

Mutual attraction between high-frequency verbs and clause types with finite verbs in early positions: corpus evidence from spoken English, Dutch, and German

Gerard Kempen^a and Karin Harbusch^b

^aMax Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; ^bFaculty of Computer Science, University of Koblenz-Landau, Koblenz, Germany

ABSTRACT

We report a hitherto unknown statistical relationship between the corpus frequency of finite verbs and their fixed linear positions (early vs. late) in finite clauses of English, Dutch, and German. Compared to the overall frequency distribution of verb lemmas in the corpora, high-frequency finite verbs are overused in main clauses, at the expense of nonfinite verbs. This finite versus nonfinite split of high-frequency verbs is basically absent from subordinate clauses. Furthermore, this “main-clause bias” (MCB) of high-frequency verbs is more prominent in German and Dutch (SOV languages) than in English (an SVO language). We attribute the MCB and its varying effect sizes to faster accessibility of high-frequency finite verbs, which (1) increases the probability for these verbs to land in clauses mandating early verb placement, and (2) boosts the activation of clause plans that assign verbs to early linear positions (in casu: clauses with SVO as opposed to SOV order).

ARTICLE HISTORY

Received 15 October 2018
Accepted 3 July 2019

KEYWORDS

Verb frequency; SVO and SOV; finite verb position; sentence production; Lexical accessibility

1. Introduction: the accessibility–anteriority link

Speakers tend to *prioritise* sentence constituents whose content and form were planned with little processing effort: They are more likely to assign earlier positions to easy constituents than to constituents that were harder to plan, provided the grammar allows sufficient linear-order flexibility. Three oft-studied cases are the following. An important determinant of processing cost is *conceptual accessibility* (Bock & Warren, 1985). A well-known example is *animacy*: Animate/human referents are processed faster than inanimate referents, and constituents denoting the former often precede constituents denoting the latter (Kempen & Harbusch, 2003, 2004; Osgood & Bock, 1977; Vogels, Krahmer, & Maes, 2019). A second source of linear-order preferences concerns the *topic-comment* distinction. Constituents expressing topical (old, presupposed, known) conceptual content tend to receive sentence positions that precede constituents conveying a newly conceptualised comment. In the linguistic literature, the topic-comment distinction is often discussed under the heading of *information structure* (Lambrecht, 1994, Ch. 4; Vallduvi, 1992). The third source of processing complexity concerns word frequency (*lexical accessibility*).

Unless prevented by strict rules of grammar, frequent words tend to occupy earlier positions than rare words. For example, Fenk-Oczlon (1989) found that in “frozen” coordinations (i.e. fixed expressions such as *facts and figures*, *dead or alive*) the more frequent member tends to come first. (Recently, Berg, 2018 wrote a detailed overview of frequency effects in a variety of syntactic constructions.) Lexical accessibility is prone to online fluctuations under the influence of priming and visual cueing – conditions leading to pre-activation of lexical items and to the construction of sentences that afford early placement of pre-activated items (Hwang & Kaiser, 2015 and Sauppe, 2017 review the literature and report new data).

A property shared by these phenomena is that the words or constituents occupying early positions in the evolving sentence, have arguably been composed with less effort than their counterparts at later positions. Assuming that less effort implies shorter processing time, “low-effort” constituents can be inserted into a sentence frame earlier than high-effort constituents, if the grammar supplies suitable landing places in early positions. MacDonald (2013, p. 3) calls this the “*Easy First*” bias.

The greater processing effort required by fillers of later (posterior) compared to earlier (anterior) positions

CONTACT Gerard Kempen  gerard.kempen@mpi.nl

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/23273798.2019.1642498>

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

does not imply that the *overall* processing load recruited by a sentence increases while the speaker is adding more and more constituents. Sentence planning includes not only composing and ordering individual constituents but also assembling them into an overarching structure. Indeed, an experimental study in which speakers had to combine spontaneous language production with a secondary (“dual”) task did not provide evidence for cognitive processing load increasing towards the end of finite sentences (Ford & Holmes, 1978; see also Harley, 2014, pp. 431–432). Actually, the cross-sentence processing load seemed to be more or less stable. A plausible explanation is based on (effort-free and simultaneous) pre-activation of (1) future structural elements entailed by the current incomplete sentence frame, and (2) lexical items associatively or semantically linked to lexical items used in the current incomplete sentence, that happen to fit the syntactic fragments under construction.

This observation is in line with recent data showing that speakers, when confronted with a peak in the processing load due to a high-effort sentence continuation, may insert optional function words in front of the problematic upcoming fragment. The extra word, e.g. an “easy” clause-initial optional complementiser such as the subordinating conjunction *that*, lengthens the interval available for planning the difficult fragment, thus lowering the average (per-word) processing load leading up to the fragment (Jaeger, 2010; Levy & Jaeger, 2007). Conversely, when the upcoming fragment (the onset of the complement clause) is highly predictable, the probability of *not* using the optional word (“reduction”) increases. The effect is known as *Uniform Information Density* (UID).

In the next section, we describe an unexpected observation in spoken German, Dutch and English corpora that suggests an *accessibility–anteriority relation regarding finite verbs*, and propose a theoretical account. Section 3 describes the data collection methodology applied in the study. The statistical analyses of the data are reported in Section 4. Finally, Section 5 explores relations between our findings and other sentence production phenomena discussed above (Easy-First phenomena, Uniform Information Density), and proposes topics for further investigation.

2. An accessibility–anteriority link for finite verbs?

Empirical studies of effects of accessibility on linear order have focused on (pro)nominal constituents – presumably because nouns and pronouns enjoy much intraclausal freedom of position. In the present paper, we show that finite verbs, too, are subject to effects of accessibility, in spite of very restricted placement options. We

present corpus data showing that accessibility effects can manifest themselves not only in more anterior placement of constituents but also, if the grammar does not allow anterior placement, in alternative lexical choices for the constituent. These data, extracted from spoken language treebanks, suggest that the tendency for high frequency lexical items to occupy early sentential/clausal positions may have a double origin: not only in prioritisation of constituents that have multiple placement options, as sketched above, but also in a bias favouring the selection of high-frequency lexical material for early constituents that have a single, fixed position in the surface structure of a clause.

