GLOSSA psycholinguistics

Cortés Rodriguez, Á., & Felser, C. (2026). Processing efficiency constraints on relative clause extraposition in German: Evidence from speeded judgments and production. *Glossa Psycholinguistics*, 5(1): 15, pp. 1–32. DOI: https://doi.org/10.5070/G6011.61343

eScholarship University of California

Processing efficiency constraints on relative clause extraposition in German: Evidence from speeded judgments and production

Álvaro Cortés Rodríguez, Independent Researcher, acr.academic@gmail.com

Claudia Felser, University of Potsdam, felser@uni-potsdam.de

Results from previous corpus and untimed judgment studies indicate that the distribution and acceptability of relative clause extraposition (RCE) in German are modulated both by the extraposition distance and the length of the extraposed relative clause. The observed complexity effects have been argued to reflect processing efficiency constraints, such as a short-before-long ordering preference or a bias towards optimizing dependency lengths within a sentence. However, it is still unclear whether such processing-related biases affect the perception and production of RCE in the same way or differently, and how complexity measures interact with other factors that are known to modulate processing difficulty, such as relative clause expectancy. The present study used speeded acceptability judgment and timed sentence generation tasks to examine the processing and production of RCE in German. Our results show that complexity measures affect RCE perception and production in a similar way, with smaller VP-to-RC length ratios linked to higher RCE acceptability and increased production likelihood. Relative clause expectancy affected participants' response times, but not RCE acceptability. Despite the high overall acceptability of RCE in German, even long relative clauses were preferably positioned adjacent to their head noun in our production task, however. This suggests that a bias towards minimizing the distance between an RC and its head noun, rather than a short-before-long ordering preference, is the primary constraint on German speakers' production choices.

1. Introduction

Word-order preferences, including the supposedly universal tendency towards keeping the distance between syntactically dependent constituents short (e.g., Futrell et al., 2015; Futrell et al., 2020; Gildea & Temperley, 2010; Liu et al., 2017; Temperley & Gildea, 2018; Yadav et al., 2022), have often been argued to reflect processing economy constraints (e.g., Hawkins, 2004, 2014).

In languages such as English or German, placing dependent clauses or heavy modifiers to the right periphery of a clause or sentence can help optimize the lengths of syntactic dependencies within the sentence. In (1a) below, for example, the bracketed relative clause (RC) modifying the noun phrase (NP) *the woman* has been extraposed to the right. From the perspective of incremental processing, this allows the subject-predicate relationship in the main clause to be established prior to processing the RC. In the corresponding integrated word-order variant (1b), in contrast, the subject-verb dependency can only be completed once the intervening RC has been processed.

(1) a. Last night a woman called [$_{RC}$ who wanted to sell an insurance policy].
b. Last night a woman [$_{RC}$ who wanted to sell an insurance policy] called.

While relative clause extraposition (RCE) in sentences such as (1a) serves to minimize the distance between the head noun of the subject and the verb, it increases the distance between the RC and its head noun. We might thus expect speakers to prefer word orders that allow the relevant dependency lengths to be optimized, or which reflect some kind of compromise between minimizing verb phrase size and minimizing extraposition distance, as proposed by (Hawkins, 2004, 2014). According to Hawkins' (2004) principle of Domain Minimization, the degree to which RCE is preferred over integrated RCs can be predicted from the degree to which RC length exceeds extraposition distance or VP length. RCE is expected to be facilitated by a small VP-to-RC length ratio, i.e., for sentences containing short VPs and long RCs. The factor VP-to-RC length ratio was indeed identified as the best predictor for the occurrence of RCE in English corpora (Francis, 2010; Francis & Michaelis, 2014). In German, the likelihood of an RC being extraposed, or of an extraposed RC being deemed acceptable, is also affected both by an RC's length and its extraposition distance (e.g., Gosemann et al., in press; Konieczny, 2000; Strunk, 2014; Uszkoreit et al., 1998). The derived factor VP-to-RC length has not previously been statistically examined for German, however.

Most of the evidence in support of processing efficiency constraints affecting clausal extraposition comes from written corpora and offline judgment data, neither of which provides a direct measure of processing difficulty or efficiency, however. As noted by Gosemann et al. (in press), RCE distribution frequencies in corpus data might be affected by stylistic genre

norms, discourse-pragmatic factors, and audience design considerations. Although judgment tasks are primarily comprehension-based, offline judgments may be affected by metalinguistic or stylistic considerations and may also be influenced by production-related constraints, due to speakers' comparing different word-order alternatives in their mind before providing a judgment (Konieczny, 2000).

As the results from corpus and experimental studies do not necessarily align (Francis, 2010; Gosemann et al., in press; Uszkoreit et al., 1998), it is still unclear whether complexity measures such as VP length or extraposition distance, RC length, or VP-to-RC length ratio affect sentence production and comprehension in comparable ways. One possibility is that the choice for or against RCE is primarily determined by production-related constraints. According to Wasow's (2002) Principle of End Weight (PEW), short or simple phrases tend to be placed before longer or more complex phrases. This short-before-long preference has been argued to reflect sentence planning and production economy constraints (Arnold et al., 2000; MacDonald, 2013; Stallings et al., 1998; Wasow, 1997).

As maintaining unresolved dependencies across long distances can be cognitively demanding, minimizing or optimizing dependency lengths has also been argued to benefit comprehension (e.g., Gibson, 1998). It is conceivable, however, that word-order variants which are easy to comprehend are not necessarily easy to produce, or vice versa. Building on and extending previous research on the role of processing efficiency constraints on RCE in German, we compare RCE production and perception, using experimental tasks that seek to minimize the influence of unwanted top-down factors, such as stylistic or audience design considerations.

Another poorly understood issue is how complexity metrics interact with other processing-related factors, such as RC expectancy (Levy et al., 2012). It is conceivable, for example, that the tendency for RCE to be judged as less acceptable than its corpus frequency suggests (Gosemann et al., in press; Konieczny, 2000; Uszkoreit et al., 1998) reflects comprehenders' difficulty integrating an unexpected extraposed RC. Here, we investigate the above issues through speeded acceptability judgment and timed sentence production tasks. Before presenting our experiments, we briefly review previous studies reporting effects of complexity measures and expectancy on RCE, focusing on German.

2. Effects of complexity and expectancy on RCE in German

While German noun phrases are predominantly right-branching, German is a verb-final language with relatively free word order. Since finite verbs must be placed in the second structural (V2) position in main clauses, composite tense forms or verbs containing separable prefixes will normally be linearized discontinuously. In both (2a) and (2b) below, the finite auxiliary *hat* 'has' and its associated participle *geklingelt* 'rung' form a so-called "verbal bracket". While the RC

modifying the subject noun phrase *das Telefon* 'the phone' appears adjacent to its host NP within the verbal bracket in example (2a), it has been extraposed beyond the right verbal bracket in example (2b).[1]

(2) a. *Wieder* ***hat*** *das Telefon, [$_{RC}$ das defekt war], nicht* ***geklingelt.***
again has the phone that broken was not rung

b. *Wieder* ***hat*** *das Telefon nicht* ***geklingelt,*** *[$_{RC}$ das defekt war].*
again has the phone not rung that broken was
Both: 'The phone which was broken did not ring again.'

While RCE, as in (2b), serves to decrease the size of the verbal bracket, thus allowing for the verb phrase to be completed earlier, it increases the distance between the dependent clause and its head noun. As we will outline in the following subsections, it is conceivable that different length or complexity measures affect RCE processing and production in different ways.

2.1 Production

The extraposition of long or complex modifiers is thought to reflect a general bias towards positioning long constituents after shorter ones, which has been argued to facilitate sentence planning and production (MacDonald, 2013; Wasow, 1997, 2002). Placing a heavy modifier outside the verbal bracket in German will allow for the verb phrase's main constituents to be planned and produced early, while the planning and production of complex or non-essential constituents can be delayed.

