Language and Cognition **10** (2018), 514–543. doi:10.1017/langcog.2018.13

Comparing explanations for the Complexity Principle: evidence from argument realization*

DIRK PIJPOPS
Research Foundation Flanders (FWO); Research Unit QLVL, University of Leuven

DIRK SPEELMAN
Research Unit QLVL, University of Leuven

STEFAN GRONDELAERS
Centre for Language Studies, Radboud University of Nijmegen

AND

FREEK VAN DE VELDE
Research Unit QLVL, University of Leuven

(*Received* 23 *April* 2018 – *Revised* 03 *August* 2018 – *Accepted* 08 *August* 2018)

ABSTRACT

The likelihood with which language users insert optional words or morphemes that explicitly mark syntactic structure tends to increase in complex grammatical environments. This positive correlation between explicitness and complexity, best known as the Complexity Principle, has been observed for a multitude of case studies in both naturally occurring language and experimental settings. Researchers have sought the explanation for this Complexity Principle in three different domains: cognitive comprehension processing, the language channel, and cognitive production processing. Based on these accounts, we formulate predictions regarding the action radius of the Complexity Principle in the alternation between a direct and prepositional object of the Dutch verb *zoeken* 'search'.

[*] We would like to thank the participants of the workshop *Sprachliche Kodierungs-Asymmetrien, Gebrauchsfrequenz und Informativität* [Lingual coding assymetries, usage frequency and informativity] at the 39th DGfS conference for valuable feedback on an early version of this work, Austin Frank for the use of some of his R code, Robbert De Troij for useful comments on a previous draft, and two anonymous reviewers whose remarks have greatly helped to improve this paper. The present study was supported by the Research Foundation Flanders (FWO) [grant number: 11ZZO16N]. Address for correspondence: e-mail: dirk.pijpops@kuleuven.be; dirk.speelman@kuleuven.be; s.grondelaers@let.ru.nl; freek.vandevelde@kuleuven.be

These predictions are tested against corpus observations. Our results confirm accounts according to which optional elements indicate production difficulties, as well as those that explain the Principle as a result of restrictions on the language channel. In addition, our results indicate that the Principle is sensitive to context-determined restrictions that are the result of its underlying cause. This may present a possible caveat for alternation studies.

KEYWORDS: complexity, language processing, argument realization, Uniform Information Density, corpus.

1. Introduction

A growing body of research is interested in the relation between processing complexity and grammatical explicitness (Ferreira & Dell, 2000; Haspelmath, 2008; Hawkins, 2002, 2004; Jaeger, 2006, 2010; Rohdenburg, 2016). This interest stems from two branches of linguistic inquiry: corpus-based alternation studies and psycholinguistics. On the one hand, corpus linguists turn to processing complexity as a possible explanation for the distributions they find in bodies of natural language use, and for the constraints they posit in probabilistic grammars (Bresnan, Cueni, Nikitina, & Baayen, 2007; Gries, 2002, 2003; Grondelaers, 2000; Grondelaers & Speelman, 2007; Shank, Plevoets, & Bogaert, 2016). On the other hand, psycholinguists are ex officio concerned with language processing and employ grammatical alternations as useful case studies to test processing hypotheses (Arnold, Wasow, Asudeh, & Alrenga, 2004; Ferreira & Hudson, 2011; Ferreira & Schotter, 2013). These two research traditions are increasingly converging, with corpus linguists asking questions on language processing (Grondelaers, Speelman, Drieghe, Brysbaert, & Geeraerts, 2009; Jaeger, 2011), and psycholinguists turning to corpus research as a methodology that is complementary to experimental work (Gennari & Macdonald, 2009; Roland, Elman, & Ferreira, 2006). The present investigation is of the first type, that is, a corpus-based alternation study primarily interested in the mechanisms causing the correlation between complexity and explicitness. This correlation is most famously expressed in Rohdenburg's Complexity Principle:

> In case of more or less explicit grammatical options, the more explicit one(s) will tend to be favored in cognitively more complex environments. (Rohdenburg, 1996, p. 151)

We then aim to answer two questions:

(i) What drives the correlation between complexity and explicitness as we find it in corpora?
(ii) Does the correlation hold in all linguistic contexts, and if not, in which ones?

Concerning the first question, the different explanations for the cause of the Complexity Principle can be divided into three viewpoints. The first viewpoint asserts that the Complexity Principle is chiefly caused by cognitive processing during language **production** (e.g., Ferreira & Dell, 2000; MacDonald, 2013). Explicit coding would present a convenient way to buy time for the language producer when processing demands are high, such as in complex linguistic environments. The second viewpoint states that the Complexity Principle is primarily the result of restrictions on the physical language **channel** (e.g., Fenk & Fenk-Oczlon, 1993; Fenk-Oczlon, 2001; Jaeger, 2010). These restrictions introduce noise into the language channel that may disrupt the flow of information, and as a result, additional coding is required to smooth out the peaks in information density that typically arise in complex environments. Finally, the third viewpoint proposes that the correlation emerges primarily due to cognitive **comprehension** processing (e.g., Bolinger, 1980; Clark & Murphy, 1982; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Hawkins, 2004). That is, explicit coding is first and foremost aimed at optimizing the addressee's comfort. More complex environments would then be coded using the more explicit grammatical option, because the explicit coding of the syntactic structure simplifies parsing. Our case study will allow us to differentiate between on the one hand the production and channel perspective and on the other the comprehension perspective, but not between the production and channel perspective.

Concerning the second question, we will specifically look at different word orders. We will argue that the various explanations for the correlation make different predictions about how the correlation behaves in particular word order contexts in our case study. In this way, answering the second research question can lead to an answer to the first question.

Most research on the topic has looked into the English *that*-alternation, as in (1), as a case study (a.o. Bolinger, 1972; Ferreira & Dell, 2000; Ferreira & Hudson, 2011; Ferreira & Schotter, 2013; Jaeger, 2005, 2010, 2011; Jaeger & Wasow, 2005; Roland et al., 2006; for an overview, see Shank et al., 2016, pp. 202–213). We will turn to a similar alternation in Dutch that has thus far not been looked at, namely the alternation between a direct object and prepositional object of the verb *zoeken* 'to search', as in (2). Just as the English conjunction *that* may be used to introduce a complement clause, the Dutch preposition *naar* 'to' may optionally introduce the theme of this verb (Haeseryn, Romijn, Geerts, de Rooij, & van den Toorn, 1997, p. 1168).[1]

[1] In this paper, we will use *theme* to refer to the participant that is searched for. This is only meant as a practical designation, and we do not mean to attribute any specific theoretical status to this term, either as a semantic role or similar concept. In the following example sentences in this paper, the optional preposition and the theme are underlined.

We thus consider the form with *naar* to be the explicit variant. Still, this alternation does differ from the English *that*-alternation in a number of important aspects, which will enable us to differentiate between the viewpoints.

(1) I would guess (that) Al Gore will not endorse anyone. (COCA, cited in Shank et al., 2016, p. 208)
(2) Men zoekt (naar) een alternatief. (WR-P-P-G-0000001757.p.1.s.5)
One searches (to) an alternative
'They are searching for an alternative.'

The following section discusses the contrasting viewpoints introduced above in further detail. The third section presents the employed case study, the corpus, our operationalization of complexity, and the composition of the dataset. The fourth section works out the predictions made by each viewpoint regarding our case study, and composes a mixed regression model to test these predictions. The final section summarizes the conclusions, discusses the relevance of the findings for several strands of research, and ends with a number of suggestions for further study.

2. Production, channel, or comprehension?

1.1. PRODUCTION

The most direct way in which complexity can affect explicitness is through cognitive production processing. Making sentence structure explicit by including the optional complementizer *that* or the preposition *naar* evidently requires some effort from the producer, but this effort would buy time for the producer to formulate a complex complement clause or noun phrase, thereby relieving pressure on production facilities (Ferreira & Dell, 2000, pp. 298–300). The primary cause of the correlation between complexity and explicitness would then be the cognitive effort of the producer. It is still possible that the comprehender also benefits from the use of explicit coding in complex contexts, but only in a derived or secondary way. Two production accounts that allow for this are the PDC-model (Production–Distribution–Comprehension) proposed in Gennari and Macdonald (2009), MacDonald (2013), and MacDonald and Thornton (2009), and the 'collateral signals' account (cf. Clark, 2004, pp. 373–381, as well as Brennan & Williams, 1995; Clark & Fox Tree, 2002; Collard, Corley, MacGregor, & David, 2008; Corley & Hartsuiker, 2003; Fox Tree & Clark, 1997; Smith & Clark, 1993, and references cited therein).

According to the PDC-model, pressures in production processing determine the distributions that we find in language use. In turn, these distributions shape an individual's grammar, and finally, this probabilistic grammar is employed in comprehension. This means that the comprehender will expect

the form of new sentences to confirm to this grammar, and thus to the form of previously heard sentences, whose realization was optimized for production. When a newly heard sentence then contradicts the comprehender's expectations by not being optimized for production, but rather for comprehension, this would – seemingly paradoxically – cause comprehension difficulties.