The data we collected concern frequency distributions of finite verbs in main and subordinate clauses of three Germanic languages with fairly rigid placement options for verbs, especially finite verbs. In English, a Subject-Verb-Object (SVO) language, the finite verb standardly occupies a relatively early position in main as well as subordinate clauses. In German and Dutch, however, there is a split between main and subordinate clauses: The finite verb is obligatorily placed clause-finally in subordinate clauses (SOV) but in clause-medial or earlier position in main clauses (SVO). Due to this split, the positional contrast of finite verbs in main versus subordinate clauses is larger in German and Dutch (VO vs. OV) than in English (always VO). (Given the data classifications used below, there is no need to discuss word order rules in more detail.)

The work to be reported began as fallout from a corpus study on a different topic (Kempen & Harbusch, 2017). As part of that study, we had computed lemma frequencies of verbforms functioning as head of a main clause (always finite), of a finite subordinate clause, or of a nonfinite subordinate clause (often called Verb Phrase, VP: infinitival, participial, gerund). In line with general linguistic practice, we had treated finite auxiliaries and modal verbs (*be, have, may, can, will, do*) as heads of finite clauses. They govern complement clauses whose heads are nonfinite verbs. This also holds for other verbs (*go, want, like, try*) that take nonfinite complements. These corpus data revealed a remarkable interaction, in all three languages, between total lemma frequency and clause type. In subordinate clauses, the probability of a finite verb to function as clause head turned out to be more or less constant across the frequency spectrum – as expected *a priori*. However, the probability of a finite verb to head a main clause appeared to rise with the total lemma frequency of the verb – a rise that went hand in hand with a corresponding fall of the probability of the verb’s lemma to head a nonfinite clause. Stated differently, high-frequency verbs tend to be *overused* (over-

represented) in main clauses, whereas their presence in finite subordinate clauses tends to align with their total lemma frequency (i.e. without overuse of high-frequency verbs). We refer to this pattern as the *main-clause bias of high-frequency finite verbs* (for short: *MCB effect*). By implication, high-frequency verb lemmas are *underused* as heads of nonfinite clauses. In the present paper, we provide a closer analysis of this intriguing data pattern, focusing on contrasts between the three languages.

The hypotheses to be tested presuppose that finite clauses are important units of sentence planning. This conclusion is based on the fact that during spontaneous speech large proportions of them are immediately preceded by disfluencies (hesitations, pauses, repetitions, repairs, false starts, etc.). This suggests that if a sentence consists of a sequence of several finite clauses, the speaker probably has planned them sequentially (see the literature surveys by Harley, 2014; Levelt, 1989). However, this does not rule out the possibility for a finite subordinate clause (e.g. a relative clause) to be center-embedded in a higher clause, or that a finite subordinate clause precedes the main clause in the finally delivered sentence. In fact, we assume that the time course of sentence planning corresponds to the hierarchy of clauses, such that the onset of planning a clause higher in the hierarchy precedes the onset of planning any embedded clause. Within a clause, parallel planning of multiple immediate syntactic constituents is possible, although due to incremental conceptualisation of the to-be-expressed message and differing accessibility conditions, some constituents will be initiated and completed earlier than other ones (cf. the “Easy First” phenomenon discussed in Section 1). Once a suitable shape has been determined, a constituent is ready to fill one of the linearly ordered positions (slots) defined by the grammar of the clause; but revisions of a constituent may continue until it is needed in an upcoming slot and passed on to phonological and articulatory processing stages.

This course of events implies that anterior clause positions exert more time pressure on the lexical selection of suitable fillers than posterior positions (see Sauppe, 2017 for experimental evidence from German). With respect to finite verbs, this means that easy and fast accessibility is in greater demand in main clauses than in subordinate clauses, given that the large majority of finite subordinate clauses is attached to the next higher (often main) clause as the latter’s most posterior constituent (tail recursion, creating the impression that finite clauses are like beads on a string).

In view of these general assumptions about the inter-clausal and intraclausal time course of the sentence planning process, and the temporal head-start enabled by

high accessibility of lexical items, we predict¹ the main-clause bias (MCB): The head-start of high-frequency verbs makes them privileged candidates (1) to become heads of clauses planned early (main rather than subordinate), and (2) within these clauses, to fill obligatory early positions. These two effects can be summarised succinctly as follows: *Clauses with early standard position of the finite verb “attract” high-frequency verbs.*

This is not the only effect, though. The data to be reported below show that the size of the MCB effect varies between the target languages. In order to account for this cross-language difference, we stipulate an attraction effect in opposite direction, which may influence clauses planned subsequently to the main clause at the top of the hierarchy.

As said, in all three languages the planning process starts with an obligatory main clause. However, finite clauses planned subsequently can be main or subordinate. Depending on the grammatical and pragmatic context, the choice may or may not be obligatory. One determining factor is the conceptual relation between the new clause and the preceding one at the top of the hierarchy. This relation is lexically realised by a coordinating conjunction (*and, but, for*) or by an introductory constituent marking the clause as subordinate: in case of complement or adverbial clauses by a subordinating conjunction (*because, if, that, while, when, although, etc.*), or by an introductory *wh*-phrase in case of relative clauses or dependent interrogative clauses. (The introducing constituent is optional in certain contexts, as in *I wasn’t aware (that) I was speeding.*) Sometimes the lexicalisation of the conceptual relation is (co-)determined by the illocutionary force of the propositions involved in the relationship. For instance, the causal coordinating conjunction *for* requires that both related clauses have independent illocutionary force (e.g. make an assertion), as in *I got a ticket for I was speeding.* In that case, the clauses can also be realised as main clauses of two separate sentence (*I got a ticket. I was speeding.*), or the speaker might have used the subordinating conjunction *because*. However, *because* does not require that the main clause expresses an assertion. Variant *I got a ticket because I was speeding* can be used if independent illocutionary force is associated with the sentence as a whole, not necessarily with the main clause (which may express a presupposition). This example shows that the choice between a main or a adverbial finite clause may be optional. Another example is the possibility to realise a sentence-final non-restrictive relative clause as the second conjunct of a main-clause coordination (e.g. *I got a speeding ticket, which I should pay within a month vs. I got a speeding ticket and I should pay it within a month*).

Speakers may thus revise their initial (tentative) choice, especially as long as no subordination marker has been overtly expressed, and select the main-clause format if licensed by the grammatical and pragmatic context. We refer to such clause format switches as *covert crossovers*. They may occur when the main clause resulting from the crossover comes after the main clause at the top of the hierarchy. (For corpus evidence in support of crossover scenarios in German and Dutch, see Kempen & Harbusch, 2017, 2018.)