The influence of complexity on the production likelihood of RCE in German has been examined in several corpus studies. Besides revealing relatively high RCE occurrence frequencies, these studies have shown that the likelihood of RCE production is modulated both by RC length and by the extraposition distance (Gosemann et al., in press; Shannon, 1992; Strunk, 2014; Uszkoreit et al., 1998).[2] Uszkoreit et al. (1998) and Strunk (2014) found RCE to be favored over integrated RCs across short extraposition distances, but not across longer extraposition distances, for example. Using a large newspaper corpus that covered a period of almost 50 years (1970-2018), Gosemann et al. (in press) found that 74.6% of all RCs whose antecedent was located within the verbal bracket were extraposed. RCE likelihood decreased with increasing extraposition distance, but increased with increasing RC length. RCE was most likely if extraposition crossed a single word only, and VP length affected RCE likelihood only for longer extraposition distances. Given that the production of written texts may involve several stages of revision and editing, and may be

[1] Note that in German, both restrictive and nonrestrictive integrated RCs must be separated from their antecedents by a comma.

[2] The occurrence of RCE in German corpora is also affected by factors other than word-based length or distance measures which are not investigated here (see Strunk, 2014, p. 98).

influenced by stylistic or audience design considerations, it is unclear to what extent data from written corpora reflect sentence planning or economy-of-production constraints, however.

RCE production has also been investigated experimentally. Bader (2014) reports the results from two oral production experiments, which showed that five-word-long object-modifying RCs were extraposed between 11%–54% of the time, depending on how much and what kind of sentence material intervened between the RC and its host NP. In another oral production task, Bader (2024) found that the likelihood of speakers producing an extraposed RC was (weakly) affected by information-structural factors, with 47% of RCs being extraposed in a topic context, and 56% in a focus context. The extraposition distance was always minimal (one word only), though.

Using sentences with subject-modifying RCs, Gosemann et al. (in press) carried out an untimed written production task manipulating both VP length and RC length. Participants were asked to indicate their preferred fragment orderings by typing a sequence of numbers. The results showed that integration was strongly preferred over extraposition. Even for the shortest possible extraposition distance of one word, fewer than 12% of the RCs that participants produced were extraposed. While RC extraposition was more likely for shorter VPs than for longer VPs, the length of an RC did not affect extraposition likelihood. Together, these findings suggest that RC integration tends to be favored over extraposition in elicited production tasks.

Note, however, that it is unclear to what extent previous production tasks were able to tap into sentence planning processes. In both Bader's (2014, 2024) and Gosemann et al.'s (in press) experiments, the RCs to be produced were already part of the stimulus materials and, thus, did not have to be generated by the participants themselves. Bader (2014) presented complex sentences that participants then had to repeat from memory, and Bader (2024) and Gosemann et al. (in press) used sentence anagram tasks in which participants previewed and then re-ordered a set of sentence fragments. No time pressure was applied in any of these tasks, and potential effects of dependency length optimization were not examined.

2.2 Judgment and comprehension

Similar to what has been reported in corpus studies, the acceptability of RCE is also influenced by length-based complexity measures, notably, by extraposition distance and RC length. Using a magnitude estimation acceptability judgment task, Uszkoreit et al. (1998) found that extraposed RCs were generally considered less acceptable than integrated RCs, except for long RCs that were separated from their host NP by one word only. Also using magnitude estimation, Konieczny (2000) reports similar findings. The acceptability of RCE was negatively affected by increasing extraposition distance, and sentences containing short extraposed RCs were judged as significantly worse than sentences containing longer ones. Using an offline scalar acceptability judgment task, Gosemann et al. (in press) confirmed the detrimental effect of longer extraposition distances on

RCE acceptability, but did not investigate potential effects of RC length or of VP-to-RC length ratio.

These findings are in line with theoretical proposals that predict grammatical weight and the distance between dependent constituents to interact, such as Hawkins' (2004, 2014) principle of Domain Minimization. Few studies have used processing tasks to examine how complexity measures affect the online comprehension of sentences with extraposed RCs, however. Konieczny (2000) reports the results from a word-by-word self-paced reading experiment investigating how RC position, RC length, and RC extraposition distance affect local processing difficulty. However, participants' reading times at the matrix verb were not modulated by either RC length or extraposition distance, leading Konieczny to speculate that dependency length might primarily affect processing cost in language production, rather than comprehension. Hawkins (2014, p. 55f.) also considers the possibility that complexity factors might primarily guide speakers' production choices, but points out that slower or faster processing of individual words does not necessarily bear on the validity of global processing-efficiency constraints such as his Domain Minimization principle.

From the perspective of incremental sentence comprehension, extraposition may give rise to added processing cost, in comparison to integrated word orders. Encountering an unexpected RC should trigger a backwards-looking search for a suitable host. Identifying the RC host is likely to become harder with increasing extraposition distance, and/or with an increasing number of intervening discourse referents (Bader, 2014). Given that language comprehension involves generating predictions about upcoming sentence material (Huettig, 2015), linking an extraposed constituent to its licensing head might be easier if the extraposed constituent is, in fact, expected. For English, Levy et al. (2012) found that, while sentences with RCE were more difficult to process than sentences with integrated RCs, this difficulty was neutralized if the RC was predictable.

For German, evidence for the role of RC expectancy comes from an offline acceptability judgment task conducted by Strunk (2014). Strunk observed that RCE is facilitated by antecedent NPs introduced by a cataphoric demonstrative determiner, such as *diejenige* 'the one' in (3) (adapted from Strunk, 2014, p. 103).

(3) *Jens hat diejenige Musikerin ausgelacht, die das einfache Stück nicht spielen konnte.*
Jens has the.one musician laughed.at who the easy piece not play could
'Jens laughed at the musician who couldn't play the easy piece.'

Strunk (2014) found an interaction between extraposition distance and determiner type, showing that the use of a cataphoric demonstrative – as opposed to a definite – antecedent was effectively able to neutralize the negative effect of a longer extraposition distance on RCE acceptability. This finding might be attributed to the fact that a cataphoric determiner triggers a strong expectation that a relative clause will follow, thus reducing the degree of surprisal associated

with encountering a non-integrated RC (compare Levy et al., 2012). Little is known about how RC expectancy interacts with complexity measures in processing tasks, however.

3. The present study

We conducted two controlled experiments to examine how length-based complexity measures affect German speakers' perception and production of RCE, and how RC expectancy interacts with complexity measures in the processing of RCE. Our study builds on and extends previous studies by asking the following research questions:

i. Do extraposition distance (which, in the current study, corresponds to VP length), RC length, and/or VP-to-RC length ratio affect the acceptability and production likelihood of RCE under time pressure in comparable ways?
ii. How does RC expectancy modulate the effects of length-based complexity measures on speakers' online judgments?

In both experiments, we used a wider-than-usual range of VP length and RC length combinations, and the factor VP-to-RC length ratio is examined for the first time for RCE in German. VP length and RC length were counted in words, and the derived complexity measure of VP-to-RC length ratio was obtained by dividing VP length by RC length.

Treating VP-to-RC length ratio as a single predictor differs from modeling VP length and RC length separately: it embodies the idea that what matters is the balance between the two domains, rather than the absolute size of either constituent on its own (Hawkins, 2004). Under this view, sentences with the same ratio but different absolute lengths are expected to pattern alike, whereas accounts such as the PEW would predict an effect of RC length over and above any influence of VP length. If Hawkins' (2004) Domain Minimization principle applies equally to sentence perception and production, we expect VP-to-RC length ratio to be the best predictor for both RCE acceptability and RCE production likelihood.

Note that RC predictability is more likely to affect sentence comprehension than production, given that identifying the head noun of a distant modifier is not an issue in production. We thus manipulated predictability in our speeded acceptability judgment experiment only. Our judgment experiment (Experiment 1) will be presented first, followed by our sentence generation experiment (Experiment 2).

3.1 Experiment 1: Speeded acceptability judgments

Our first experiment tests the prediction made by Hawkins' (2004) Domain Minimization principle and the potential effects of RC expectancy on the acceptability of RCE in German. As offline judgments may be influenced by prescriptive or stylistic considerations and may also

involve aspects of production (Konieczny, 2000), we used a speeded judgment task to minimize the likelihood of participants comparing different word-order options before making a decision. Timed or speeded judgment tasks are frequently used to obtain intuitive, unreflected judgments (Godfroid et al., 2015), with word-by-word presentation forcing readers to process stimulus items incrementally. According to Blackwell et al. (1996), speeded judgment tasks tap into the same cognitive processes as do incremental processing tasks. While we were primarily interested in the factors influencing participants' judgments, we also analyzed their response times as a potential indicator of relative processing difficulty.