According to the collateral signals account, the use of optional markers informs the comprehender about the state of production. For example, production difficulties may be a cue to the comprehender that the following words are difficult to integrate in the existing context. The comprehender can then prepare for this by cancelling his or her expectations about upcoming material (Grondelaers et al., 2009, pp. 159–160).

1.2. CHANNEL

The channel perspective is rooted in Shannon Information Theory (Cover & Thomas, 1991; Shannon, 1948). It searches the root cause of the Complexity Principle not in any kind of cognitive processing by either producer or comprehender, but rather in the physical language channel between producer and comprehender (Fenk & Fenk-Oczlon, 1993; Fenk-Oczlon, 2001; Jaeger, 2010; Levy & Jaeger, 2007). As such, it is different from both the production and comprehension perspective.

This perspective states that human language use constitutes a form of information exchange, and that the language channel is a type of information channel. Like any kind of information channel, the language channel is prone to noise. This noise introduces the risk of information loss. The more information is packed into a signal, e.g., into a string of words, the more information will be lost if the signal is damaged by noise. In other words, the more dense the information flowing through a channel, the higher the risk of noise causing substantial information loss. Meanwhile, the less dense the information flowing through the channel, the less efficiently the channel is being used. As a result of these competing pressures, any information channel has an associated optimal level of information density that balances risk of information loss with efficiency of use. The users of a channel will attempt to approximate this level at all times, resulting in a more or less constant density of the information flow through the channel. This has been called the principle of Uniform Information Density (Jaeger, 2010).

The channel of natural language has been noted to be particularly prone to noise (Levinson, 2000, p. 28). For example, in the case of spoken language, background noises may cause some words to become unrecognizable to the comprehender. If the producer then chooses to express his message in as few words as possible, such noises may already cause too much information to be lost and may thus render the original message irretrievable. In the case of written

language, sources of noise include typos, imperfect eyesight, bad printing quality, and illegible handwriting. In the case of sign language, they may include sore muscles and visual clutter.

Optional markers that make syntactic structure explicit, such as English *that* or Dutch *naar*, may then present a way to tune the information density of an utterance. Such markers will be low in inherent information content, as they can apparently be added or removed without drastically altering the message expressed by the sentence. Additionally, they explicitly flag what follows as respectively a complement clause or a theme argument, hence rendering it more predictable. According to Information Theory, information equates with the negative logarithm of predictability. As such, these markers effectively reduce the information density of the following complement clause or theme argument. As a result, since complex elements tend to be high in information density and simple elements tend to be low, these markers would more often appear with complex elements. This then constitutes the correlation described by the Complexity Principle (Jaeger, 2010, pp. 26–28).

In this text, we present the channel-driven account separately from both the comprehension and production perspective for two reasons. First and foremost, it is fundamentally different from both the comprehension and production perspective in stating that the root cause of the Complexity Principle is not to be sought in any kind of cognitive processing, but rather in the physical limitations on the language channel. Second, if one would have to include it under either the production or comprehension perspective, it is not clear which one would be more appropriate. On the one hand, the channel-driven account pivots on successful communication. The question is whether the information contained in the message reaches the comprehender, and one could therefore include it under the comprehension perspective (cf. Jaeger, 2013). On the other hand, the noise in the language channel and therefore the cause of maximal information density stems for a large part, though not completely, from properties of the producer, namely the limitations of our physical articulators (Levinson, 2000, p. 28). Moreover, Ferreira and Schotter (2013, p. 1569) have argued for a strong affinity between the channel- and production-driven accounts, viewing them as merely "different levels of description of the same sort of phenomenon". According to this viewpoint, the production-driven account would be seen as the cognitive implementation of the principle of Uniform Information Density, which makes sure that language producers in practice always approximate the optimal level of information density.

1.3. COMPREHENSION

Finally, explicitly encoding the syntactic structure of a sentence evidently simplifies parsing and thus comprehension. In the case of *that*, the optional

marker signals to the comprehender that the producer is entering a complement clause. In the case of *naar*, the optional word is a fixed preposition with the verb *zoeken* 'to search' and it could be argued, according to the comprehension perspective, that it therefore expedites the linkage between the verb and its complement by flagging the following noun phrase as its complement with an explicit formal marker.

Still, the choice whether or not to use such optional elements of course rests with the producer, not the comprehender. There are then two ways in which comprehension processing can still affect this choice. The first is speaker's altruism or strong audience-design (Hawkins, 2002, 2004; Kirby, 1999, p. 60). This states that, if the producer is going to utter a complex phrase, s/he will choose the structure that is easiest to parse for the comprehender, even if this requires more effort from his/her part. Of course, the producer then needs to have some way of knowing which structure is easiest to parse, i.e., s/he needs to have access to some metric of parsing effort.

Note that this account of speaker's altruism is not a case of true altruism, as the producer may also indirectly benefit from forming easily comprehensible sentences. For one, comprehenders may be more inclined to listen to and act on the messages formulated by such producers. Moreover, communication is fundamentally a collaborative task, meaning that producers have to make at least some effort in order to be comprehensible (Zipf, 1949).[2] It then only seems a minor step to say that they also make the effort to use optional markers in order to be **easily** comprehensible.

The second way in which comprehension processing may affect choices in production is hearer selection (Kirby, 1999, pp. 31–62, see Ferreira & Schotter, 2013,p. 1568, for a similar proposal). This differs from speaker's altruism in that comprehension steers production in a more indirect way. It proposes that only constructions which lead to successful comprehension become entrenched in grammar, or that those which lead to more effortless comprehension become more strongly entrenched than those which require more effort. Once entrenched in grammar, these constructions can in turn affect the production of the language user in question. In other words, tendencies that obstruct comprehension processing are selected against in language evolution. While this account dispenses with the assumption that some metric of parsing effort is directly taken into account during production, it does require that entrenchment be dependent on successful or easy comprehension. This proposal can be seen as the reversal of the PDC-model from the production perspective. Figure 1 presents a comparison of the two.

[2] Unless, of course, when the producer is consciously trying not be comprehensible, in which case communication ceases to be a collaborative task. However, it is generally assumed that such situations present the exception, rather than the rule (Clark, 1996).

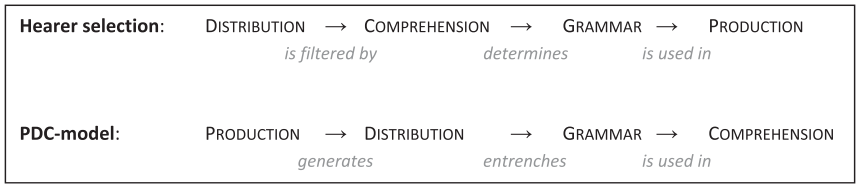

Fig. 1. Hearer selection versus the PDC-model.

NOTE: DISTRIBUTION here refers to natural language usage as we find it in corpora, and GRAMMAR stands for the cognitive organization of one's experience with language (Bybee, 2006, p. 711).

So far, empirical findings from experiments and corpora appear to favor the channel and production perspectives over the comprehension perspective. Ferreira and Dell (2000) find no evidence that language users employ explicitness to simplify comprehension in controlled experiments, while they do find evidence that lexical availability during production plays a role. Likewise, Elsness (1984) and Roland et al. (2006) find no indications that, in corpora, language users use the optional complementizer *that* to facilitate comprehension processing. Further indications from experiments and corpora in favor of the channel perspective are presented in Fenk-Oczlon (2001), Jaeger (2010), and Levy and Jaeger (2007); and in favor of the production perspective in Ferreira and Hudson (2011), Ferreira and Schotter (2013), Gennari and Macdonald (2009), Kraljic and Brennan (2005), MacDonald (2013), and MacDonald and Thornton (2009). For other studies investigating the differences between cognitive processing in language production and comprehension, see Bock, Irwin, and Davidson (2004), Tanner and Bulkes (2015), Tanner, Nicol, and Brehm (2014), and references cited therein.

3. Data

3.1. CASE STUDY AND CORPUS

The employed case study concerns the alternation between a direct and prepositional object of the Dutch verb *zoeken* 'to search'. The theme of this verb may be overtly marked by the preposition *naar*, and thus be realized as a prepositional object, as in (3), or this preposition may be dropped, and the theme realized as a simple direct object, as in (4). The reference grammar *Algemene Nederlandse Spraakkunst* explicitly states that the two variants of the alternation are synonymous (Haeseryn et al., 1997, p. 1168).

(3) Kee schildert en **zoekt** naar sponsors. (WR-P-P-G-0000011665.p.3.s.5)
Kee paints and searches to sponsors
'Kee paints and looks for sponsors.'

(4) Vzw De Scute **zoekt** daarom nog sponsors.
non-profit organization De Scute searches therefore still sponsors
(WR-P-P-G-0000350208.p.4.s.2)
'That's why non-profit organization De Scute is still looking for new sponsors.'