The hypothesis of covert crossovers means that *competition* may arise between a main- and a subordinate-clause realisation. (Such competition is absent when the top-level main clause of the sentence is being planned.) The outcome of this competition will be sensitive to the frequency of the finite verb if early availability of the verb not only leads to acceleration of the planning process for the clause under construction but also to suppression and abandonment of the originally envisioned clause type. If one of the competing clause types requires a finite verb in a more anterior position than the other clause type, the outcome of the competition favours the former: *High-frequency verbs* may thus be said to *attract clauses with an early position for the head verb*. In German and Dutch subordinate clauses, the finite verb is clause-final (SOV) whereas it occupies verb-second position in main clauses (SVO). Therefore, in these languages high-frequency head verbs attract main clauses. In English, with SVO order in both clause types, high-frequency is not expected to favour covert subordinate-to-main crossovers.

Note that each such crossover has two effects on the distribution of main-finite and subordinate-finite proportions in a corpus. First, it increases the difference between the *overall* main-clause and subordinate-clause proportions. In addition, as a crossover is more likely with higher-frequency verbs, it also tends to steepen the slope of the MCB curve.

In sum, the reasoning developed here postdicts not only the MCB effect in languages with early finite verb positions in main clauses, but also the larger effect size

in languages with different, than in languages with similar, word orders in main vs. subordinate clauses. As regards the current three target languages, we hypothesise that the MCB effect in English is due to main clauses attracting high-frequency verbs; in German and Dutch the MCB is stronger due to the additional attraction of main clauses by high-frequency verbs triggering covert subordinate-to-main crossovers.

3. Methodology

Our data sources are the three syntactically annotated corpora listed in Table 1. To our knowledge, these were the largest treebanks for spoken German, Dutch and English available in the literature when we began the research project described here (2014). The spoken materials consist of sentences extemporaneously produced in varied dialogue situations (face-to-face or telephone conversations).

From the German VM dialogues, we used the sentences syntactically annotated in the TüBa-D/S treebank (Stegmann et al., 2000). TüBa-D/S uses tags that allow easy classification of clauses as VO (designating a main, always finite clause: henceforth Main-Fnt), or OV (in finite subordinate clauses: Sub-Fnt). Nonfinite verbforms (Non-Fnt) are identified by special part-of-speech tags. CGN contains spoken sentences from various different domains (news, telephone conversations, speeches, etc.). However, not all of them were produced spontaneously. In total, we discarded about 3800 sentences with read speech. The sentences had been annotated with relatively theory-neutral dependency graphs (Hoekstra et al., 2001; van der Beek, Bouma, & van Noord, 2002). The corpus specifies features that directly allow classifying clauses as Main-Fnt or Sub-Fnt. As in VM, part-of-speech tags enable classification of verbforms as nonfinite. SWB is a large corpus of dialogues comprising about 2500 phone conversations by 500 speakers from around the USA. SWB does not specify features enabling straightforward identification of clauses as Main-Fnt or Sub-Fnt. We rectified this by adapting TIGERSearch (König & Lezius, 2003), and writing our own JAVA software. In sum, all three treebanks were analysed by means of TIGERSearch along with JAVA programmes we developed ourselves (see also Dipper & Kübler, 2017 for a more detailed description of the TüBa-D/Z and TIGER annotation schemes, and for TIGERSearch as used in corpus studies of written German.)

In all three treebanks we had to *lemmatise* the verbforms, i.e. to assign them to a citation form ("lemma"; represented by the infinitive, except in case of English modal auxiliaries and a few defective verbs). A major subtask here concerned separable verbs:

Table 1. The spoken treebanks used in the present study: some important details.

Language	Full name of treebank and key references	Abbreviated name
German	VERBMOBIL Corpus Stegmann, Telljohann, and Hinrichs (2000); Wahlster (2000)	VM
Dutch	Corpus Gesproken Nederlands 2.0 Hoekstra, Moortgat, Schuurman, and van der Wouden (2001); van Eerten (2007)	CGN
English	SWITCHBOARD Corpus Godfrey, Holliman, and McDaniel (1992)	SWB

Table 2. Treebanks used in the present study. First data column: number of trees containing at least one finite or nonfinite verbform. Second data column: number of extracted verbform tokens. Rightmost column: number of different (unique) verbs (lemmas).

Language treebank		Number of		
		sentences	verbforms	verb lemmas
German	VM	38,328	50,676	1083
Dutch	CGN	126,787	162,985	3884
English	SWB	110,504	167,272	2564

combining the particle with the core verb. For lemmatisation purposes, we used published computational-linguistic databases containing lemmatised verbforms, and software developed in-house;² but we carefully checked the results manually. When reporting verb frequencies, we will always use the citation forms (lemmas). In order to obtain the *total lemma frequency* of a verb, we added the frequencies of all its finite and nonfinite forms. Excluded from all calculations were verbs within sentence fragments tagged as repairs or revision (Table 2).

We did not try to *disambiguate* verbforms. That is, if a verbform can be allocated to more than one infinitive (e.g. *lay* as finite form of *lie* or *lay*), we arbitrarily chose one (always the same). If the citation form itself is ambiguous, that is, belongs to multiple subclasses of verbs (e.g. intransitive or transitive, full verb or auxiliary), we adopted the verb-class tag already attached to the verbform in the treebank; we did not try to disambiguate polysemous or homophonous verbs (e.g. *lie*). Informal inspection of published word frequency counts (e.g. Leech, Rayson, & Wilson, 2014) reveals that the incidence of these types of ambiguities is too low to seriously affect the data patterns we are focusing on in any of the three target languages. In sum, we largely relied on the parse-tree information stored in the treebanks, deviating from it only in case of obvious parsing errors or lacunae.

As final preparatory step we assigned a *clause type* to each individual verbform token: *main finite* (Main-Fnt, including parentheticals such as *you know*, and imperatives), *subordinate finite* (Sub-Fnt: complement, adverbial, and relative clauses), and *nonfinite* (Non-Fnt: infinitival, participial, gerund). In the present paper, nonfinite forms will receive only cursory attention. For each of

the various clause types, and for each treebank separately, we defined a set of search queries based on the treebank's morphological, lexical and syntactic tagging system and on the relative positions of these tags and other node labels in the syntactic trees.