We expect to replicate previous findings from offline judgment tasks showing that RCE acceptability is influenced both by VP length and RC length. According to Hawkins' (2004) Domain Minimization principle, VP-to-RC length ratio should be a better predictor for RCE acceptability than either of the two individual length measures. Regarding the role of RC expectancy, surprisal theory (Levy, 2008) predicts that RCs which are expected are easier to process and to integrate, compared to those which are not, as was previously shown in Levy et al.'s (2012) reading-time study on RCE in English.

3.1.1 Method

Participants

We recruited 53 adult native speakers of German from the University's participant pool. They all had normal or corrected-to-normal vision and did not report any language-related or other behavioral or neurological disorders. All participants provided informed, voluntary written consent prior to participating in our experiment. They were naïve with respect to the ultimate purpose of the experiment and received either course credit or a small monetary compensation of 5€ for their participation.

Prior to the statistical analysis, we excluded the data from seven participants who had responded incorrectly to half or more of our attention-checking items (see below), and the data from one participant who judged more than 75% of our ungrammatical or unacceptable fillers as acceptable. Two duplicated data sets were also removed. The data from the remaining 43 participants (34 female, 9 male; mean age: 23.55 (4.98); range: 18–40) were included in the analysis.

Materials

All experimental sentences included subject-modifying, extraposed subject RCs that were positioned immediately after the right verbal bracket. We manipulated three independent variables: (i) the type of DETERMINER introducing the RC head noun (definite vs. demonstrative), (ii) RC LENGTH (5–12 words), and (iii) VP LENGTH (3–9 words). The RC extraposition distance always equaled VP LENGTH minus one word. Both RC LENGTH and VP LENGTH represent continuous variables and were incorporated in a fully crossed design. All possible combinations of the three factors yielded

112 experimental sentences.[3] We also calculated VP-to-RC length ratio (henceforth, LENGTH RATIO) as a derived complexity measure by dividing VP LENGTH by RC LENGTH. Two example items representing opposite ends of the RC and VP length scales are shown in (4) and (5).

(4) *Neulich ist der Helikopter bei einem Trainingsmanöver aufgrund starker Schneefälle*
recently is the helicopter in a.DAT training_maneuver due_to heavy snowfall
abgestürzt, der im Skigebiet eingesetzt wurde.
crashed which in.the ski_resort used was
'Recently, the helicopter that was used at the ski resort crashed during a training maneuver due to heavy snowfall.'
[VP Length = 9; RC Length = 5; Length Ratio: 1.8]

(5) *Bedauerlicherweise ist derjenige Tierpfleger gestorben, der sich allein und unvorsichtig*
unfortunately is the.one zookeeper died who REFL alone and incautiously
im Gehege der Tiger und Löwen bewegte.
in.the compound of.the tigers and lions moved
'Unfortunately, the zookeeper died who moved alone and incautiously in the tigers and lions' compound.'
[VP Length = 3; RC Length = 12; Length Ratio: 0.25]

Example (4) includes a short RC of five words and a long VP of nine words, yielding a length ratio of 1.8, which represents the highest length ratio in our materials set. In contrast, example (5) includes a long RC of 12 words and a short VP of three words. This corresponds to a length ratio of 0.25, representing the lowest ratio in our experiment. For each length combination, the RC head noun was introduced either by the definite article *der* 'the' (n = 56) or by the cataphoric demonstrative *derjenige* 'the one' (n = 56). Our 112 experimental sentences were continuously distributed across the length ratio scale, and their total length varied between ten and 23 words.

All experimental items contained a sentence-initial adverb, and all RC head nouns were placed within the verbal bracket. All RC head nouns were singular and of masculine gender, so that all RCs were introduced by the case- and number-unambiguous, masculine singular relative pronoun *der* 'who'. Any other noun phrases intervening between the RC and its head noun were either feminine and/or plural, to ensure that the relative pronoun only had one possible antecedent in each sentence. A full list of our experimental sentences is available at the Open Science Framework (OSF) website (https://osf.io/xv97q/).

In addition to the critical stimulus sentences, 84 filler sentences were created, including 32 acceptable sentences and 52 ungrammatical or unacceptable sentences. Our filler items included sentences of varying length and syntactic complexity, including 21 grammatical or ungrammatical sentences that contained RCs. Thirty-seven of our filler sentences contained

[3] Due to a programming error, item 59 (DETERMINER = demonstrative; VP LENGTH = 5; RC LENGTH = 5) was not displayed to participants.

non-extraposed subclauses. We also added four attention-checking items, in which participants were asked to answer *yes/no* comprehension questions about the previous sentence by clicking on the corresponding button. Together with our unacceptable fillers, these allowed us to verify whether participants read the stimulus items attentively. All experimental and filler items were uniquely randomized for each participant at runtime.

Procedure

The experiment was designed as a web-based, speeded binary acceptability judgment task, using the experimental platform *PCIbex* (Zehr & Schwarz, 2018). Participants received a link to the experiment, where they were first asked to provide consent to participate in the study. They then answered some biographical questions before beginning the experiment. After reading the instructions, participants received three practice items to familiarize themselves with the experimental task. Participants were instructed to read each stimulus sentence carefully and then to decide, as fast as possible, whether or not the sentence presented to them was acceptable. Sentences were displayed word by word in the center of the screen in black 30-point *Lucida Grande* font against a white background, following the presentation of an initial fixation cross. Each word was shown for 300 ms and was then replaced by the next word, with the final word indicated by a full stop. Punctuation marks were displayed together with the preceding word; in our RCE items, the comma separating the main clause from the extraposed RC thus appeared attached to the participial verb at the end of the VP. After all words were presented, the following question appeared: *War der Satz akzeptabel?* 'Was the sentence acceptable?'. Participants were allowed a maximum of 2000 ms to provide their answer by pressing either the "F" key for 'yes' or the "J" key for 'no'. The experiment took approximately 90 minutes to complete, on average (ranging from 37 to 172 minutes).

Data cleaning and analysis

While we were primarily interested in participants' behavioral choices, we additionally analyzed their response times, as these can provide information about how our complexity measures affected the relative difficulty of reaching a decision (Donkin & Brown, 2018). Before running any statistical models, all timed-out responses were removed, as well as trials with a response time of 0 ms. This procedure eliminated 50 observations, incurring a data loss of 1.05%. The remaining 4722 observations were included in the statistical analysis. The statistics software *R* (Version 4.3.1) (R Core Team, 2023) was employed to analyze both the raw judgment data, using generalized linear mixed-effects models for binomial distributions, and the response times, using linear mixed-effects models. We employed the *glmer* and *lmer* functions of the *lme4* package (Bates et al., 2015), respectively. We analyzed models with the continuous predictors RC LENGTH and VP LENGTH and the categorical predictor DETERMINER as fixed effects, including potential interactions between these factors. All continuous predictors were scaled and mean-centered for standardized comparison. Additionally, a separate model incorporated the derived predictor LENGTH RATIO and examined its interaction with the factor DETERMINER. To identify the optimal

random slope structure, we followed Matuschek et al. (2017), including random slopes only if they improved model fit based on the Akaike Information Criterion (AIC), which balances model complexity with explanatory power to explain the variance of the data (Venables & Ripley, 2002). Lower AIC values indicated better model fit. We began with a maximal model containing random slopes for all factors and their interactions, then iteratively removed the random slopes by participant or item that contributed the least variance (Barr et al., 2013), until we identified the model with the lowest AIC possible.

3.1.2 Results

Judgment data

Our experimental sentences elicited a high proportion of positive judgments overall (79.21%, SD = 0.59). The generalized linear mixed-effects model, including the factors RC LENGTH, VP LENGTH, and DETERMINER and their corresponding interactions, revealed no significant effects for RC LENGTH or DETERMINER, but a significant main effect for VP LENGTH. The VP LENGTH effect reflects the fact that items with longer VPs yielded significantly lower proportions of *yes* responses. None of the interaction terms between the factors reached significance. The interaction model with the best-fitting random effect structure is given in **Table 1**.

We also ran an alternative model with the derived factor LENGTH RATIO and DETERMINER. The results of the best-fitting model for these factors showed a main effect of LENGTH RATIO, but no effect of DETERMINER (**Table 2**). Sentences with a higher VP-to-RC length ratio elicited significantly lower proportions of *yes* responses, irrespective of determiner type, as illustrated by **Figure 1**. No significant interaction was obtained between the two factors.

Table 1: Summary of statistical analysis of the judgment data, Experiment 1 (Model I).