As the source of the data, we employ the SoNaR corpus of written Dutch (Oostdijk, Reynaert, Hoste, & Schuurman, 2013b), more specifically, the version that is syntactically annotated by the Alpino-parser (van Noord, 2006, see Bouma & Kloosterman, 2002, 2007, on how to best access XML-treebanks).[3] We have two main reasons for choosing a corpus of written language. First and foremost, the Complexity Principle has been observed not only in spoken language, but also in written language (a.o. Bouma, 2017; Rohdenburg, 1996, 2016; Shank et al., 2016). In this paper, we are primarily looking to explain why the Complexity Principle holds in written language. Still, we currently see no compelling reasons to assume that there are fundamentally different explanations for the Principle in spoken versus written language. In fact, there is ample research showing that findings based on written language are generally in accordance with findings from spoken language with regard to the relation between explicitness and complexity (Grondelaers, 2000; Grondelaers, Speelman, & Geeraerts, 2003; Jaeger, Levy, Wasow, & Orr, 2005; Jaeger & Wasow, 2005). Still, even if future research would reveal fundamental differences, it is not the case that written language processing is a priori less interesting than spoken language processing; this would simply limit the relevance of our research to the former.

Second, because we analyze observational data from corpora of natural language rather than experimental data from lab settings, we will need to deal with a number of confounds.[4] This means we will need sufficient datapoints to be able to control for these. The only way to acquire a sufficiently large dataset is to turn to a corpus of written text. In this choice for written data, we follow earlier studies on complexity, including Bloem, Versloot, and Weerman (2017), Gennari and Macdonald (2009), Gries (2002), Grondelaers et al. (2009), Jaeger (2011), MacDonald and Thornton (2009), Rohdenburg, (2016), Roland et al. (2006), and Willems and De Sutter (2015).

What does this choice for written data mean for the three perspectives introduced in the previous section? In general, the choice is favorable for the comprehension perspective. First, it benefits the comprehension perspective

[3] We have not made use of the material from text messages, tweets, chat logs, and discussion lists because the quality of the syntactic parses of these components was a priori deemed too low (Oostdijk, Reynaert, Hoste, & Schuurman, 2013a, pp. 49–50).

[4] The advantages of observational corpus data over experimental data mostly relate to ecological validity. See Jaeger (2010, p. 26) for a discussion.

in that we expect writers to bear in mind the ease with which their readers read their texts, at least to a greater extent or more explicitly than speakers would take into consideration the comprehension processing required from their hearers. As such, written language would be more prone to tendencies that reduce effort in comprehension processing.

Second, the choice for written data is disadvantageous to the production and channel perspectives. Regarding the production perspective, its reasoning primarily relates the spoken language. As such, we need the extra assumption that the (probabilistic) grammar of language users is first and foremost shaped by their experiences in spoken language, since this forms the majority of the linguistic input, and that this same grammar is then employed when processing written language. The correlation between complexity and explicitness in written language would then be a second-order effect, i.e., an effect that is retained even when its original cause is not directly present, because it has become entrenched in probabilistic grammar. Such second-order effects have also been demonstrated in morphology (Pijpops & Van de Velde, 2016, 2018). Regarding the channel perspective, the information channel of written language is probably less prone to noise than the channel of spoken language. Therefore, it would arguably be associated with a higher optimal level of information density.[5] As such, the channel would generate less pressure to use optional markers in complex environments in the case of written language than in the case of spoken language. Still, the channel of written language would still have some optimal level of information density. As such, the reasoning behind this perspective still holds. To sum up, the choice of written data results in a conservative research design regarding the production and channel perspectives.

3.2. OPERATIONALIZATION OF COMPLEXITY

There are at least two principal ways of contrasting the three perspectives using corpus data. The first is to formulate three separate operationalizations of complexity, each tailored to each perspective. For instance, one operationalization would be more suited for production complexity, while another operationalization would better measure information density, etc. (see Menn & Duffield, 2014, for a discussion on several operationalizations of complexity). Next, we could investigate which is the best correlate of explicitness, viz. in our case, the best correlate of the probability of the prepositional variant. However, these

[5] We say *arguably* because spoken language allows for many more multimodal ways of reinforcing the signal, e.g., through the use of gestures and facial expressions, which could in principle result in a higher optimal level of information density in spite of there being more noise (Cover & Thomas, 1991).

various operationalizations of complexity are likely to strongly correlate with one another, making it hard to disentangle them. We therefore opt for the second way, which is to employ a single operationalization of complexity that works for all perspectives. We can then compare the contexts in which it is correlated with explicitness.

As an operationalization of complexity, we use the variable THEME COMPLEXITY, counted as the natural logarithm of the number of words of the theme argument. While this may not constitute the most advanced operationalization of complexity, it is robust, reliable, largely independent of the employed parsing formalism, and it works for each perspective, as we will now argue.[6]

For the production and comprehension perspective, the choice for THEME COMPLEXITY is quite straightforward. Regarding the production perspective, the optional preposition *naar* always appears right in front of the theme argument, at exactly the moment when the producer needs to plan the theme. When the theme is long and the producer is hence under high processing pressure, this would be the most opportune moment to buy extra processing time. Regarding the comprehension perspective, having to parse a long noun phrase puts a large strain on cognitive comprehension facilities. As such, it would be most useful to have a formal marker right in front of this noun phrase that explicitly marks it as the theme argument.

The operationalization of THEME COMPLEXITY for the channel perspective requires some more clarification. It is based on the presumption that longer themes tend to be more specific than short themes. In turn, more specific themes are harder to predict, which means they contain more information. For instance, the theme in (5) is a lot more specific and hence contains more information than the theme in (6). As argued above, themes that contain more information have a greater need for a preceding preposition *naar* 'to' to reduce their information density.

(5) De provincie **zocht** naar een educatieve oplossing om toch
The county searched to an educational solution to still
enige greep te krijgen op minderjarige overtreders van het
some handle to get on underage transgressors of the
verkeersreglement.
traffic regulations

(WR-P-P-G-0000619732.p.4.s.2)

[6] Still, the operationalization is dependent on the question where to draw immediate constituent borders, but we assume that these are largely unproblematic, in that most current linguistic theories would by and large agree on them. In this, we follow the constituent borders assigned by the Alpino-parser, which is based on HPSG theory (van Noord, 2006; van Noord et al., 2013). As mentioned below, we only disagreed on the theme borders of 18 out of 1,000 randomly selected instances in the dataset.

'The country administration is searching for an educational solution to get some handle on underage transgressors of traffic regulations.'

(6) De provincie Gelderland **zoekt** een oplossing.
The county Gelderland searches a solution
(WR-P-P-G-0000037003.p.1.s.4)
'The country of Gelderland is searching for a solution.'

3.3. DATASET

All 79,410 instances of *zoeken* that appeared with a theme argument were extracted from the corpus and annotated with information based on the Alpino-parses. Not all of these data could be used, however.

First, the dataset still contained a number of instances of fixed collocations, viz. *zijn heil zoeken bij* 'flee to, turn to' as in (7) (959 instances), *ergens niets/niks te zoeken hebben* 'have no reason to be somewhere' as in (8)–(9) (728 instances) and *zijn toevlucht zoeken* 'seek refuge' as in (10) (957 instances). The meaning of these fixed collocations is non-compositional and they never appear in the variant with *naar*. As such, these 2,644 instances were excluded from the dataset.

(7) Velen **zochten** hun heil bij familie in Zuid-Servië en
many searched their salvation with family in South-Serbia and
Kosovo.
Kosovo
(WR-P-P-G-0000023743.p.4.s.4)
'Many fled to family in southern Serbia and Kosovo.'

(8) Zij hebben hier niets te **zoeken.**
They have here nothing to search
(WR-P-P-G-0000012841.p.2.s.6)
'They have no reason to be here.'

(9) Zo lang dit geen VN-operatie is, hebben wij daar niks te
So long this no UN_operation is have we there nothing to
zoeken.
search
(WR-P-P-G-0000017711.p.4.s.1)
'As long as this is no UN operation, we have no reason to be there.'

(10) Deze winkelier heeft inmiddels zijn toevlucht **gezocht** in
this shopkeeper has meanwhile his refuge searched in
het buitenland.
the outside_country
(WR-P-P-G-0000041098.p.3.s.3)
'Meanwhile, this shopkeeper has sought refuge abroad.'

Second, prepositional objects in Dutch enjoy greater liberties in positioning than direct objects. Dutch word order functions a lot like German word order, and is also characterized by a bipolar structure (i.e., the so-called *Klammernstruktur*; see König & Gast, 2009, Ch. 10; Zifonun, Hoffmann, & Strecker, 1997, p. 1498; Zwart, 2011, p. 26). This can be seen in Table 1. Bare noun phrases such as subjects or direct objects are grammatically limited to the prefield before the first verbal pole (1a) or the midfield between the poles (1b). They cannot grammatically be placed in the postfield, i.e., the position behind the second verbal pole (1c), the only exception being when they are realized as a subordinate clause. By contrast, prepositional phrases such as the prepositional object have access to the prefield, midfield, and postfield (2a–c). This means that when the prepositional object is placed in postfield position, the preposition *naar* cannot be dropped without overhauling the sentence structure. As such, these 6,454 instances were also removed from the dataset.

Finally, we will want to control for the country of origin, which can be either the Netherlands or Belgium. However, the country of origin was not known for 4,726 instances, which were therefore removed from the dataset. This left us with 65,586 observations. Table 2 shows how they are distributed among countries and corpus components.