An important principle we adhere to in our counts is “one head verb, one clause”: every verb token is head of a clause; and vice-versa, every clause has one head verb. (We treat finite auxiliary and modal verbs as heads of their clauses, not the nonfinite verbs they govern.) This means we may use the phrases “number of head verbs” and “number of clauses” interchangeably. Although a finite clause contains exactly one finite head, it may include constituents that themselves consist of a hierarchy of one or more nonfinite clauses (as in *He will [try [to sell his bike]]*). As regards coordinate structures: Two or more clauses participating in a coordination were counted separately; clauses featuring Gapping (as in *John loves Mary, and Peter Jane*) or other elliptical constructions, were not included in the counts. Given our focus on finite verbs, we discarded all verb lemmas whose corpus occurrences (tokens) consisted of nonfinite verbforms only.³ Table 3 shows some key figures of the resulting dataset.

For each lemma we counted how often its forms occurred as head of a main, a finite subordinate, or a nonfinite clause (Main-Fnt, Sub-Fnt, Non-Fnt), and calculated its total lemma frequency by adding the three numbers. Dividing the Main-Fnt, Sub-Fnt and Non-Fnt tokens of a lemma by its total frequency yields *clause-type proportions*, which define the *cross-clause-type* distribution of the lemma (a measure of the lemma's relative attraction to each of the clause types). The three clause-type proportions of a verb lemma *v* can be expressed as conditional probabilities: Prob(Main-Fnt|*v*), Prob(Sub-Fnt|*v*), and Prob(Non-Fnt|*v*). In order to remove the effects of differing corpus sizes, we normalised the total frequency of each lemma by dividing it by the sum of all lemmas per language (i.e. the numbers in the rightmost column in Table 4 below). This gives the *normalised total frequency* (NormTotFreq) of each lemma.

When computing the average clause-type distribution of a *group* of lemmas, we can do this on the basis of the

Table 3. Numbers of unique verb lemmas, and the distribution of verb tokens across clause types (based on verbs with at least one finite token in the corpus, i.e. a subset of the verbs in Table 2).

Language	Corpus	Verb lemmas	Number of clauses			Corpus total
			Main-Fnt	Sub-Fnt	Non-Fnt	
German	VM	650	34,744	4407	10,728	49,879
Dutch	CGN	2212	91,481	26,183	41,635	159,299
English	SWB	1469	75,475	36,913	52,639	165,027
Grand totals		4331	201,700	67,503	105,002	374,205

Table 4. Verb lemmas with a total lemma frequency ≥ 20 (a subset of the numbers in Table 3). The first data column shows the number of unique lemmas; the other columns give the number of tokens (verbforms). The percentages indicate the clause-type distributions of all verbforms belonging to the lemmas with total lemma frequency ≥ 20 .

Language	Corpus	Verb lemmas	Number of clauses			Corpus total
			Main-Fnt	Sub-Fnt	Non-Fnt	
German	VM	171	33,836 71.0%	4077 8.6%	9758 20.5%	47,671
Dutch	CGN	437	89,015 59.2%	24,593 16.3%	36,828 24.5%	150,436
English	SWB	321	74,190 46.5%	35,925 22.5%	49,505 31%	159,620
Grand totals		929	197,041 55.1%	64,595 18.1%	96,091 26.9%	357,727

abovementioned “raw” clause-type proportions (the conditional probabilities), or we allow higher-frequency lemmas in the group to exert a stronger influence on the outcome than lemmas with lower frequencies. The latter we call “*weighted*” clause-type proportions (abbreviated WMain and WSub) which we obtain by multiplying the clause-type proportions of a verb by its normalised total frequency (NormTotFreq). Note that this weighting transformation affects group totals and group averages but leaves the clause-type distribution of the individual lemmas intact.

4. Results

In Section 2, we formulated three hypotheses which, in the terminology introduced above, read as follows. First, the Main-Fnt proportions of a verb lemma are expected to increase with increasing total lemma frequency (because main clauses assign early positions to their finite head). Second, no such increase (or only a small one) should occur in the Sub-Fnt proportions (because subordinate clauses are planned later than main clauses, meaning that early accessibility of finite verbs is not urgent). The combination of these hypotheses we call the MCB effect. Third, German and Dutch will exhibit larger MCB effect sizes than English (because they have differing word orders in main compared to subordinate clauses, whereas English has basically the same word order in both clause types).

Before testing these predictions statistically, we need to remove a methodological artefact from the dataset. Consider Figure 1, which shows the distribution of finite and nonfinite verb tokens across the three clause types as a function of the total frequency of verbs occurring with at least one finite form (cf. Table 3). The three corpora reveal the same overall pattern: clause-type curves in the form of inverted U-shapes. As our primary interest is the effect of total lemma frequency on the distribution of *finite* forms, we discarded all verb lemmas

whose occurrences were all nonfinite – a decision that more likely impacts on low-frequency than on high-frequency verbs, thus indirectly raising the proportion of finites in low-frequency verbs. The turning point in nearly all U-shapes is around a total lemma frequency of 20. In the statistical and theoretical analyses reported

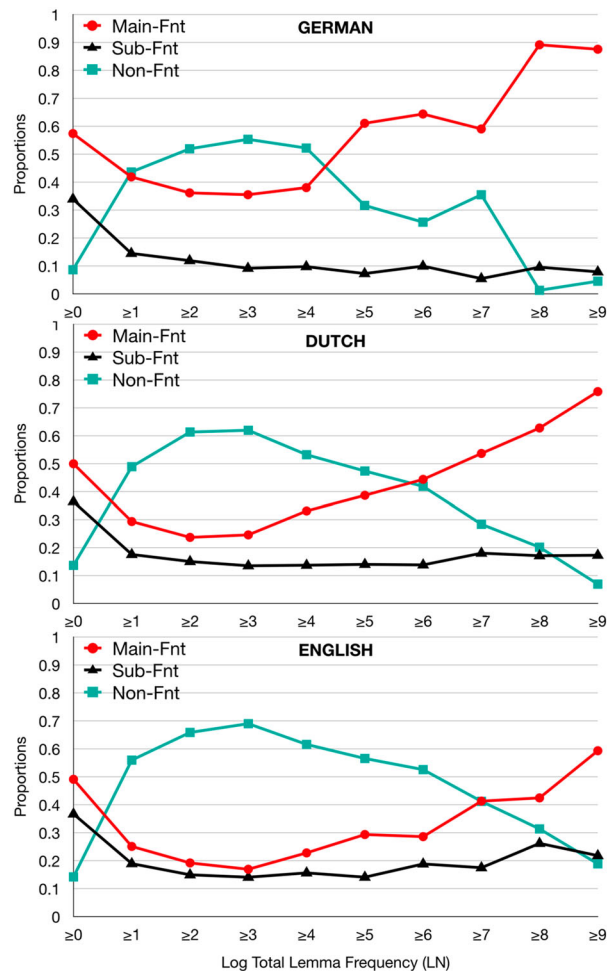


Figure 1. The unweighted clause-type distributions of verb lemmas as a function of the natural logarithm of their total frequency, for the three target languages. In each graph, the three proportions depicted above a given abscissa value add up to 1.

