Auxiliary verb in ascending cluster	2445.8
	Main verb in ascending cluster	1155.6
-	Overall PPW	1679.8

Table 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 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 shown in table 3. 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:

$$UID(C) = |log_2(P(V1)) - log_2(P(V2))| \tag{4}$$

Again, $P(V1)$ and $P(V2)$ are the estimated probabilities of the two verbs in the cluster (as estimated in Equation 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 4 is used, instead of that in Equation 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. (in press). 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 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 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 χ^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 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 complentary to them instead.

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 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 and 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 and 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.

References

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. *IWCS 2015*, page 118.

Emmon Bach, Colin Brown, and William Marslen-Wilson. 1986. Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 1(4):249–262.

Jelke Bloem, Arjen Versloot, and Fred Weerman. 2014. Applying automatically parsed corpora to the study of language variation. In Jan Hajic and Junichi Tsujii, editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1974–1984, Dublin, August. Dublin City University and Association for Computational Linguistics.

Jelke Bloem, Arjen Versloot, and Fred Weerman. in press. Verbal cluster order and processing complexity. *Language Sciences*.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer, and J. Zwarts, editors, *Cognitive foundations of interpretation*, pages 69–94. KNAW, Amsterdam.

Timothy Colleman. 2009. Verb disposition in argument structure alternations: a corpus study of the dative alternation in Dutch. *Language Sciences*, 31(5):593–611.

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

Gert De Sutter, Dirk Speelman, and Dirk Geeraerts. 2007. Luisteren schrijvers naar hun innerlijke stem? De invloed van ritmische factoren op de woordvolgorde in geschreven werkwoordelijke eindgroepen. *Neerlandistiek*, 2007.

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

Gert De Sutter. 2007. Naar een corpusgebaseerde, cognitief-functionele verklaring van de woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen. *Nederlandse Taalkunde*, 12(4):302–330.

Gert De Sutter. 2009. Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. In A Dufter, J Fleischer, and G Seiler, editors, *Describing and Modeling Variation in Grammar*, pages 225–255. Walter De Gruyter.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

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

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8. Association for Computational Linguistics.

John A Hawkins. 2004. *Efficiency and complexity in grammars*. Oxford University Press.

John A Hawkins. 2014. *Cross-linguistic variation and efficiency*. OUP Oxford.

T Florian Jaeger and Harry Tily. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.

Dan Jurafsky. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production.

Roger P Levy and TF Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Caitlin Meyer and Fred Weerman. 2016. Cracking the cluster: The acquisition of verb raising in Dutch. *Nederlandse Taalkunde*, 21(2):181–212.

Maarten van Gompel and Antal van den Bosch. 2016. Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software*, 4(1).

G.J.M. van Noord, P Mertens, C Fairon, A Dister, and P Watrin. 2006. At Last Parsing Is Now Operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. Leuven University Press.

G.J.M. van Noord. 2009. Huge parsed corpora in LASSY. In F. Van Eynde, A. Frank, K. De Smedt, and G. van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, volume 12, pages 115–126. LOT.

Thomas Wasow, T Florian Jaeger, and David Orr. 2011. Lexical variation in relativizer frequency. In Horst J. Simon and Heike Wiese, editors, *Expecting the unexpected: Exceptions in grammar*, pages 175–195. Walter de Gruyter, Berlin.

Daniel Wiechmann and Arne Lohmann. 2013. Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, 25(01):65–88.