	Estimate	**Std. Error**	***z*-value**	***p*-value**	
(Intercept)	1.697	0.151	11.246	<0.001	***
c.(RC LENGTH)	0.091	0.057	1.601	0.110	
c.(VP LENGTH)	–0.155	0.057	–2.715	0.007	**
DETERMINER	–0.047	0.096	–0.489	0.625	
c.(RC LENGTH):c.(VP LENGTH)	0.031	0.057	0.545	0.586	
c.(RC LENGTH): DETERMINER	0.002	0.057	0.041	0.968	
c.(VP LENGTH):DETERMINER	–0.086	0.057	–1.516	0.130	
c.(RC LENGTH):c.(VP LENGTH):DETERMINER	0.072	0.057	1.269	0.204	

Formula: glmer(Proportion of YES responses ~ c.(RC_Length) * c.(VP_Length) * Determiner + (1 + Determiner | Subject) + (1 | Item), family = “binomial”).
The significance levels reported here are the following: $p < 0.05$ = *; $p < 0.01$ = **; $p < 0.001$ = ***.

Table 2: Summary of statistical analysis of the judgment data, Experiment 1 (Model II).

	Estimate	Std. Error	*z*-value	*p*-value	
(Intercept)	1.697	0.151	11.249	<0.001	***
c.(LENGTH RATIO)	–0.187	0.056	–3.340	<0.001	***
DETERMINER	–0.046	0.096	–0.486	0.627	
c.(LENGTH RATIO):DETERMINER	–0.089	0.056	–1.586	0.113	

Formula: glmer(Proportion of YES responses ~ c.(Length_Ratio) * Determiner + (1 + Determiner | Subject) + (1 | Item), family = "binomial").
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

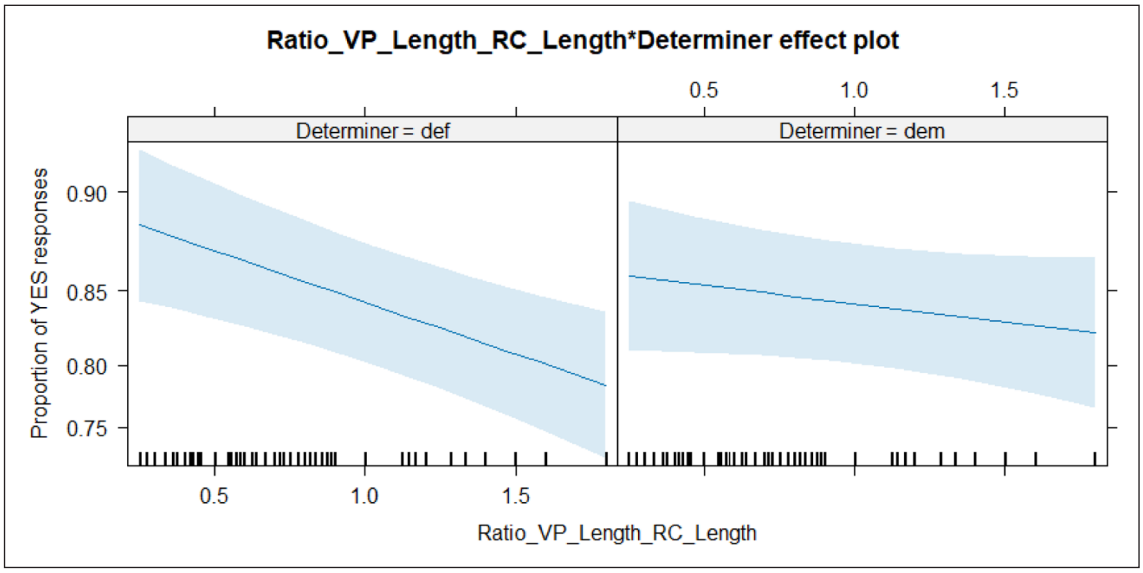

Figure 1: Effect plot for judgments per determiner type, Experiment 1 (Model II).

The model containing the factors LENGTH RATIO and DETERMINER (**Table 2**) provided a better fit for the data compared to the model with the factors RC LENGTH, VP LENGTH, and DETERMINER (**Table 1**), as indicated by the lower AIC value for Model II (AIC = 4215.8) compared to Model I (AIC = 4223.0). Lower AIC values suggest a more parsimonious model with better explanatory power relative to the data, favoring the inclusion of the LENGTH RATIO factor. The AIC difference ($\Delta_i = AIC_i - AIC_{min}$) between the two models of 7.2 lends little support for Model I (see Burnham & Anderson, 2002). The model probabilities calculated by Akaike Weights (w_i), show a probability of 2.66% that Model I is the best model, while Model II shows a probability of 97.34%.

Response times

Statistical models parallel to the ones reported above were run for the log-transformed response time results. A linear regression model (Model III) including the factors RC LENGTH, VP LENGTH, and DETERMINER yielded significant effects for RC LENGTH and DETERMINER (**Table 3**). Judgments of sentences containing longer RCs were provided significantly faster than judgments of sentences with shorter RCs, and sentences with a cataphoric demonstrative were responded to faster than those with a definite determiner. No significant effect was obtained for VP LENGTH, and neither the two-way nor the three-way interactions proved significant.

Table 3: Summary of statistical analysis of the response time data, Experiment 1 (Model III).

	Estimate	**Std. Error**	***df***	***t*-value**	***p*-value**	
(Intercept)	5.768	0.065	48.23	88.751	<0.001	***
c.(RC LENGTH)	–0.094	0.019	109.39	–4.859	<0.001	***
c.(VP LENGTH)	0.007	0.019	109.22	0.371	0.711	
DETERMINER	–0.080	0.028	108.68	–2.916	0.004	***
c.(RC LENGTH):c.(VP LENGTH)	–0.012	0.019	109.47	–0.629	0.531	
c.(RC LENGTH): DETERMINER	0.028	0.028	108.62	1.028	0.306	
c.(VP LENGTH):DETERMINER	–0.040	0.028	108.56	–1.414	0.160	
c.(RC LENGTH):c.(VP LENGTH): DETERMINER	0.029	0.027	108.63	1.056	0.293	

Formula: lmer(log_Response_Time ~ c.(RC_Length) * c.(VP_Length) * Determiner + (1 | Subject) + (1 | Item), REML = FALSE).
The significance levels reported here are the following: $p < 0.05$ = *; $p < 0.01$ = **; $p < 0.001$ = ***.

In addition to Model III, we also fitted an alternative model (Model IV) with the derived factor LENGTH RATIO and DETERMINER to explore a potential interaction motivated by previous work on expectancy effects in RCE (e.g., Levy et al., 2012). The results showed main effects for both factors, as well as an interaction between them (**Table 4**). Regarding the main effect of LENGTH RATIO, slower response times were obtained for items with higher VP-to-RC ratios. The effect of DETERMINER indicated overall faster response times for items where the head noun in the antecedent was introduced by a cataphoric demonstrative.

Table 4: Summary of statistical analysis of the response time data, Experiment 1 (Model IV).

	Estimate	**Std. Error**	***df***	***t*-value**	***p*-value**	
(Intercept)	5.769	0.066	49.96	87.943	<0.001	***
c.(LENGTH RATIO)	0.076	0.021	109.67	3.534	<0.001	***
DETERMINER	–0.082	0.030	109.14	–2.717	0.008	**
c.(LENGTH RATIO): DETERMINER	–0.061	0.030	109.01	–2.020	0.046	*

Formula: lmer(log_Response_Time ~ c.(Length Ratio) * Determiner + (1 | Subject) + (1 | Item), REML = FALSE).
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

However, Model III, with the factors RC LENGTH, VP LENGTH, and DETERMINER provided a better fit for the data (AIC = 11419.6) compared to Model IV, with the factors LENGTH RATIO and DETERMINER (AIC = 11432.3), yielding an Akaike difference of 12.7 points. Model III, with an Akaike weight of 0.9983, thus has a 99.83% probability of being the better model. In contrast, Model IV has an Akaike weight of only 0.0017, corresponding to a 0.17% probability. This large difference suggests that Model IV has virtually no support, making Model III the preferred one.