below, we disregard verbs with a total lemma frequency smaller than 20. The consequence is that the analyses will be based on 929 high- and mid-frequency verbs, which cover 21.5 percent of the verb *lemmas* in the corpora but 73.7 percent of the verbforms (*tokens*).

Application of these filters yields the scatter diagram of Figure 2, which clearly exhibits the hypothesised pattern: an increasing difference between the Main-Fnt and Sub-Fnt with increasing total lemma frequency (the MCB effect), and steeper slopes (see the trendlines) for the German and Dutch corpora than for the English one. The equations underlying the trendlines, and the corresponding R^2 values are listed in Table 5.

Figure 2 shows not only steeper MCB slopes for German and Dutch than for English, but also a bigger overall ratio of the Main-Fnt to the Sub-Fnt proportions. The ratios are 2.1 to 1 in the English, and 3.6 to 1 in the Dutch corpus. Although many factors unrelated to the current theory may have contributed to this difference (variability of the topic domains, corpus sizes, and conversational setting, for instance), we interpret the ratio difference as in line with the predicted difference regarding the probability of covert subordinate-to-main clause crossovers (higher in Dutch than in English). The much higher ratio in the German corpus (8.3 to 1) must be due to unrelated factors (see also next section).

In order to test the statistical significance of the predicted differences, we applied Beta Regression (Cribari-Neto & Zeileis, 2010),⁴ which assumes a beta-distributed dependent variable in the (0,1) range (i.e. proportions bigger than 0 and smaller than 1), using the Betareg software package in R (mean model with logit link; see Appendix C for details of the analyses). We tested models with three different dependent variables: (1) the weighted Main-Fnt proportions (called WMain in the Appendix), (2) the weighted Sub-Fnt proportions (WSub), and (3) the difference of the Main-Fnt and Sub-Fnt proportions (DiffWMainWSub). In all models, the independent variables were the normalised total frequency of the lemmas (NormTotFreq) and Language – the former entered as a continuous, the latter as a categorical predictor. In order to make sure that all Main-Fnt minus Sub-Fnt differences were greater than zero (as required by Beta Regression), we added 0.1 to each difference. (This raised the minimum, maximum and mean values of the frequency predictor to 0.1,

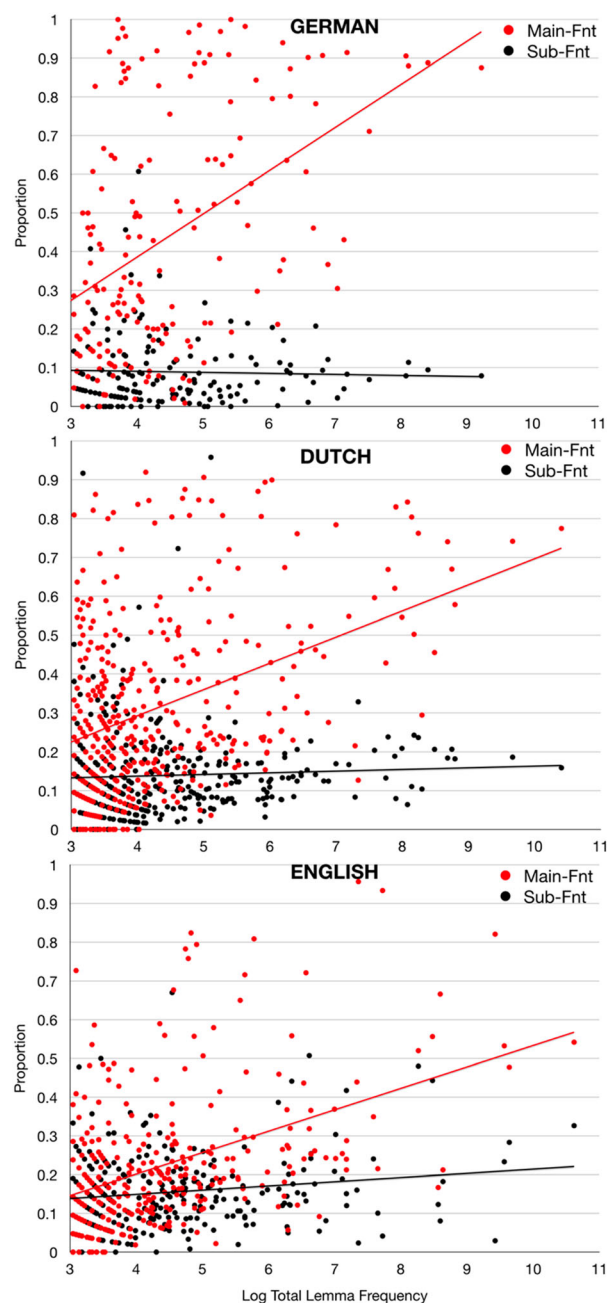


Figure 2. Clause-type distributions of finite verbs in the target languages (weighted proportions). Only verbs with a total lemma frequency of 20 or more (i.e. natural logarithm approximately 3) are shown. The equations underlying the trendlines, and corresponding R^2 values are in Table 5.

0.2697258, and 0.1014064, respectively. As for the two dependent variables: All zero values of WMain became

Table 5. Equations underlying the trendlines in Figure 2, and corresponding R^2 values.

	Main-Fnt		Sub-Fnt	
	Equation	R^2	Equation	R^2
German	$y = 0.1115x - 0.0602$	0.2013	$y = -0.0026x + 0.1013$	0.0013
Dutch	$y = 0.0673x + 0.0237$	0.1470	$y = 0.0043x + 0.1204$	0.0024
English	$y = 0.0553x - 0.0197$	0.1886	$y = 0.01080x + 0.1064$	0.0226

2.220446e-16; the maximum of WMain became 0.1865495, and the mean 0.001901277. The corresponding values for WSub: minimum 2.220446e-16; maximum 0.08365493; mean: 0.0005102989.)