From the final dataset, 1,000 instances were randomly selected and subjected to manual checking. Of these, we identified 18 cases in which we disagreed with the Alpino-parses on the exact demarcation of the theme, which was judged to be an acceptable level of noise. Earlier research has also shown that automatically generated datasets do not compromise the reliability of the results, while offering important advantages in reproducibility and scalability (cf. Bloem, 2016; Bloem, Versloot, & Weerman, 2014; Bouma, 2017; Theijssen, Boves, Halteren, & Oostdijk, 2010).

4. Hypotheses and analysis

In order to differentiate between the three perspectives, we will distinguish between those instances where the verb *zoeken* precedes the theme, as in (11), and those where the theme precedes the verb, as in (12). Instances where the initial part of the theme precedes the verb and the remainder follows it, as in (13), are counted amongst those where the theme precedes the verb, since the preposition *naar*, if it is present, would also precede the verb, as does the syntactic head of the theme.[7] We will now argue that the production and channel perspectives predict a negative correlation between the complexity of the theme

[7] For instances such as (17), the part of the theme following the verb is taken up in the calculation of THEME COMPLEXITY.

TABLE 1. *Various placement options for the theme when realized as direct or prepositional object in Dutch. Only when realized as a prepositional object can the theme be placed in postfield position*

		Prefield	1st pole	Midfield	2nd pole	Postfield
(1)	(a)	*Dat boek*	***heb***	*ik*	***gezocht***	∅
		that book	have	I	searched	∅
	(b)	*Ik*	***heb***	*dat boek*	***gezocht***	∅
	(c)	**Ik*	***heb***	∅	***gezocht***	*dat boek*
(2)	(a)	*Naar dat boek*	***heb***	*ik*	***gezocht***	∅
	(b)	*Ik*	***heb***	*naar dat boek*	***gezocht***	∅
	(c)	*Ik*	***heb***	∅	***gezocht***	*naar dat boek*

'I have looked for that book.'

TABLE 2. *Number of instances in our dataset from each country and each corpus component*

	Belgium	the Netherlands
autocues	3339	408
books	14	6163
brochures	119	12
e-magazines	1018	335
guides & manuals	1	10
legal texts	2	6
newspapers	23765	8832
periodicals & magazines	13171	1996
policy documents	21	5
printed newsletters	0	6
proceedings	18	1
reports	44	211
subtitles	5637	0
teletext pages	93	0
texts for the visually impaired	0	161
websites	116	72
written assignments	0	10

and the propensity for the explicit prepositional variant in cases such as (12), and positive correlation in cases such as (11). Meanwhile, the comprehension perspective will be argued to predict a positive correlation in both cases, and an even stronger positive correlation in cases such as (12) than in cases such as (11).

(11) We **zoeken** naar de oorzaak, maar hebben nog geen idee.
we search to the cause but have still no idea
(WR-P-P-G-0000039610.p.2.s.5)
'We are looking for the cause, but we have no idea so far.'

(12) We zijn dus wel gedwongen nu al naar een
We are thus PART forced now already to a
goede vervanger te **zoeken.**
good substitute to search
(WS-U-T-B-0000000070.p.13.s.3)
'We are thus forced to already look for a good substitute.'

(13) … als je naar een oplossing **zoekt** die perfect aansluit bij je
if you to a solution search that perfectly fits to your
bancaire behoeften.
banking needs
(WR-P-P-G-0000229626.p.13.s.1)
'… if you are looking for a solution that fits your banking needs perfectly.'

4.1. PRODUCTION HYPOTHESIS

The production perspective proposed that *naar* presents a way to buy time for the producer to formulate a complex theme. When the theme precedes the verb, however, this purchase comes at a serious cost. Only a handful of Dutch verbs combine with a prepositional object with *naar*. Using *naar* would therefore force the producer to already decide on which verb s/he is going to use. The planning scope of producers is limited, and the longer and more complex the upcoming theme argument, the less cognitive resources are available to simultaneously consider the choice of verb (see Gleitman, January, Nappa, & Trueswell, 2007; Konopka, 2012, and references cited therein). Meanwhile, if the producer chooses to realize the theme as a bare noun phrase, s/he can postpose the choice of verb until after the theme in completed. Moreover, if the producer has already decided on the future verb while building the upcoming complex theme, s/he will be forced to retain this verb in working memory until s/he has completed the formulation of the theme. Leaving this choice until later would allow him/her to free up this working memory.

An example with a complex preverbal theme is given in (14). When the producer includes *naar* in (14), his or her choice of verb will be limited to *zocht* 'searched' and perhaps *streefde* 'strove'. In other words, s/he would have to consider the choice of verb, exactly when facing the arduous task of planning the complex theme. Meanwhile, if the producer does not include *naar*, the choice of verb can be left for the future. In (14), reasonable options to finish the sentence would include *zocht* 'searched', but also *wilde volgen* 'wanted to follow', *probeerde te vinden* 'tried to find', *nastreefde* 'pursue', etc.

(14) De Wereldraad van Kerken heeft dat niet gedaan, omdat hij
The World Council of Churches has that not done, because he
van begin af aan (naar) een derde weg tussen het
of start off on (to) a third way between the

communistische oostblok en het vrije, kapitalistische westen **zocht.**
communist Eastern bloc and the free, capitalist West searched.
(WR-P-P-G-0000103341.p.3.s.3)
'The World Council of Churches has not done that, because, from the very beginning, it was searching for a third way between the communist Eastern bloc and the free, capitalist West.'

To sum up, when the theme precedes the verb, more complex themes are likely to elicit the use of the variant without *naar*. Conversely, in instances where the theme is not complex, the producer is hardly under any processing pressure, and s/he might very well contemplate the choice of verb early on and choose to include *naar*. We therefore make the following prediction.

Production Hypothesis: There should be a negative correlation between THEME COMPLEXITY and the likelihood of *naar* when the theme precedes the verbs, and a positive correlation when the verb precedes the theme.

4.2. CHANNEL HYPOTHESIS

Taking the channel perspective, a parallel reasoning can be made. In cases where the theme precedes the verb, the presence of the preposition *naar* limits the number of verbs that may follow. Hence *naar* makes the following verb *zoeken* more predictable and therefore reduces its information content. Of course, since *naar* does not actually change the meaning of the sentence, this information does not just disappear; it is rather transferred over from the verb to the preposition. To sum up, the preposition signals a lot of information about the verb that is to follow.

This means that, in instances where the preposition precedes the verb, the preposition already carries a lot of information. Combining such an informationally heavy preposition with a complex, informationally heavy theme would lead to a peak in information density, which should be avoided. Instead, combining the heavy preposition with a simple, informationally light theme would smooth out the information density.

Of course, this reasoning only holds for instances where the theme precedes the verb. When the verb precedes the theme, the preposition evidently cannot signal any information about the verb, because the verb is already known at that point. As such, the now informationally light preposition can nicely combine with complex, informationally heavy themes. This leads to the following prediction, which is identical to the prediction made by the production hypothesis.

Channel Hypothesis: There should be a negative correlation between THEME COMPLEXITY and the likelihood of *naar* when the theme precedes the verbs, and a positive correlation when the verb precedes the theme.

In objection to the reasoning above, it could be claimed that, when the theme precedes the verb, the addition of *naar* doesn't actually make the verb that much more predictable. Perhaps the theme by itself already narrows down the list of possible verbs to a large degree, and *naar* doesn't do much to narrow it down even further. We can then ask, for all instances in our dataset where the theme precedes the verb, how much more predictable *naar* would actually make the verb, if it were included. We estimate this in the following way. First, we look at the lemmas of the syntactic heads of the themes. We will refer to these lemmas as 'theme lemmas'. For each theme lemma in the subset of our dataset where the theme precedes the verb, we count the number of times it appears as the syntactic head of a noun phrase with *zoeken*, in the SoNaR corpus.[8] Next, we count the number of times it appears as the syntactic head of a noun phrase with any verb. Finally, we divide the former by the latter. This yields for each theme lemma the probability that it combines with the verb *zoeken*. The average of these probabilities over the subset of our dataset where the theme precedes the verb is 0.0279.[9] This means that, given the theme lemma, there's on average a 2.79% chance that the upcoming verb is *zoeken*.

We now do the same calculations for the variant with *naar*. We count for each theme lemma in the same subset the number of times it appears as the syntactic head of a prepositional phrase introduced by *naar* with *zoeken*, in the SoNaR corpus. Next, we count the number of times it appears as the syntactic head of a prepositional phrase introduced by *naar* with any verb. Finally, we divide the former by the latter. The average of these probabilities over the same subset is 0.2364. This means that, given the theme lemma and the preposition *naar*, there's on average a 23.64% chance that the verb will be *zoeken*. Including *naar* thus makes us on average 8.5 times more confident in our guess that the following verb is *zoeken*, which we regard as a considerable increase.

The next paragraph outlines this reasoning more formally. Although the following procedure may seem to take a different outlook on the issue, the calculations are fundamentally the same. The procedure is based on

[8] Again, the tweets, chat logs, text messages, and discussion lists were excluded.