3.1.3 Summary and discussion

We found that the derived factor VP-to-RC length ratio proved to be the best predictor for RCE acceptability, such that acceptability increased with the extent to which RC length exceeded the length of the VP. Unlike Uszkoreit et al. (1998) and Konieczny (2000), whose offline judgment results showed effects of both extraposition distance (here, VP length) and RC length, we found that RC length on its own had no significant effect. The type of determiner – that is, whether an upcoming RC could be anticipated – also had no significant effect on participants' judgments. This finding is unexpected from the perspective of surprisal theory (Levy, 2008) and in view of the results from Strunk's (2014) offline judgment task, which showed that RC expectancy could effectively cancel out any negative effects of phrasal complexity on RCE in German. It is possible that the high occurrence rate of extraposed RCs in our materials reduced the degree of surprisal when an extraposed RC was encountered, to the extent that a cataphoric determiner did not measurably increase sentence acceptability.

Both RC length and the type of determiner significantly affected participants' response times, however. Response time in speeded choice tasks is commonly thought to reflect processing time, which primarily includes the time it takes for sufficient evidence to be accumulated before a decision can be made (Donkin & Brown, 2018). Our experimental sentences were responded to

more quickly when the extraposed RC was longer, and response times were faster for sentences containing cataphoric demonstrative determiners vs. definite determiners.

Sentences with longer extraposed RCs might have been responded to faster because they are more acceptable than sentences with shorter extraposed RCs, even though we did not find a significant main effect of RC length in our judgment data. Alternatively, the fact that sentences with longer RCs elicit faster response times than those with shorter RCs may be due to participants' making up their mind about a sentence's acceptability before it has been fully presented, probably as soon as the full VP has been presented. Therefore, longer RCs provide more time for coming to a decision than shorter ones do, so that the *yes* or *no* button can be pressed more quickly at the end of a trial if the RC is longer.

Faster responses to sentences containing a cataphoric determiner are likely to reflect a processing facilitation for extraposed RCs whose appearance was expected. As our stimulus presentation was externally paced, any such potential processing speed-up would only be detectable in participants' end-of-trial response times. We will return to the results of Experiment 1 in Section 4.

3.2 Experiment 2: Sentence production

Our second experiment examined how length-based complexity factors would affect RCE production. According to Hawkins' (2004) Domain Minimization principle, the likelihood of German speakers producing RCE should be influenced by VP-to-RC length ratio rather than by VP or RC length on their own. Conversely, the PEW (Wasow, 2002) predicts that RC length should be the primary factor triggering RCE, with long RCs preferentially being extraposed.

Our production task required participants to combine two simple sentences into a single complex one, which involved a silent sentence generation phase followed by the corresponding positioning of sentence fragments.[4] A timeout was added to reduce the possibility of participants re-evaluating different word order options or changing their minds about the positioning of RCs.

3.2.1 Method

Participants

We recruited 55 (19 female, 36 male; mean age: 32.85, range: 19–60), self-reported native German-speaking adults through Prolific (www.prolific.com). All participants indicated that German was their primary language and that they were born, raised, and currently residing in Germany. They all had normal or corrected-to-normal vision and reported no literacy difficulties.

[4] We used a written sentence anagram task rather than a spoken production task to minimize potential effects of sentence planning difficulty on structural choices and completion times, so that global effects of domain minimization would be easier to detect.

All participants gave their informed, voluntary, written consent before taking part in the study. They were naïve about the experiment's ultimate purpose and received monetary compensation of 5€ for their participation.

Materials

Our experimental items were adapted from the materials used in Experiment 1 and partly based on the materials used in Gosemann et al.'s (in press) untimed sentence anagram task. All experimental items comprised two parts. First, two full sentences were presented (henceforth referred to as *context sentences*). One always contained a singular, masculine indefinite subject NP (such as *ein Vater* 'a father' in **Figure 2**), while the other one provided additional information about that subject. Secondly, four sentence fragments were presented to be rearranged in the participants' preferred order. The materials set comprised a total of 56 items and 56 filler sentences. We manipulated two independent variables: (i) RC LENGTH (5–12 words), and (ii) VP LENGTH (3–9 words). As in Experiment 1, RC LENGTH and VP LENGTH represented continuous variables and were incorporated in a fully crossed design. Hence, the possible length combinations yielded 56 experimental sentences. Their LENGTH RATIOS again ranged from 0.25 to 1.8, and the RC extraposition distance was always VP LENGTH minus one.

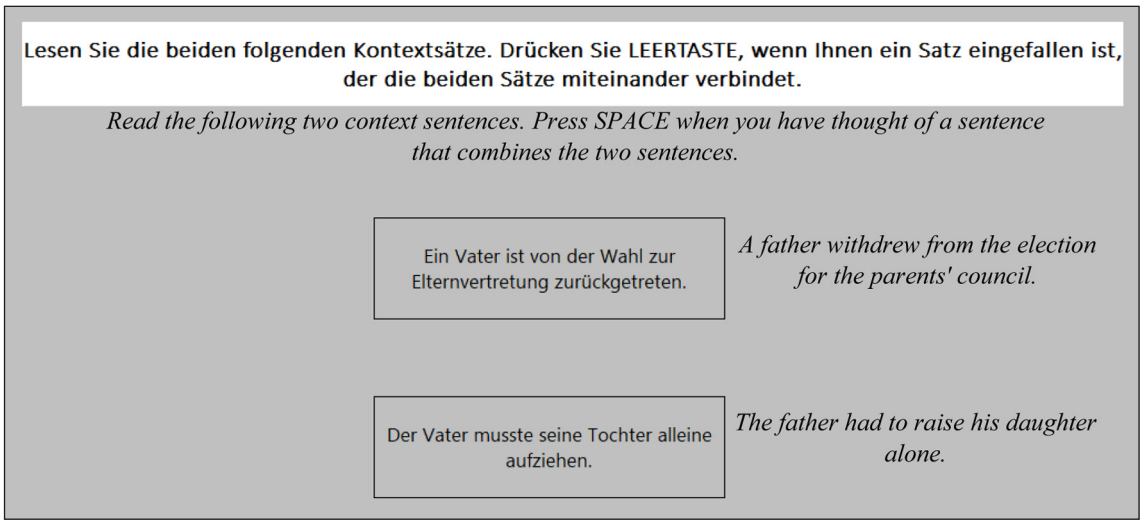

Figure 2: Context sentences, as seen by the participants during the experimental trials. Text in italics is added here to provide an English translation.

The example in **Figure 2** shows the two context sentences in the format that they were presented to the participants (English translations have been added here for the reader's convenience). The sentence presented on top was always intended as the target sentence's main clause, and the one below was intended to be turned into a subject-modifying RC during silent sentence generation.

Figure 3 illustrates the subsequently presented sentence fragments. The lexical items that appeared in these fragments were the same as those in the two context sentences, except for the addition of an adverb. This adverb was capitalized and appeared in a color-marked frame, which participants were instructed should always be used as the initial fragment. One of the frames always included a finite auxiliary and a definite singular subject noun phrase. Due to the V2 requirement, starting each target sentence with an adverbial helped ensure that the fragment containing the auxiliary and subject NP would always be positioned immediately after the adverb. The sentence fragments were segmented in such a way that the RC could either be positioned adjacent to the subject (6a) or after the lexical main verb, i.e., in extraposed position (6b). Only those two word orders were considered coherent.

(6) a. ADVERB > AUXILIARY & SUBJECT > RC > VERB PHRASE
b. ADVERB > AUXILIARY & SUBJECT > VERB PHRASE > RC

All RCs were introduced by the nominative singular, masculine relative pronoun *der* 'who' and were presented as a single fragment. A fourth fragment contained the remainder of the predicate phrase. Any other noun phrases intervening between the head noun and its corresponding RC were either feminine and/or plural, to ensure that the relative pronoun only had one possible antecedent in the sentence. A full list of our experimental sentences is available at the study's OSF website (https://osf.io/xv97q/).

Figure 3: Sentence fragments to be rearranged by dragging and dropping them in the white box below. The countdown timer was set to 30 seconds (here, showing 26 s). The English translations in italics have been added for the reader's benefit only.

In addition to the critical stimulus sentences, 56 filler sentences were included in the experimental setup. The set of filler items comprised biclausal sentences of varying length and type, including 17 finite complement clauses which also allowed for both integrated and extraposed word orders. The experimental and filler items were uniquely randomized for each participant at runtime.

Procedure

The experiment was programmed using the E-Prime 3.0 software and administered online via E-Prime Go (Psychology Software Tools, Pittsburgh, PA). Participants were informed about the minimum technical requirements their machines had to fulfil to be able to run the experiment. They were provided with a link to access and download the experiment. Upon accessing the experiment link, participants were first asked to provide informed consent for their participation, followed by some biographical questions.