In all three models, the main effects of NormTotFreq and Language were highly significant (all p -values $< .001$; see Appendix C). Here, we focus on the interactions between NormTotFreq and Language. With weighted Main-Fnt proportions as dependent variable (first Betareg model in Appendix C), the NormTotFreq values of Dutch and German verbs both had a stronger effect on the growth of the MCB effect than those of English verbs (both z -values > 3 , $p < .002$). However, the growth with NormTotFreq of the weighted Sub-Fnt proportions was not modulated by Language (second model; z -values between -1 and $+1$). These differential effects of Language on the growth rates of the Main-Fnt vs. Sub-Fnt proportions suggests a significant NormTotFreq * Language interaction when the difference between the clause-type proportions is the dependent variable. This expectation is confirmed in the third model (z -values > 30 , $p < .0001$).

The graphs in Figure 2 show that the Main-Fnt slope computed for the German data is steeper than the slope of its Dutch counterpart – an effect *not* predicted by the theory. We surmise that this contrast is somehow related to the narrower range of topics addressed by the German speakers, given the task assigned to them during the recorded sessions. (Each pair of speakers was asked to prepare a joint business trip. This accounts for frequent mention of verbs related to scheduling and travel (e.g. in Appendix A: *fahren* “drive”; *passen* “suit”; *ankommen* “arrive”; *zurückkommen* “return” *stattfinden* “take place”, etc. As a result, the number of different (unique) lemmas was small compared to the numbers in the other corpora, where the speakers usually talked without specific domain instructions.)

To sum up, the Betareg analyses confirm that the differences entailed by the three hypotheses put forward above, are statistically significant: The overuse of high-frequency verbs is largely restricted to main clauses (the MCB effect), and the extent of this overuse – as measured by the difference between weighted Main-Fnt and Sub-Fnt proportions – is more prominent in German and in Dutch than in English.

5. Discussion

We have attributed the main-clause bias (MCB) effect and the cross-language difference of its effect size to two different manifestations of what we have called the *accessibility–anteriority link*. First, speakers of the

Germanic language that we investigated, tend to overuse high-frequency finite verbs in the main clauses they produce. We found hardly any overuse of such verbs in finite subordinate clauses, meaning that a model based on “Easy First” alone is inadequate. The crucial factor seems to be time pressure: Finite verbs are demanded at anterior positions in main clauses, and high-frequency verbs can meet this demand more readily than lower-frequency verbs. For finite subordinate clauses, whose planning is likely to lag behind main-clause planning, more planning time is available. This eliminates much of the time pressure and allows less frequent but perhaps more appropriate verbs to conquer a clause position.

The second manner in which we suggest the accessibility–anteriority link can become manifest, is by biasing the format/status of a clause-under-construction from subordinate to main. We argue that covert subordinate-to-main crossovers occur in all three target languages, conditionally upon pragmatic factors such as the type of illocutionary force associated with the clause, and on this clause being syntactically licensed to follow the topmost main clause. When these conditions are met, *competition* arises between main- and subordinate-clause formats, with the format that assigns an earlier position to the finite verb surfacing as the likely winner. Given the different SVO/SOV word order patterns in the target languages, we argue that the MCB effect should be larger in German and Dutch than in English – a prediction that is statistically confirmed.

Is the proposed explanation of the MCB patterns in the three languages the most parsimonious one? The current interpretation of the MCB effect is based on mutual attraction by high-frequency verbs and clauses with finite verbs in early positions. In addition, we have assumed that subordinate clauses, which tend to be planned later than main clauses, are immune to time pressure arising if the clause needs a finite verb in an early position. However, the latter assumption can perhaps be dispensed with if a more central role can be assigned to covert crossovers. Consider the possibility that *all* finite clauses attract high-frequency verbs (with a force depending on the earliness of the head verb); that is, overuse of high-frequency verbs occurs not only in main but also in finite subordinate clauses. However, only the latter clauses may undergo covert subordinate-to-main crossovers due to competition between two clauses formats. (Such competition does not affect clauses at the top of the clause hierarchy because at that planning level main-clause format/status is the only option.) If this assumption is correct, we could stipulate that the observed (near-)absence of overuse in finite

subordinate clauses is due to covert crossovers, which leads to flatter slopes in the Sub-Fnt proportions and simultaneously to steeper slopes in the Main-Fnt slopes. Although this alternative account seems more parsimonious than the one proposed above, it fails to explain why the slope of the proportions in Sub-Fnt clauses is (close to) zero rather than just flatter than the slope of the Main-Fnt proportions. Mathematical modelling may be able to decide whether the simpler account can fit the data patterns in the three languages.

In the remainder of the present section, we explore relations between the MCB findings and other phenomena discussed in the Introduction (Easy-First phenomena, Uniform Information Density), and propose topics for further investigation.

But first we need to justify why we did not go into factors causing some verbs to be more frequent than others. The reason is that we view the accessibility-anteriority link as the “proximal cause” of the MCB, and that a variety of factors underlying accessibility levels are “distant causes”. Most verb frequency differences presumably reflect conceptualisation frequencies – how often the conceptual content that drives lexical access and retrieval, is “on the speaker’s mind”. Another factor is the verb’s multifunctionality: whether it can be used in multifarious pragmatic, semantic or syntactic contexts. Among the reasons why certain verbs are used in main rather than subordinate clauses are pragmatic and cognitive/communicative factors such as propositional attitude, evidentiality, and epistemicity. (See also the verb listings in Appendix B.)

We count the MCB effect as an Easy-First phenomenon, but one with a special touch: It does not increase the likelihood of selecting the earlier member of a set of placement options open for a clause constituent; instead, it boosts the probability for high-frequency lexical items to select the sole placement option for that constituent (here, the fixed clause position of the finite verb). Nonetheless, covert subordinate-to-main crossovers may be compared to the rather “drastic” Easy-First variant that involves changing the voice of a clause from active to passive. In both cases, a late constituent receives a much earlier position: SOV becomes SVO in the former case; in the latter, a late object becomes an early subject (and the transformation usually introduces a high-frequency finite auxiliary verb appearing at an early position).