[9] Instances for which one of the probabilities was equal to zero, e.g. because the theme lemma of a prepositional object instance never appeared in a bare noun phrase in the entire SoNaR corpus, were not taken up in the calculation of the average.

Jaeger (2010, p. 28), who estimates the Shannon information of a complement clause in a similar way, i.e., by taking the negative logarithm of the probability that a complement cause would follow, given the matrix verb lemma. These calculations are necessarily only approximate measurements (Jaeger, 2010, p. 28).

We assume that the sentences in (15) are all synonymous, i.e., that they all contain the same information in total. Now we want to estimate the difference in Shannon information of *naar* when the theme precedes the verb (15a) versus when the verb precedes the theme (15b), as expressed in (i). Under the current reasoning, we expect this difference to be positive, because that would mean that *naar* is informationally heavier when it precedes the verb than when it follows the verb. The difference corresponds to the degree to which *naar* in (15a) makes the verb more predictable.[10] That is, it corresponds to the difference in information of *zoeken* in (15a) versus in (15c), as expressed in (ii). We now assume that it is primarily the theme lemma that makes *zoeken* in (15c) more predictable and we therefore estimate the information of *zoeken* in (15c) as its information given the theme lemma. Correspondingly, we estimate the information of *zoeken* in (15a) as its information given the theme lemma and *naar*. We now have (iii). Shannon information can be calculated as the negative logarithm of the probability, which gives us (iv).[11] For each instance of a preverbal theme in our dataset, we then calculate the probabilities in (iv) as described above, which gives us the estimated ΔI_{naar}. Finally, we take the average of these, as in (v). This means that *naar* is on average estimated to be 3.70 bits heavier when the theme precedes the verb than when the verb precedes the theme.

(15) (a) Ik heb gisteren naar een schaar **gezocht.**
I have yesterday to a scissors searched
(b) Ik heb gisteren **gezocht** naar een schaar.
I have yesterday searched to a scissors
(c) Ik heb gisteren een schaar **gezocht.**
I have yesterday a scissors searched
'I have searched for a pair of scissors yesterday.'

(i) $\Delta I_{\text{naar}} = I(\text{naar} \mid \textit{preverbal theme}) - I(\text{naar} \mid \textit{postverbal theme})$

(ii) $= I(\text{zoeken} \mid \textit{preverbal theme} \,\&\, \textit{direct obj.}) - I(\text{zoeken} \mid \textit{preverbal theme} \,\&\, \textit{prep.obj.})$

(iii) $\approx I(\text{zoeken} \mid \textit{preverbal theme lemma}) - I(\text{zoeken} \mid \textit{preverbal theme lemma} \,\&\, \text{naar})$

(iv) $= -\log_2 p(\text{zoeken} \mid \textit{preverbal th.lemma}) + \log_2 p(\text{zoeken} \mid \textit{preverbal th.lemma} \,\&\, \text{naar})$

(v) $\textit{Average}\, \Delta I_{\text{naar}} \approx 3.70\ \textit{bits}$

[10] Of course preverbal *naar* in (2) also makes the theme lemma more predictable, but this holds equally for postverbal *naar* in (1).

[11] We used a logarithm with base 2, as in the seminal paper by Shannon (1948).

This would mean that the information content of the preposition *naar* is dependent upon its position relative to the verb. If the verb precedes *naar*, the preposition evidently cannot contain any information about the verb, since the verb is already known when *naar* is heard or read. Since *naar* is thus informationally light, it can nicely combine with a complex, informationally heavy theme. Meanwhile, if *naar* precedes the verb, it is burdened with a large chunk of the information content otherwise contributed by the verb, thus rendering it informationally heavy. In that case, it would be preferable not to combine it with a complex, informationally heavy theme.

4.3. COMPREHENSION HYPOTHESIS

Finally, the comprehension perspective stated that for the comprehender *naar* functions as a signpost that simplifies the parsing of a complex theme. Such a signpost would be especially useful if a complex theme precedes the main verb, since in that case it already gives considerable information about the verb that is to follow, as argued above. Because the main verb for a large part determines the structure of the entire sentence, knowledge of this verb would further simplify parsing to a great extent (Müller, 2006; Müller & Wechsler, 2014). As such, we formulate the following hypothesis.

> **Comprehension Hypothesis**: There should be a strong positive correlation between THEME COMPLEXITY and the likelihood of *naar* when the theme precedes the verbs, and a weaker positive correlation when the verb precedes the theme.

To sum up, our three perspectives make different predictions about how the correlation between complexity and explicitness behaves in different linguistic contexts. In particular, we have argued that the relevant distinction will be one between a context where the theme precedes the verb and one where the verb precedes the theme. We will now check this, which will yield an answer to the second research question. If we find any of the hypotheses above to be confirmed, this will also answer our first research question.

4.4 ANALYSIS

To test the hypotheses made in the previous subsections, we compose a mixed logistic regression model that has as the dependent variable the presence or absence of *naar* and THEME COMPLEXITY as a fixed effect.[12] THEME

[12] For the application of (mixed) logistic regression models in corpus research, see Baayen (2008), Gries (2015), Speelman (2014), and Speelman et al. (2018b).

COMPLEXITY is a numeric variable, so it can be directly implemented as a parameter in the model.

We also add the variable VERB–THEME ORDER as a fixed effect, as well as an interaction between THEME COMPLEXITY and VERB–THEME ORDER. This variable distinguishes between the contexts where the theme precedes the verb, and those where the verb precedes the theme. For a categorical variable such as VERB–THEME ORDER, we need an additional coding step to implement it into the regression formula. For this, we use user-defined sum-to-zero contrasts, also called user-defined sum coding. This type of coding has a number of advantages over more traditional treatment contrasts or dummy coding. Most notably, the odds ratios can be interpreted as deviations to the group mean instead of to a reference level that sometimes needs to be chosen arbitrarily. For a more in-depth discussion of user-defined contrasts in linguistics, see Heller (2018, pp. 85–88). VERB–THEME ORDER has only two levels, viz. *theme–verb* and *verb–theme*, so can be implemented into the regression formula with just one parameter. For the instances of *theme–verb*, this parameter is set to 1. For those of *verb–theme*, it is set to –1.

The variable VERB–THEME ORDER has some correlates that we want to control for. These are the variables CLAUSE TYPE and VERB FINITENESS, which are both added as fixed effects. CLAUSE TYPE distinguishes between main clauses and subordinate clauses. It is implemented as a single parameter that was set to 1 if the occurrence appears in a main clause, and –1 if it appears in a subordinate clause. VERB FINITENESS distinguishes between instances where the main verb *zoeken* 'search' is a finite form, an infinitive, or a participle. Because this variable has three levels, it is implemented with two parameters. The first parameter distinguishes between the finite and non-finite forms. It is set to 1 for finite forms, and –0.5 for infinitives and participles. The second parameter then distinguishes between the infinitives and the participles. It was set to 0 for finite forms, to 1 for infinitives, and –1 for participles.

Previous research on a similar alternation in Dutch, the *er*-alternation, has revealed that, although the processing motivation for *er* is highly comparable in Belgian and Netherlandic Dutch, there are considerable differences between the Belgian and Netherlandic models (Grondelaers, van den Bosch, Speelman, & van Hout, 2015; Grondelaers, Speelman, & Geeraerts, 2008; van den Bosch, Grondelaers, & Speelman, forthcoming). We therefore also include the variable COUNTRY as a fixed effect in the model. This variable is implemented as a single parameter, set to 1 for Belgian occurrences, and –1 for occurrences from the Netherlands.

As mentioned above, the variants with and without *naar* of the verb *zoeken* 'search' are considered synonymous (Haeseryn et al., 1997, p. 1168). Still, subtle semantic differences have been proposed for similar alternations in English

(Goldberg, 1995, pp. 118–119, 1999, pp. 198–200; Perek, 2015) and we want to err on the right side of caution.[13] Based on theoretical accounts such as Goldberg (1995), Hopper and Thompson (1980), and Langacker (1991), it could be suggested that the prepositional variant implies a form of directionality, or movement to a place, while the transitive variant implies an undergoer being affected. These notions relate to the theme argument. For example, when the theme is a place, as in (16), the act of searching typically implies an attempt to move to that place, and we could therefore theorize a preference for the prepositional variant. Meanwhile, in (17), the act of searching implies an attempt to formulate the formulas, i.e., to bring the formulas into being. In other words, the formulas are deeply affected by the act of searching. As such, we could theorize a preference for the transitive variant.

(16) We **zochten** naar een aanlegplaats en zagen er geen, dus
We searched to a landing_place and saw there no so
we bleven op een afstand liggen.
we stayed on a distance lay
(WR-P-P-B-0000000170.p.2076.s.3)
'We were searching for a landing place, but didn't see any, so we kept our distance.'

(17) *Vicsek* ***zoekt*** *wiskundige formules die hun vormen beschrijven*.
Vicsek searches mathematical formulas that their form describe
(WR-P-E-G-0000004047.p.139.s.1)
'Vicsek searches for mathematical formulas that describe their shapes.'