After reading the instructions, participants received four practice items, to familiarize themselves with the experimental task. They were informed that the stimulus items comprised two parts. In the first part, two context sentences appeared centered on the screen against a light-grey background in a 13 pt *Leelawadee UI* black font (**Figure 2**). Participants were asked to think about how to combine the two context sentences into a single coherent sentence. Once they had formed a sentence in their mind, they could proceed by pressing the spacebar.

In the following screen, four sentence fragment frames appeared vertically aligned and centered against a light grey background (**Figure 3**). The order of these fragments was randomized for each trial. The frame containing the adverb was slightly smaller and had an olive-green background. The other three sentence fragment frames were of the same size and shown with a transparent background. The text in the four frames was presented in 11 pt *Leelawadee UI* black font.

Participants were instructed to arrange the sentence fragments in their preferred order by mouse-dragging and dropping each fragment into the horizontal white area on top of the countdown timer. Participants were instructed that the olive-green fragment containing the adverb should be positioned first. Participants were asked to rearrange the fragments as fast as possible and were shown a countdown set to 30 seconds. After this time, participants were timed out, and the next item appeared. The timeout was introduced to reduce the possibility of participants (re-)evaluating different potential word orders before positioning the fragments. Once they were done, participants had to click the green *FERTIG* ('done') button at the bottom of the screen to submit their answer and proceed to the next item.

Data cleaning and analysis

First, we checked for inattentive or uncooperative participants, by verifying whether the fragment containing the adverbial was placed in the first position, as instructed. All participants were accurate at least 98% of the time here. Then we checked whether the fragment containing the auxiliary and subject was placed in second position. Only one participant had an accuracy lower than 95% here (89.3%), and no participant fell below our exclusion threshold of 75%.

In a second step, we removed any incoherent orderings, including one observation where the adverbial was not positioned first and eight observations where the auxiliary and subject fragment was not used as the second fragment, amounting to a data loss of 0.29%. No datapoints had to be excluded due to time-outs. The remaining 3071 observations were included in the statistical analysis.

The statistical analysis was performed in R 4.3.1 (R Core Team, 2023), using a generalized linear mixed-effects model (GLMM) for binomial distributions, as logistic regression models are appropriate given the binomial nature of the dependent variable (Jaeger, 2008). We used the *glmer* function of the *lme4* package (Bates et al., 2015). The models analyzed the binary outcome produced as modulated by the different dependency length measurements. The two continuous predictors RC LENGTH and VP LENGTH were scaled and centered around their means before being included in the model.

To identify the best-fitting model, we initially constructed a model that incorporated all experimental factors and interactions as fixed effects, along with random intercepts and random slopes by item and participant. Following Matuschek et al. (2017), we determined the optimal random effect structure by including them only when they enhanced the model's goodness of fit, as measured by the AIC (Akaike, 1974). We began by comparing the maximal model with simpler, minimally reduced versions, choosing the one with the lower AIC, as this indicates a better fit. Starting with a maximal model that included random slopes for the relevant factors and their interactions, we refined it by iteratively adding or removing slopes, based on their effect on the AIC. If the model did not converge, we followed Barr et al. (2013) by systematically eliminating random slopes by participant or item that accounted for the least variance, until we achieved a stable, optimal model. The final best-fitting models, along with their formulas, are provided in 3.2.2.

3.2.2 Results

Production choices

Participants strongly favored placing the RC adjacent to its head noun, with 89.1% of the responses reflecting this integrated structure. Only 10.9% of the word orders produced featured an extraposed RC. The best-fitting GLMM model, including the factors RC LENGTH and VP

LENGTH, showed a significant effect for VP LENGTH but no main effect for RC LENGTH and no significant interaction (**Table 5**).

Table 5: Summary of the statistical analysis for production choices, Experiment 2 (Model I).

	Estimate	**Std. Error**	***z*-value**	***p*-value**	
(Intercept)	–3.234	0.274	–11.813	<0.001	***
c.(RC LENGTH)	0.180	0.107	1.681	0.093	
c.(VP LENGTH)	–0.453	0.159	–2.850	0.004	**
c.(RC LENGTH):c.(VP LENGTH)	0.193	0.106	1.828	0.068	

Formula: glmer(Word Order ~ c.(RC_Length) * c.(VP_Length) + (1 + c.(VP_Length) | Subject) + (1 | Item), family = "binomial").
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

The significant effect for VP LENGTH shows that a higher proportion of RCE was obtained for shorter VPs than for longer VPs. In other words, fewer instances of RCE were produced with increasing extraposition distance. Longer RCs also tended to facilitate the production of RCE, but neither the effect of RC LENGTH nor the interaction between the two factors reached significance.

The results of the alternative GLMM model for the derived factor LENGTH RATIO (see **Table 6**) showed a significant main effect, indicating that fewer instances of RCE were obtained for items with higher VP-to-RC ratios (**Figure 4**).

Table 6: Summary of the statistical analysis for production choices, Experiment 2 (Model II).

	Estimate	**Std. Error**	***z*-value**	***p*-value**	
(Intercept)	–3.274	0.290	–11.591	<0.001	***
c.(LENGTH RATIO)	–0.670	0.191	–3.495	<0.001	***

Formula: glmer(Word Order ~ c.(Length Ratio) + (1 + c.(Length_Ratio) | Subject) + (1 | Item), family = "binomial").
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

The model for LENGTH RATIO (Model II) has a slightly larger AIC (1618.5) than Model I, which includes the non-derived factors RC LENGTH and VP LENGTH (1618.0). The Akaike difference between the two models is 0.5 points, with Model I having an Akaike weight of 0.56 (probability 56.22%), and Model II an Akaike weight of 0.44 (probability 43.78%). This small difference between the models suggests that both models show equivalent support for the analyzed data.

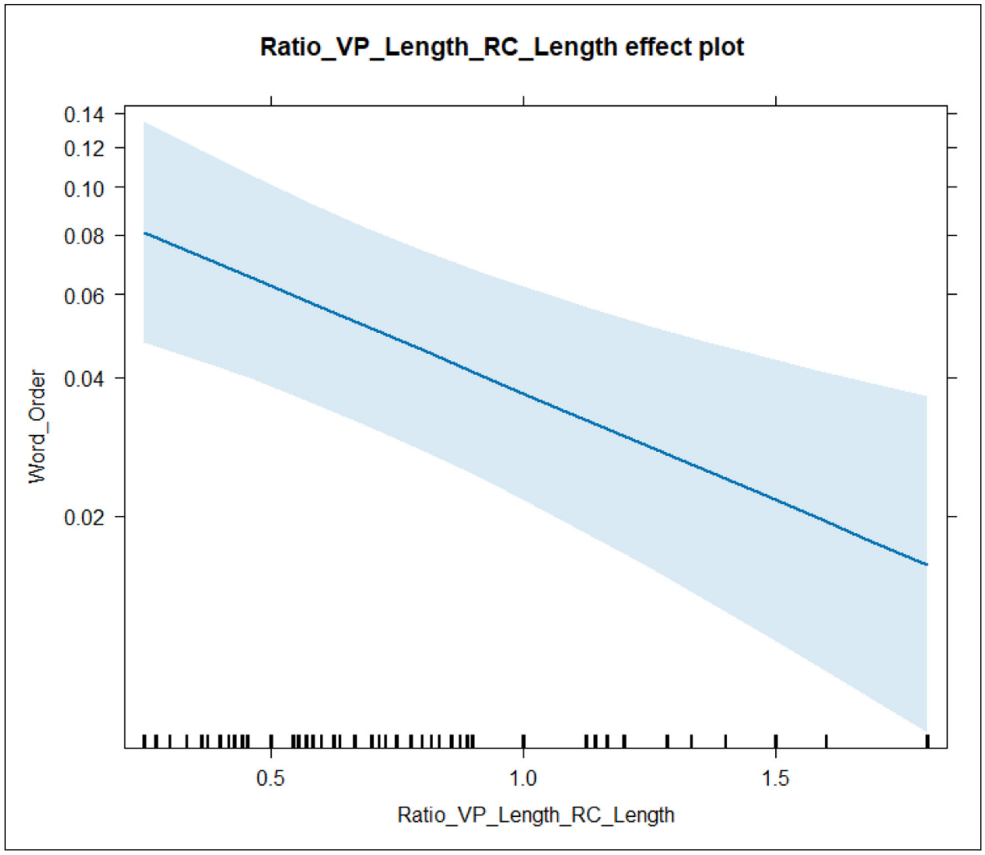

Figure 4: Effect plot for production choices, Experiment 2 (Model II).