The accessibility-anteriority link we hold responsible for the MCB also has the effect of promoting uniform information density (UID). At the onset of planning a mono- or pluriclausal sentence, the speaker has to decide on values for a large number of parameters that together make up the shape of the upcoming utterance.

(Incremental production can alleviate this task to a considerable extent.) In the course of this planning process, the number of degrees of freedom decreases gradually (often with interrupts at the transitions between finite clauses). In languages where a clause hierarchy is planned top-down, main clauses are usually planned prior to subordinate clauses, as supported by the fact that the majority of main clauses precedes finite subordinate clauses in the surface structure of the sentence. If indeed the peak of the cognitive processing capacity mounted by the planning process thus tends to affect main clauses more heavily than subordinate clauses, the speaker can help to prevent overload by choosing easily accessible verbs (and other “easy” clause constituents with early placement options). The result will be a less skewed distribution of planning capacity recruited across the clause hierarchy.

In information-theoretical terms, one therefore expects the predictability (uninformativity) of the finite verb to be higher, on average, in main clauses than in subordinate clauses. In order to estimate how predictable a finite verb v is in the given clause type, we took an approach similar in spirit to one popular in work on predicting properties of speech sounds (Cohen Priva, 2008; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 2011; Seyfarth, 2014). It is based on the conditional probabilities $\text{Prob}(v|\text{Main-Fnt})$ and $\text{Prob}(v|\text{Sub-Fnt})$. The surprisal value (*Shannon information*) associated with these probabilities can be approximated by their logarithmic transforms $\text{LN}(1/\text{Prob}(v|\text{Main-Fnt}))$ and $\text{LN}(1/\text{Prob}(v|\text{Sub-Fnt}))$, where LN denotes the natural logarithm. The resulting values are weighted by the verb’s Main-Fnt and Sub-Fnt proportions: $\text{Prob}(\text{Main-Fnt}|v)$ and $\text{Prob}(\text{Sub-Fnt}|v)$, respectively. Within each clause type, the resulting products (i.e. the predictability estimates) tend to be high for verbs whose presence in the clause type yields low surprisal values (being often encountered “inhabitants” of that clause type compared to rare inhabitants), and to have a stronger attraction to that clause type (relative to other clause types). In each treebank, the average predictability scores thus calculated for verbs in main clauses are higher than those for subordinate clauses (2.67 and 0.53 for German; 2.35 and 1.04 for Dutch; 1.67 and 1.16 for English): This result (unsurprising, given the results discussed in the previous section) shows that verbs in posterior clauses (subordinate) tend to be less predictable than those in more anterior clauses (main). This is in line with UID on the presupposition that *non-lexical* aspects of planning at sentence onset are rather unpredictable.

However, the latter assumption seems at variance with the results of a recent corpus study by Temperley

(2019) with written newspaper English (Wall Street Journal). This study shows that, as summarised in the title, “rare constructions are more often sentence-initial”. The rare constructions investigated by Temperley belong to the class of “Main Clause Phenomena”. Examples are participle preposing (*Standing next to me was the president of the company*); preposing around *be* (*More significant would be the development of a semantic theory*); locative inversion (*On the wall hangs a portrait of Mao*), and topicalization (*This book you should read*). His explanation of the predominantly sentence-initial position of such constructions (as opposed to rarer occurrences in non-sentence-initial positions, e.g. in complement clauses) is based on the assumption that

syntactic processing at the very beginning of a sentence [is] easier (requiring less computation) than elsewhere. It therefore makes sense that a language might evolve to allow additional syntactic possibilities at the beginning of the sentence; this causes additional processing complexity, but this is counterbalanced by the reduction in processing load due to the absence of previous syntactic context. (Temperley, 2019, p. 6)

Temperley continues this passage with remarking that his account is in line with the UID hypothesis.

Can Temperley’s interpretation be reconciled with our account of the MCB effect? We suggest lexical rather than syntactic factors could be responsible for the lower processing complexity of sentence-initial compared to later clauses (due to higher-frequency verbs and other parts of speech – the latter to be checked in the corpus). We expect that while planning progresses, syntactic planning problems decrease, on average, due to preactivation of suitable syntactic alternatives. An obvious alternative attempt at resolving the conflicting interpretations could be based on the fact that Temperley’s study used written rather than spoken corpus materials. Such an attempt is likely to fail, though, because we found an MCB in the Wall Street Journal corpus as well, be it with a smaller effect size than in the Switchboard corpus consulted in the present study (cf. Kempen & Harbusch, 2017).

As for future empirical work, if our theoretical account of the MCB and the observed cross-language variability is correct, it entails predictions regarding MCB effects and effect sizes in other languages. We expect moderate MCB effects, comparable to English, in other Germanic languages, given they are largely SVO in both main and subordinate clauses. No MCB is expected in languages such as Japanese where verbs are final in all clause types, and main clauses follow rather than precede subordinate clauses. The advantage of early available verbs is thus eliminated. (Note that Japanese does exhibit intra-clausal Easy-First effects of the type we described in Section 1; see Lohmann & Takada, 2014; Tachihara,

Pitcher, & Goldberg, 2019.) Empirical studies of MCB effects in different language families (in particular, studies comparing SVO and SOV languages) can provide new insights into the planning process from which multiclausal sentences originate. One topic we find worthwhile pursuing is whether strongly head-final (SOV) languages favour bottom-up planning of clause hierarchies, in contrast to the predominantly top-down course we have assumed for SVO languages, and if so, how this difference interacts with other aspects of grammatical encoding.

Notes

1. We realise that the term “postdiction” is more in line with history: The data pattern came first, and the interpretation is *post hoc*.
2. The databases we consulted are the following. For all three languages: CELEX (Baayen, Piepenbrock, & Gulikers, 1995). In addition, we used the morphological software MORPHY (Lezius, 2000), and lemmatisations from these written treebanks: TIGER (Brants et al., 2004), TüBa-D/Z (Telljohann, Hinrichs, Kübler, & Kübler, 2004), and ALPINO (van der Beek, Bouma, Malouf, & van Noord, 2002). The in-house software mentioned here and elsewhere in the paper is available on request from the second author.
3. This resulted in the removal of – mostly low-frequency – verbs: 440 German lemmas (with a total of 845 verbform tokens); 1689 Dutch lemmas (3749 tokens), and 1143 English lemmas (2349 tokens). We verified (using the statistical methods of the next section) that their removal did not alter essential aspects of the data patterns and significance levels.
4. Cribari-Neto and Zeileis (2010, p. 1) introduce Beta Regression as follows:

[Beta Regression] is based on the assumption that the dependent variable is beta-distributed and that its mean is related to a set of regressors through a linear predictor with unknown coefficients and a link function. The model also includes a precision parameter which may be constant or depend on a (potentially different) set of regressors through a link function as well. This approach naturally incorporates features such as heteroskedasticity or skewness which are commonly observed in data taking values in the standard unit interval, such as rates or proportions.