We will attempt to control for such a semantic differentiation by looking at the theme lemmas. For instance, in (16), the theme lemma is *aanlegplaats* 'landing place', and in (17), it is *formule* 'formula'.[14] In particular, we take over a technique first proposed by Levshina and Heylen (2014) and elaborated upon by Speelman, Heylen, and Grondelaers (forthc.), which involves adding

[13] In fact, a semantic distinction has also been proposed for the English *that*-alternation (Elsness, 1984, p. 526; Thompson & Mulac, 1991), which is nonetheless used in much of the research on the Complexity Principle (see above).

[14] Actually, we used a combination of the Alpino *root*-tag and the *pos*-tag, e.g. *aanlegplaats**noun* in (16), in keeping with Levshina and Heylen (2014) and Speelman, Heylen, and Grondelaers (forthc.). Furthermore, the formal realization of some third person pronouns is dependent on whether it appears in the transitive or prepositional variant. For instance, the demonstrative pronoun *dat* can be realized as the pronominal adverb *daarnaar* 'thereto' in the prepositional variant, but never in the direct object variant. As such, the theme lemmas of these pronouns were collapsed with those of their corresponding pronominal adverbs. We did not distinguish between the stressed and non-stressed versions of the personal pronouns (e.g., *jij* and *je* 'you').

a variable called SEMANTIC CLUSTER. For each full nominal theme lemma, we calculated a distributional vector based on the SoNaR corpus, and then clustered these vectors into 50 semantic groups.[15] The variable SEMANTIC CLUSTER then distinguished between these clusters. In this way, the theme lemma *aanlegplaats* 'landing place' is grouped in a semantic cluster with other places, such as *opvanglocatie* 'shelter location', *weideland* 'pasture', and *slaapplek* 'sleeping place', while the theme lemma *formule* 'formula' ends up in a cluster with lemmas like *methode* 'method', *tactiek* 'tactic', and *techniek* 'technique'. The pronominal theme lemmas were not clustered, but rather directly added as individual levels of this variable.

SEMANTIC CLUSTER was then introduced into the regression model as a random effect with random intercepts. The variable was added as a random effect rather than a fixed effect because: (i) it only functions as a control variable, and we are currently not directly interested in its effects; (ii) it has 96 distinct levels; and (iii) the levels of the variable are in principle not exhaustive, i.e., the verb *zoeken* 'search' could be used with a theme lemma that does not fit into any of the clusters in our present dataset (Speelman, Heylen, & Geeraerts, 2018a, p. 3).

To control for the influence of register, the corpus component was also added as an additional random effect with random intercepts. It would also have been possible to consider register a fixed effect. In that case, one would typically use coarse-grained levels that are exhaustive, such as *formal register* vs. *informal register*. However, we prefer to directly use the more fine-grained distinction between individual corpus components, which means that the levels are not repeatable when a follow-up study would use a different corpus. As such, we opt for random effects. For a discussion of the merits of both approaches, see Speelman et al. (2018a, p. 3).

[15] For background on distributional vectors, see Turney and Pantel (2010). The vectors used dependency-based features of eight possible relations discerned by the Alpino-parser: subject, direct object, prepositional object, conjunction, apposition, adverbial prepositional phrase, post-modifying prepositional phrase, and adjective. Examples of such dependency-based features are *subject-of-zien* 'subject of to see', *modified-by-adjective-bruin* 'modified by the adjective brown', etc. This type of vector was found most successful by Levshina and Heylen (2014). Only the 5000 most frequent dependency-based context features were used in the vectors, and the frequencies were weighted through positive point-wise mutual information. Context features with Alpino POS-tags that correspond to function words were not used. The vectors were clustered using the hierarchical clustering algorithm *ward.D* in R, based on cosine distances. The occurrences of the theme lemma with *zoeken* 'search', i.e. the occurrences that can be found in the dataset, were excluded in the calculation of the vectors to avoid circularity. Those theme lemmas that uniquely appear with *zoeken* in the SoNaR-corpus (i.e. some proper names and multi-word expressions) and for which, as a result, no vector could be calculated, were added as a separate level called *unique name* of the variable SEMANTIC CLUSTER. Likewise, the numerals were added as a single, separate level, as were the subordinate clauses.

This model was then fitted to the data. Multicollinearity was not found to be a problem, with the condition number ($\kappa = 6.28$) below the conventional threshold of 15 (Wolk, Bresnan, Rosenbach, & Szmrecsanyi, 2013, p. 401). The model has a C-index of 0.734 (Somer's Dxy = 4.67), indicating acceptable discrimination (Hosmer & Lemeshow, 2000, p. 162). The model specifications can be found in Table 3.[16] The presence of *naar* is the success level of the response variable, so the intercept represents the odds of *naar*, with all parameters of the model set to 0. That is, it represents the odds of *naar* for themes that are 1 word long (THEME COMPLEXITY = log(1) = 0), averaged over all categorical variables. The Odds Ratio of 1.20 for CLAUSE TYPE indicates that the odds of *naar* increase with factor 1.20 in main clauses, compared to the mean of both clause types, for themes of 1 word, averaged over both countries, etc.

Figure 2 shows the effect plot of the interaction between THEME COMPLEXITY and VERB–THEME ORDER, which visualizes the estimated probability of *naar* as a function of THEME COMPLEXITY for the observations where the theme precedes the verb, and for those where the verb precedes the theme. We find a negative correlation when the theme precedes the verb and a positive correlation when the verb precedes the theme. This confirms the Production and Channel Hypotheses.

5. Discussion and conclusions

We can now formulate an answer to the research questions, which are repeated below.

(i) What drives the correlation between complexity and explicitness as we find it in corpora?
(ii) Does the correlation hold in all contexts, and if not, in which ones?

As for the first question, our results indicate that, with regard to the alternation under scrutiny, the correlation between complexity and explicitness is primarily motivated by either production processing or channel constraints. This dovetails with the majority of the literature on the influence of production vs. comprehension processing, including Ferreira and Dell, (2000), Kraljic and Brennan (2005), Roland et al. (2006), Ferreira and Hudson (2011), Ferreira and Schotter (2013), Gennari and Macdonald (2009), Jaeger (2010), Levy and Jaeger (2007), MacDonald (2013), and MacDonald and Thornton (2009).

[16] For the analyses, we used R (R Core Team, 2014) and the packages *lme4* (Bates, Maechler, Bolker, & Walker, 2013), *effects* (Fox, 2003), and *Hmisc* (Harrell, 2017).

TABLE 3. *Mixed effects logistic regression model predicting the presence of preposition naar*

AIC: 51,146.4	Observations without *naar*: 55,232
C-index: 0.734	Observations with *naar* (success level): 10,354

Fixed effects	**Level**	**Odds Ratio**	**Confidence interval**		**P-value**
			2.5%	**97.5%**	
	intercept	0.13	0.09	0.19	< .0001
THEME COMPLEXITY		0.93	0.9	0.97	0.0001
VERB–THEME ORDER	*theme–verb (vs. verb–theme)*	1.07	1.01	1.13	.0214
CLAUSE TYPE	*main (vs. subordinate)*	1.20	1.16	1.25	< .0001
VERB FINITENESS	*finite (vs. infinitive & participle)*	1.07	1.01	1.12	.0122
	infinitive (vs. participle)	0.71	0.67	0.74	< .0001
COUNTRY	*Belgium (vs. the Netherlands)*	0.89	0.87	0.92	< .0001
Interaction THEME COMPLEXITY and VERB–THEME ORDER	*theme–verb (vs. verb–theme)*	0.76	0.73	0.78	< .0001

Random effects	**Number of levels**	**Variance**	**Standard Deviation**
SEMANTIC CLUSTER	95	1.37	1.17
COMPONENT	17	0.12	0.35

Still, it should be noted that our results do not entail that the use of explicit coding is completely unbeneficial to the comprehender. Regarding the production perspective, we included both the PDC-model and the collateral signals account, both of which hold that the comprehender does benefit, albeit in an indirect way. There is, in fact, strong evidence that the comprehender interprets both disfluencies and grammatical markers such as the English subordinator *that* and Dutch existential *er* as signals of upcoming production difficulties or unpredictable material (Grondelaers et al., 2009; Jaeger, 2005, see also Clark & Fox Tree, 2002; Collard et al., 2008; Corley & Hartsuiker, 2003). Perhaps it is more relevant for the comprehender to receive notifications on the current state of production than to procure sentences that are (marginally) easier to parse. Additional research would need to confirm whether the comprehender does indeed interpret the Dutch preposition *naar* as such a signal, but our current results certainly do not exclude it; they rather indicate that it is possible. Regarding the channel perspective, it may very well be in the interest of the comprehender to burden his or her own cognitive processing if a more important goal is safeguarded. For example, it is in the