Completion times

Statistical models parallel to the ones reported for production choices were conducted for completion times, but with the factor WORD ORDER (which was the dependent variable in the models above) now entered in the models as a fixed factor.

A linear mixed-effects model including the factors WORD ORDER, RC LENGTH, and VP LENGTH did not yield a significant effect for WORD ORDER, but we found significant main effects for VP LENGTH and RC LENGTH (see **Table 7**). Sentences with longer VPs required significantly longer completion times, compared to those with shorter VPs, and sentences with longer RCs required longer completion times than those with shorter RCs. A significant interaction was obtained between the factors WORD ORDER and RC LENGTH. To better interpret this interaction, we performed two independent analyses for the subsets of data for items following the WORD ORDER *integrated* and *extraposed*. In both cases, the effect of RC LENGTH was significant and in the same direction: *integrated* ($t = 2.456$; $p < 0.019$) and *extraposed* ($t = 2.781$; $p < 0.007$). The latter, however, shows a greater effect size. None of the remaining two-way or three-way interactions were significant.

Table 7: Summary of the statistical analysis for completion times, Experiment 2 (Model III).

	Estimate	**Std. Error**	***df***	***z*-value**	***p*-value**	
(Intercept)	9.167	0.033	59.286	276.919	<0.001	***
WORD ORDER*extraposed*	0.007	0.016	3030.301	0.454	0.650	
c.(RC LENGTH)	0.019	0.008	58.937	2.447	0.017	*
c.(VP LENGTH)	0.040	0.008	60.120	5.238	<0.001	***
WORD ORDER:c.(RC LENGTH)	0.043	0.013	2999.865	3.172	0.002	**
WORD ORDER:c.(VP LENGTH)	0.020	0.013	2999.767	1.616	0.106	
c.(RC LENGTH):c.(VP LENGTH)	0.010	0.008	59.678	1.252	0.215	
WORD ORDER:c.(RC LENGTH): c.(VP LENGTH)	0.018	0.013	3001.409	1.422	0.155	

Formula: lmer(log_Completion_Time ~ Word Order * c.(RC_Length) * c.(VP_Length) + (1 | Subject) + (1 | Item), REML = FALSE).
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

The results of the alternative LMEM model for the derived factor LENGTH RATIO in combination with factor WORD ORDER (see **Table 8**) showed no significant main effects and no interaction between the factors. Items with higher VP-to-RC ratios do not seem to affect sentence completion time, regardless of whether the RC was positioned integrated or extraposed.

Table 8: Summary of the statistical analysis for completion times, Experiment 2 (Model IV).

	Estimate	**Std. Error**	***df***	***z*-value**	***p*-value**	
(Intercept)	9.168	0.034	63.849	271.154	<0.001	***
WORD ORDER*extraposed*	0.005	0.016	3016.322	0.340	0.734	
c.(LENGTH RATIO)	0.015	0.010	57.289	1.476	0.146	
WORD ORDER: c.(LENGTH RATIO)	–0.010	0.014	2981.435	–0.684	0.494	

Formula: lmer(log_Completion_Time ~ Word Order * c.(Length_Ratio) + (1 | Subject) + (1 | Item), REML = FALSE).
The significance levels reported here are the following: p < 0.05 = *; p < 0.01 = **; p < 0.001 = ***.

In order to determine which of the two models presents a better fit for capturing the completion time data, the Akaike difference between the two models was calculated. This difference between the models amounts to 38.31 points. Having a ΔAIC greater than 10 (see Burnham & Anderson, 2002) lends no support for Model IV. With an Akaike weight of 0.99, Model III (AIC = –220.87) proves superior to Model IV (AIC = –182.56), which has an Akaike weight smaller than 0.001.

3.2.3 Summary and discussion

Similar to what we observed for acceptability judgments in Experiment 1, VP length and VP-to-RC length ratio – but not RC length on its own – significantly influenced the likelihood of participants choosing to produce an extraposed RC. Decreasing VP-to-RC length ratios increased the likelihood of RCE production, in line with Hawkins' (2004) Domain Minimization principle. The results from Experiment 2 are also broadly consistent with findings from German corpus studies that examined effects of RC length and extraposition distance, albeit not their length ratios (Gosemann et al., in press; Shannon, 1992; Strunk, 2014; Uszkoreit et al., 1998).

The absence of a main effect of RC length on word-order choices is unexpected from the perspective of Wasow's (2002) PEW, however. The lack of an RC length effect could potentially be due to our experimental task, which differs from spoken production tasks in which speakers must retrieve and plan the entire utterance before articulation. By contrast, our paradigm externalized the lexical content and limited possible linearization options, thereby reducing demands on memory retrieval and incremental message planning. As such, Experiment 2 primarily taps structural preferences under relatively low planning load, rather than the full range of processes involved in spontaneous spoken production.

Ordering choices were timed so as to reduce the likelihood of participants' (re-)evaluating different options or changing their mind, but in the absence of a clear hypothesis as to what completion times reflect, our completion time results must be interpreted with caution. Our finding that sentences with long VPs or long RCs took more time to complete is most likely due to sentences with lengthier constituents tending to take longer to be completed. Although our complexity measures had similar effects on both acceptability judgments and production choices, the fact that in Experiment 2, integrated RCs were overwhelmingly preferred over extraposed ones does not seem to match the high acceptability rate of RCE that we observed in Experiment 1. We will return to the apparent discrepancy between RCE perception and production in our general discussion.

4. Discussion

Starting from the assumption that minimizing or optimizing dependency lengths in a sentence increases processing efficiency, we investigated the influence of VP length, RC length, and

VP-to-RC length ratio on the acceptability and production of relative clause extraposition in German. Experiment 1 used speeded acceptability judgments and additionally manipulated RC expectancy, and Experiment 2 examined speakers' production choices with added time pressure. We asked whether the above complexity measures would affect the acceptability and production likelihood of RCE in comparable ways, and how RC expectancy modulates the effects of our length-based complexity measures on speakers' online judgments.

The current study extends previous research on RCE by systematically examining wider than usual ranges of VP and RC lengths, by examining both judgment and production using parallel materials, and by using speeded judgment and timed production tasks to minimize participants' reliance on stylistic or audience design considerations. The derived factor VP-to-RC length ratio was examined for German for the first time.

While our results primarily showed similarities between our production and comprehension-based tasks, they revealed some differences between participants' behavioral responses and the time measurements we took. **Table 9** provides an overview of the statistical effects we found in each experiment in the best-fitting models.

Table 9: Summary of significant effects and interactions per experiment and dependent variable for models with the lowest AIC value.

	Responses	**Response times**
Experiment 1	LENGTH RATIO	RC LENGTH, DETERMINER
Experiment 2	VP LENGTH/LENGTH RATIO	RC LENGTH, VP LENGTH WORD ORDER * RC LENGTH

4.1 Judgments vs. production choices

In Experiment 1, all experimental sentences contained RCE, and participants provided a high proportion (79.21%) of positive judgments, indicating that RCE is generally perceived as acceptable. Our primary finding was that the VP-to-RC length ratio was the strongest predictor of RCE acceptability. The results from Experiment 2 identified VP length and VP-to-RC length ratio as equally good predictors for participants' production choices. That is, sentences with smaller VP-to-RC length ratios (i.e., shorter VPs relative to RC length) were judged as more acceptable, and were also more likely to elicit extraposition, compared to sentences with higher length ratios. RC length on its own did not prove a significant factor in either experiment. Taken together, the response data from our two experiments indicate that RCE acceptability and production likelihood are similarly affected by dependency length optimization, with RCE being facilitated across shorter extraposition distances and by the degree to which RC length exceeds VP length. These findings are consistent with previous corpus findings from English (Francis, 2010; Francis

& Michaelis, 2014) and lend support to Hawkins' (2004) Domain Minimization principle. To our knowledge, this is the first time that the factor VP-to-RC length ratio has been systematically examined for RCE in German.