See Mangiafico (2016) for considerations regarding the selection of appropriate R packages.

Acknowledgements

The editors and three reviewers deserve many thanks for detailed comments and very useful advice. We are particularly indebted to Phillip Alday for pointing us to Beta Regression and for guidance on its application. Of course, the responsibility for the final content is entirely ours.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (release 2.5, CD-ROM). Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.
- Berg, T. (2018). Frequency and serial order. *Linguistics*, 56, 1303–1351.
- Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21, 47–67.
- Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., ... Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2, 597–620.
- Cohen Priva, U. (2008). Using information content to predict phone deletion. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th west coast conference on formal linguistics* (pp. 90–98). Somerville, MA.
- Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2). Retrieved from <https://www.jstatsoft.org/article/view/v034i02>
- Dipper, S., & Kübler, S. (2017). German treebanks: TIGER and TüBa-D/Z. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 595–639). Berlin: Springer.
- Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics*, 27, 517–556.
- Ford, M., & Holmes, V. M. (1978). Planning units in sentence production. *Cognition*, 6, 35–53.
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the international conference on audio, speech and signal processing (ICASSP-92)* (pp. 517–520).
- Harley, T. (2014). *The psychology of language: From data to theory* (4th ed.). London: Psychology Press.
- Hoekstra, H., Moortgat, M., Schuurman, I., & van der Wouden, T. (2001). Syntactic annotation for the spoken Dutch corpus project (CGN). *Language and Computers*, 37(1), 73–87.
- Hwang, H., & Kaiser, E. (2015). Accessibility effects on production vary cross-linguistically: Evidence from English and Korean. *Journal of Memory and Language*, 84, 190–204.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61, 23–62.
- Kempen, G., & Harbusch, K. (2003). Word order scrambling as a consequence of incremental sentence production. In H. Härtl & H. Tappe (Eds.), *Mediating between concepts and grammar* (pp. 141–164). Berlin: Mouton De Gruyter.
- Kempen, G., & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann & C. Habel (Eds.), *Multidisciplinary approaches to language production* (pp. 173–181). Berlin: Mouton De Gruyter.
- Kempen, G., & Harbusch, K. (2017). Sequential test of (S)OV as unmarked word order in Dutch and German clauses: A serendipitous corpus-linguistic experiment. In H. Reckman, L. L. S. Cheng, M. Hijzelendoorn, & R. Sybesma (Eds.), *Crossroads semantics. Computation, experiment and grammar* (pp. 107–123). Amsterdam: Benjamins.
- Kempen, G., & Harbusch, K. (2018). A competitive mechanism selecting verb-second versus verb-final word order in causative and argumentative clauses of spoken Dutch: A corpus-linguistic study. *Language Sciences*, 69, 30–42. doi:10.1016/j.langsci.2018.05.005
- König, E., & Lezius, W. (2003). *The TIGER language: A description language for syntax graphs, formal definition*. Stuttgart: Technical Report, IMS, University of Stuttgart.
- Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
- Leech, G., Rayson, P., & Wilson, A. (2014). *Word frequencies in written and spoken English based on the British national corpus*. Oxford: Routledge. [First edition published in 2001].
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS, Vol. 19)*. Cambridge, MA: MIT Press.
- Lezius, W. (2000). Morphy – German morphology, part-of-speech tagging and applications. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the 9th EURALEX international congress* (pp. 619–623). Stuttgart.
- Lohmann, A., & Takada, T. (2014). Order in NP conjuncts in spoken English and Japanese. *Lingua*, 152, 48–64.
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226. doi:10.3389/fpsyg.2013.00226
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126, 313–318.
- Mangiafico, S. S. (2016). *Summary and analysis of extension program evaluation in R*. Version 1.18.1. Retrieved from <http://rcompanion.org/documents/RHandbookProgramEvaluation.pdf>, http://rcompanion.org/handbook/J_02.html
- Osgood, C. E., & Bock, J. K. (1977). Salience and sentencing: Some production principles. In S. Rosenberg (Ed.), *Sentence production: Developments in research and theory* (pp. 89–140). Hillsdale, NJ: Erlbaum.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108, 3526–3529.
- Sauppe, S. (2017). Word order and voice influence the timing of verb planning in German sentence production. *Frontiers in Psychology*, 8, 1648. doi:10.3389/fpsyg.2017.01648
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133, 140–155.
- Stegmann, R., Telljohann, H., & Hinrichs, E. W. (2000). *Stylebook for the German treebank in Verbmobil*. Saarbrücken: DFKI Report 239.
- Tachihara, K., Pitcher, M., & Goldberg, A. E. (2019). *Jessie and Gary or Gary and Jessie?: Cognitive accessibility predicts order in English and Japanese*. Proceedings of the 41st Annual Meeting of the Cognitive Science Society, Montreal.
- Telljohann, H., Hinrichs, E., & Kübler, S. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)*.

- Temperley, D. (2019). Rare constructions are more often sentence-initial. *Cognitive Science*, 43, e12714. doi:10.1111/cogs.12714
- Vallduvi, E. (1992). *The informational component*. New York: Garland.
- van der Beek, L., Bouma, G., Malouf, R., & van Noord, G. (2002). The Alpino dependency treebank. In T. Gaustad (Ed.), *Computational linguistics in the Netherlands 2001* (pp. 8–22). Amsterdam: Rodopi.
- van der Beek, L., Bouma, G., & van Noord, G. (2002). Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7, 353–374.
- van Eerten, L. (2007). Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 12, 194–215.
- Vogels, J., Krahmer, E. I., & Maes, A. (2019). Accessibility and reference production: The interplay between linguistic and non-linguistic factors. In J. Gundel & B. Abbott (Eds.), *The Oxford handbook of reference* (pp. 337–364). Oxford: OUP.
- Wahlster, W. (Ed.). (2000). *Verbmobil: Foundations of speech-to-speech translation*. Berlin: Springer. doi:10.1007/978-3-662-04230-4