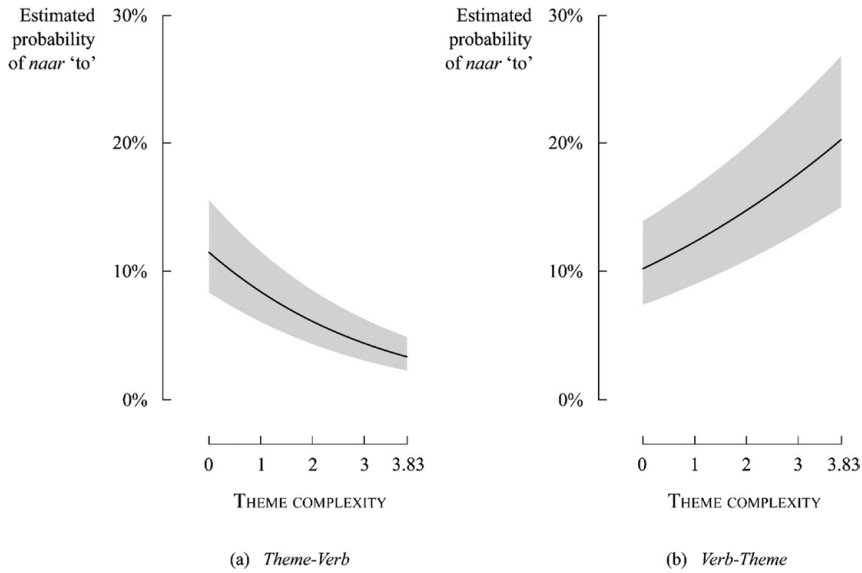

Fig. 2. Effect plot of the interaction between THEME COMPLEXITY and VERB–THEME ORDER. When the theme precedes the verb, the likelihood of *naar* decreases as the theme becomes more complex. Meanwhile, when the verb precedes the theme, the opposite effect arises. This confirms both the producer- and channel-driven explanations of the Complexity Principle.

interest of both the comprehender and the producer to make sure that as little information as possible is lost in the noisy language channel by making sure the information density does not exceed its optimal level too much or too often. If that leads to tendencies that require more cognitive effort during parsing, this may very well be a price worth paying.

Meanwhile, for the tradition of corpus-based alternation research that is not primarily concerned with language processing, the answer to the second research question is perhaps more interesting. Here, our findings indicate that the Complexity Principle should not be interpreted as a blind law, but rather as a general tendency that holds in most, but not all contexts. This is also argued by Rohdenburg (2016) and Willems and De Sutter (2015), who propose further refinements to the Complexity Principle. In order to determine in which context we can expect the Principle to hold, we need to consider its underlying mechanism, as well as the specifics of the case study. For example, we have shown that the order of theme and verb is a relevant distinction in our case study, with the effect of the Complexity Principle reversing when the theme precedes the verb. Such context-determined restrictions to the Principle present a possible caveat for alternation studies, which do not always take the underlying mechanisms of the Complexity Principle into account.

There are many possibilities for further research. One possibility is to design a clever operationalization to differentiate between on the one hand the production-driven account proposed in Ferreira and Dell (2000) and MacDonald (2013), and on the other hand, the channel-driven model underlying the principle of Uniform Information Distribution (Fenk-Oczlon, 2001; Jaeger, 2010). It is certainly possible that both influence the choice for explicit grammatical coding, but the question would then be how we can predict which mechanism is at play under which conditions, or which takes precedence when their predictions collide. Going further, we would want to differentiate between the accounts subsumed under the production perspective, viz. PDC-model and the collateral signal account.

Another possibility for further research is to repeat the same investigation on other case studies and other languages. We hope our current focus on Dutch may inspire researchers on this topic to take under scrutiny case studies outside the English language. Finally, a large-scale alternation study on the direct vs. prepositional object alternation in Dutch is still necessary. Such a study would map out all alternating verbs and aim to shed light on all other major factors governing the alternation, including those of semantic and lectal nature, and how they possibly interact.

REFERENCES

Arnold, J., Wasow, T., Asudeh, A. & Alrenga, P. (2004). Avoiding attachment ambiguities: the role of constituent ordering. *Journal of Memory and Language* **51**(1), 55–70.

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013). *lme4: linear mixed-effects models using Eigen and S4. R package version 1.4*. Retrieved from <http://cran.r-project.org/package=lme4>.

Bloem, J. (2016). Evaluating automatically annotated treebanks for linguistic research. In P. Baski, M. Kupietz, H. Lüngen, A. Witt, A. Barbaresi, H. Biber, … S. Clematide (Eds.), *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC-4)* (pp. 8–14). Manheim: Institut für Deutsche Sprache.

Bloem, J., Versloot, A. & Weerman, F. (2014). Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014: the 25th International Conference on Computational Linguistics: technical papers*. August 23–29, 2014, Dublin (pp. 1974–1984).

Bloem, J., Versloot, A. & Weerman, F. (2017). Verbal cluster order and processing complexity. *Language Sciences* **60**, 94–119.

Bock, K., Irwin, D. & Davidson, D. (2004). Putting first things first. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: eye movements and the visual world* (pp. 224–250). New York: Psychology Press.

Bolinger, D. (1972). *That's that*. The Hague: Mouton.

Bolinger, D. (1980). Wanna and the gradience of auxiliaries. In G. Brettschneider & C. Lehmann (Eds.), *Wege zur Universalienforschung: sprachwissenschaftliche Beiträge zum 60. Geburtstag von Hansjakob Seiler* (pp. 292–299). Tübingen: Gunter Narr.

Bouma, G. (2017). Om-omission. In M. Wieling, M. Kroon, G. van Noord & G. Bouma (Eds.), *From semantics to dialectometry* (pp. 65–73). Groningen: College Publications.

Bouma, G. & Kloosterman, G. (2002). Querying dependency treebanks in XML. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)* (pp. 1686–1691). Online <http://www.let.rug.nl/gosse/papers/bouma_lrec2002.pdf>.

Bouma, G. & Kloosterman, G. (2007). Mining syntactically annotated corpora with XQuery. In *Proceedings of the Linguistics Annotation Workshop (ACL 07)* (pp. 17–24). Online <http://www.let.rug.nl/~gosse/papers/law07.pdf>.
Brennan, S. & Williams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* **34**(3), 383–398.
Bresnan, J., Cueni, A., Nikitina, T. & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Krämer & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.
Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language* **82**(4), 711–733.
Clark, H. (1996). *Using language*. Cambridge: Cambridge University Press.
Clark, H. (2004). Pragmatics of language performance. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 365–382). Walden: Blackwell.
Clark, H. & Fox Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition* **84**(1), 73–111.
Clark, H. & Murphy, G. (1982). Audience design in meaning and reference. *Advances in Psychology* **9**, 287–299.
Collard, P., Corley, M., MacGregor, L. & David, D. (2008). Attention orienting effects of hesitations in speech: evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **34**(3), 696–702.
Corley, M. & Hartsuiker, R. (2003). Hesitation in speech can um help a listener understand. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 276–281). Boston: Cognitive Science Society.
Cover, T. & Thomas, J. (1991). *Elements of information theory*. Hoboken: Wiley-Interscience.
Elsness, J. (1984). *That* or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* **65**, 519–533.
Fenk, A. & Fenk-Oczlon, G. (1993). Menzerath's law and the constant flow of linguistic information. In R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 11–31). Dordrecht: Kluwer Academic Publishers.
Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45) (pp. 431–448). Amsterdam: John Benjamins.
Ferreira, V. & Dell, G. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* **40**(4), 296–340.
Ferreira, V. & Hudson, M. (2011). Saying 'that' in dialogue: the influence of accessibility and social factors on syntactic production. *Language and Cognitive Processes* **26**(10), 1736–1762.
Ferreira, V. & Schotter, E. (2013). Do verb bias effects on sentence production reflect sensitivity to comprehension or production factors? *Quarterly Journal of Experimental Psychology* **66**(8), 1548–1571.
Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software* **8**, 1–27.
Fox Tree, J. & Clark, H. (1997). Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* **62**(2), 151–167.
Garnsey, S., Pearlmutter, N., Myers, E. & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* **37**(1), 58–93.
Gennari, S. & Macdonald, M. (2009). Linking production and comprehension processes: the case of relative clauses. *Cognition* **111**(1), 1–23.
Gleitman, L., January, D., Nappa, R. & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language* **57**(4), 544–569.
Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.
Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In B. Macwhinney (Ed.), *Emergence of language* (pp. 197–212). Hillsdale: Lawrence Earlbaum Associates.

Gries, S. T. (2002). The influence of processing on grammatical variation: particle placement in English. In N. Dehé, R. Jackendoff, A. McIntyre & S. Urban (Eds.), *Verb-particle explorations* (pp. 169–288). Berlin/New York: Mouton de Gruyter.
Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: a study of particle placement*. New York: Continuum.
Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* **10**(1), 95–125.
Grondelaers, S. (2000). *De distributie van niet-anaforisch er buiten de eerste zinsplaats: sociolexicologische, functionele en psycholinguïstische aspecten van er's status als presentatief signaal*. Unpublished dissertation, University of Leuven.
Grondelaers, S., van den Bosch, A., Speelman, D. & van Hout, R. (2015). Comparing memory-based learning and regression approaches in the explanation of syntactic variation and change in Belgian and Netherlandic Dutch. Paper presented at *New Ways of Analyzing Variation (NWAV 43)*. 26 October, Chicago.
Grondelaers, S. & Speelman, D. (2007). A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* **3**(2), 161–193.
Grondelaers, S., Speelman, D., Drieghe, D., Brysbaert, M. & Geeraerts, D. (2009). Introducing a new entity into discourse: comprehension and production evidence for the status of Dutch er 'there' as a higher-level expectancy monitor. *Acta Psychologica* **130**(2), 153–160.
Grondelaers, S., Speelman, D. & Geeraerts, D. (2003). *De distributie van er in het gesproken Nederlands*. Paper presented at the workshop spraakmakende spraak. 16 May, Nijmegen.
Grondelaers, S., Speelman, D. & Geeraerts, D. (2008). National variation in the use of er 'there': regional and diachronic constraints on cognitive explanations. In G. Kristiansen & R. Dirven (Eds.), *Cognitive sociolinguistics: language variation, cultural models, social systems* (pp. 153–203). Berlin/New York: Mouton de Gruyter.
Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. & van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Groningen: Nijhoff.
Harrell, F. J., with contributions from Charles Dupont & many others (2017). Hmisc: Harrell Miscellaneous. R package version 4.0-3. Retrieved from <https://cran.r-project.org/package=Hmisc>.
Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* **19**(1), 1–33.
Hawkins, J. (2002). Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics* **28**(2), 95–149.
Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
Heller, B. (2018). *Stability and fluidity in syntactic variation world-wide: the genitive alternation across varieties of English*. Unpublished dissertation, University of Leuven.
Hopper, P. & Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language* **56**(2), 251–299.
Hosmer, D. & Lemeshow, S. (2000). *Applied logistic regression*, 2nd ed. New York: Wiley.
Jaeger, F. T. (2005). Optional *that* indicates production difficulty: evidence from disfluencies. In *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop* (pp. 103–109). Aix-en-Provence. Online: <https://www.isca-speech.org/archive_open/archive_papers/diss_05/dis5_103.pdf>.
Jaeger, F. T. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished dissertation, Stanford University.
Jaeger, F. T. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* **61**(1), 23–62.
Jaeger, F. T. (2011). Corpus-based research on language production: information density and reducible subject relatives. In E. Bender & J. Arnold (Eds.), *Language from a cognitive perspective: grammar, usage, and processing. Studies in honor of Tom Wasow* (pp. 161–197). Stanford: CSLI Publications.
Jaeger, F. T. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, **4**. Online <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00230/full>.

Jaeger, F. T., Levy, R., Wasow, T. & Orr, D. (2005). The absence of *that* is predictable if a relative clause is predictable. Paper presented at the Architectures and Mechanisms of Language Processing conference, Ghent.
Jaeger, F. T. & Wasow, T. (2005). *Production-complexity driven variation: relativizer omission in non-subjectextracted relative clauses*. Paper presented at the 18th annual CUNY Conference on Sentence Processing, 1 April, Tucson, AZ.
Kirby, S. (1999). *Function, selection, and innateness: the emergence of language universals*. Oxford: Oxford University Press.
König, E. & Gast, V. (2009). *Understanding English–German contrasts,* 2nd ed. Berlin: Erich Schmidt.
Konopka, A. E. (2012). Planning ahead: how recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language* **66**(1), 143–162.
Kraljic, T. & Brennan, S. (2005). Prosodic disambiguation of syntactic structures: For the speaker or for the addressee? *Cognitive Psychology* **50**(2), 194–231.
Langacker, R. W. (1991). *Foundations of cognitive grammar: descriptive application*. Stanford: Stanford University Press.
Levinson, S. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT press.
Levshina, N. & Heylen, K. (2014). A radically data-driven Construction Grammar: experiments with Dutch causative constructions. In R. Boogaart, T. Colleman & G. Rutten (Eds.), *Extending the scope of Construction Grammar* (pp. 17–46). Berlin: Mouton de Gruyter.
Levy, R. & Jaeger, F. T. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 849–856). Cambridge, MA: MIT Press.
MacDonald, M. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology* **4**, 226. Online <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00226/full>.
MacDonald, M. & Thornton, R. (2009). When language comprehension reflects production constraints: resolving ambiguities with the help of past experience. *Memory & Cognition* **37**(8), 1177–1186.
Menn, L. & Duffield, C. J. (2014). Looking for a 'Gold Standard' to measure language complexity: what psycholinguistics and neurolinguistics can (and cannot) offer to formal linguistics. In F. Newmeyer & L. Preston (Eds.), *Measuring grammatical complexity* (pp. 281–302). Oxford: Oxford University Press.
Müller, S. (2006). Phrasal or lexical constructions? *Language* **82**(4), 850–883.
Müller, S. & Wechsler, S. (2014). Lexical approaches to argument structure. *Theoretical Linguistics* **40**(1/2), 1–76.
Oostdijk, N., Reynaert, M., Hoste, V. & Schuurman, I. (2013a). *SoNaR User Documentation*. Online: <https://ticclops.uvt.nl/SoNaR_end-user_documentation_v.1.0.4.pdf>.
Oostdijk, N., Reynaert, M., Hoste, V. & Schuurman, I. (2013b). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: theory and applications of natural language processing* (pp. 219–247). Heidelberg: Springer.
Perek, F. (2015). *Argument structure in usage-based construction grammar: experimental and corpus-based perspectives*. Amsterdam/Philadelphia: John Benjamins.
Pijpops, D. & Van de Velde, F. (2016). Constructional contamination: How does it work and how do we measure it? *Folia Linguistica* **50**(2), 543–581.
Pijpops, D. & Van de Velde, F. (2018). Lectal contamination: how language-external variation becomes language-internal through language contact. Paper presented at *Variationist Linguistics meets Contact Linguistics*, 21 May, Ascona. Online: <https://lirias2repo.kuleuven.be/rest/bitstreams/500995/retrieve>.
R Core Team. (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Retrieved from <http://www.r-project.org/>.
Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* **7**(2), 149–182.

Rohdenburg, G. (2016). Testing two processing principles with respect to the extraction of elements out of complement clauses in English. *English Language and Linguistics* **20**(3), 463–486.
Roland, D., Elman, J. & Ferreira, V. (2006). Why is 'that'? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* **98**(3), 245–272.
Shank, C., Plevoets, K. & Bogaert, J. Van. (2016). A multifactorial analysis of *that*/zero alternation: the diachronix development of the zero complementizer with *think, guess* and *understand*. In J. Yoon & S. T. Gries (Eds.), *Corpus-based approaches to Construction Grammar* (pp. 201–240). Amsterdam/Philadelphia: John Benjamins.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**(3), 379–423.
Smith, V. & Clark, H. (1993). On the course of answering questions. *Journal of Memory and Language* **32**(1), 25–38.
Speelman, D. (2014). Logistic regression: a confirmatory technique for comparisons in corpus linguistics. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy* (pp. 487–533). Amsterdam: John Benjamins.
Speelman, D., Heylen, K. & Geeraerts, D. (2018a). Introduction. In D. Speelman, K. Heylen & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 1–10). Cham: Springer.
Speelman, D., Heylen, K. & Geeraerts, D. (2018b). *Mixed-effects regression models in linguistics*. Cham: Springer.
Speelman, D., Heylen, K. & Grondelaers, S. (forthcoming). A bottom-up, data-driven operationalization of semantic classes and predictability in syntactic alternation research. In S. Grondelaers & R. van Hout (Eds.), *New ways of analyzing syntactic variation*. Berlin: Mouton de Gruyter.
Tanner, D. & Bulkes, N. (2015). Cues, quantification, and agreement in language comprehension. *Psychonomic Bulletin & Review* **22**(6), 1753–1763.
Tanner, D., Nicol, J. & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language* **76**, 195–215.
Theijssen, D., Boves, L., Halteren, H. & Oostdijk, N. (2010). Evaluating automatic annotation: automatically detecting and enriching instances of the dative alternation. *Language Resources and Evaluation* **46**(4), 565–600.
Thompson, S. A. & Mulac, A. (1991). The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics* **15**(3), 237–251.
Turney, P. & Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**, 141–188.
van den Bosch, A., Grondelaers, S. & Speelman, D. (forthcoming). A memory based account of constructional differences between Netherlandic and Belgium Dutch. In Grondelaers, S. & van Hout, R. (Eds.), *New ways of analyzing syntactic variation*. Berlin: Mouton de Gruyter.
van Noord, G. (2006). At last parsing is now operational. *TALN*, 20–42. Online: <https://pdfs.semanticscholar.org/acc0/5b4412f3dceb49ee20f36dfe7cc6507af775.pdf>.
van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., … Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In P. Spyns & J. Odyk (Eds.), *Essential speech and language technology for Dutch* (pp. 219–247). Berlin: Springer.
Willems, A. & De Sutter, G. (2015). Reassessing the effect of the complexity principle on PP Placement in Dutch. *Nederlandse Taalkunde* **20**(3), 339–366.
Wolk, C., Bresnan, J., Rosenbach, A. & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* **30**(3), 382–419.
Zifonun, G., Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache* [Grammar of the German language]. Berlin: de Gruyter.
Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. New York: Hafner.
Zwart, J.-W. (2011). *The syntax of Dutch*. Cambridge: Cambridge University Press.