Interestingly, RC expectancy (manipulated through the use of cataphoric demonstratives) did not significantly affect the proportion of positive acceptability judgments in Experiment 1, contrary to the prediction made by surprisal theory (Levy, 2008) and Levy et al.'s (2012) processing results for English. This finding is also at odds with Strunk's (2014) offline acceptability judgment results, which showed that cataphoric determiners can neutralize the negative effects of longer extraposition distances. A possible explanation for this discrepancy could lie in differences in the way the stimulus sentences were presented. In Levy et al.'s (2012) reading-time study, stimulus presentation was participant-paced, and in Strunk's (2014) untimed judgment task, the experimental sentences were presented in their entirety. By contrast, our stimuli were presented word-by-word at a predetermined rate, and participants' responses were timed. The nature of our task, thus, did not allow for relative processing difficulty to be measured during the presentation of the RC (as in Levy et al., 2012), and being able to anticipate an upcoming RC did not allow for faster reading. At the same time, using word-by-word presentation – rather than whole-sentence presentation, as in Strunk (2014) – reduced the likelihood of participants evaluating the global pragmatic or stylistic quality of our stimuli before coming to an acceptability decision. In addition, the high proportion of extraposed RCs in Experiment 1 likely reduced the surprisal associated with encountering RCE and may have limited the scope for additional acceptability gains from cataphoric determiners.

We also found differences between RCE perception and production. Despite the high acceptability of our RCE stimuli in Experiment 1, the results of Experiment 2 revealed a strong preference for integrated (89.1%) over extraposed RCs (10.9%). The overwhelming preference for integrated RCs resembles Gosemann et al.'s (in press) written production results of 5% RCE choices, but only partially aligns with previous oral production results (e.g., Bader, 2014, 2024). Examining RCE across a minimal distance of one word, Bader (2024) found that approximately 50% of RCs were extraposed. Bader (2014), however, found that when RCE would have to cross a new discourse referent intervening between the RC and its host NP, participants would choose to extrapose an RC only 11% (Exp. 1) or 31% of the time (Exp. 2). These figures are closer to the low extraposition rate we observed in our Experiment 2. Given that all RCs in our experiments were subject-modifying, with the subject always directly following the left verbal bracket, one or more additional discourse referents intervened between the RC and its host in most cases in which the extraposition distance was greater than one word.

As noted by a reviewer, one factor that might have contributed to the low rate of RCE in Experiment 2 is the fact that the antecedent NP was always definite. Strunk (2014) found a lower

extraposition rate for RCs with definite antecedents than for RCs with indefinite antecedents in his corpus study, and Francis and Michaelis (2017) observed a similar definite/indefinite asymmetry for elicited spoken production in English. These findings differ from the results we obtained in a pilot offline judgment task, however, in which we observed no advantage for indefinite antecedents over definite antecedents of extraposed RCs. As antecedent definiteness was not manipulated here, its possible role in guiding participants' production choices remains unclear.

Our finding that even long RCs were preferentially placed adjacent to their host NP fails to lend support to the PEW (Wasow, 2002), according to which long or complex phrases should be positioned late in a sentence. The low extraposition rates observed in elicited written production tasks contrast with the relatively high frequency of RCE in written corpus data, such as in Gosemann et al.'s (in press) newspaper corpus. This discrepancy might at least partly reflect stylistic or audience design considerations that influence the production of written texts, potentially leading to inflated extraposition rates. Stylistic guidelines for German-speaking newspaper journalists do, in fact, explicitly recommend that subordinate clauses be extraposed (Schneider & Raue, 2012).

Together with the absence of a main effect of RC length, our finding that integrated RCs were strongly preferred in a production task that did not encourage the top-down evaluation of different word-order options suggests that minimizing the dependency length between the RC and its head noun is more relevant than minimizing the size of the verbal bracket. That said, note that our written sentence anagram task involved reduced sentence planning difficulty, relative to spoken production (see 3.2). Potential effects of end-weight that arise from real-time planning pressures may, thus, be attenuated in our paradigm.

While the low extraposition rate observed in Experiment 2 may seem unexpected, given that RCE was deemed acceptable most of the time in Experiment 1, it should be noted that Experiment 1 did not compare integrated vs. extraposed word orders. That is, the fact that RCE elicited a high proportion of positive judgments does not entail that integrated RCs would have scored any worse. Using an untimed scalar judgment task, Gosemann et al. (in press) found that integrated RCs were, in fact, rated better than extraposed ones, overall. RCE was also clearly deemed acceptable, however, eliciting a mean rating of 4.91 on a 7-point scale. Likewise, the fact that our participants preferentially produced integrated RCs does not entail that they considered the extraposed variants inappropriate. In sentence production, one word-order variant merely wins out over all other potentially suitable variants in a given situation. Recall that our principal research question concerned the way in which different complexity measures would affect RCE acceptability and production likelihood, and our results revealed largely parallel complexity effects for both.

4.2 Response times

Although, in both experiments, we were primarily interested in participants' responses, the time measures we obtained can serve as potential indicators of relative processing difficulty. Interestingly, the relationship between VP length and RC length, which predicted participants' judgments and production choices, did not influence our response time measures.

In Experiment 1, response times were faster for sentences with longer RCs and those containing cataphoric demonstratives. This suggests that, while these factors did not influence acceptability judgments, they did facilitate processing. Longer RCs may have given participants more time to decide on acceptability before the sentence ended. Furthermore, extraposed RC that were expected were likely to be easier to integrate with their host NP during processing or during end-of-sentence "wrap-up," which may have shortened response times without necessarily affecting or altering participants' binary acceptability decisions. The observed processing facilitation linked to the use of cataphoric determiners aligns with previous research showing that predictability facilitates RC integration (Levy et al., 2012).

The analysis of sentence completion times in Experiment 2 showed that longer RCs and VPs increased the time needed to arrange the sentence fragments provided in stage two of our task. Since longer VPs and RCs tended to increase overall sentence length, finding completion times to be affected by our length measures is hardly surprising. That said, the observed interaction between word order and RC length showed that the effect of RC length was more pronounced for extraposed RC than for integrated RCs. While sentences containing longer RCs generally took longer to be completed than those with shorter RCs, the slow-down triggered by longer RCs was greater if the RC was extraposed. Although our time measures must be interpreted with caution, this finding might indicate that extraposing long RCs is more difficult than placing long RCs adjacent to their host, which, again, is unexpected from the perspective of the PEW.

Our findings that length-based complexity measures and RC expectancy selectively affected either participants' responses or their response times indicate that the factors modulating processing difficulty do not necessarily impact sentence acceptability or production choices in the same way or to a corresponding extent.

5. Concluding remarks

Our findings highlight the interplay between syntactic complexity and processing efficiency in shaping language use. Using parallel materials in a speeded judgment task and a timed production task, we found that VP-to-RC length ratio emerged as a critical factor across both experiments for the response variables investigated. Our findings confirm the predictions made by Hawkins' (2004) Domain Minimization principle and support its hypothesized status as a processing-based constraint on word-order variation. In Experiment 1, higher VP-to-RC length ratios reduced the

proportion of positive acceptability judgments of RCE, while in Experiment 2, they decreased the likelihood of participants producing RCE. This suggests that both comprehension and production are sensitive to the balance between VP and RC lengths, with speakers preferring structures that minimize or optimize dependency distances. The results from our sentence generation experiment do not support the PEW (Wasow, 2002). The strong preference for integrated RCs that we observed replicates the findings from Gosemann et al.'s (in press) untimed production experiment, suggesting that the relatively high occurrence rates of RCE in written corpora may be inflated, due to audience design considerations or the application of stylistic norms.

Contrary to our expectation, increasing RC expectancy through the use of a cataphoric determiner did not increase RCE acceptability in Experiment 1. RC expectancy did influence the time it took participants to provide their judgments, however. This finding is consistent with the prediction made by surprisal theory (Levy, 2008), and suggests that sentences containing expected RCs are easier to process. Our finding that participants' responses and response times were differentially modulated by our complexity measures and by RC expectancy calls for further research on how processing difficulty influences the perception and production of different word-order variants, however. Theoretical hypotheses that propose a tight link between processing efficiency and word-order preferences (e.g., Hawkins, 2004, 2014) are largely based on the analysis of written corpus and untimed judgment data, neither of which provides any direct measure of processing ease or efficiency. Future studies on word-order variation might want to make more use of online methods to further explore how processing difficulty and ordering preferences are related.

Data accessibility statement

The data and analysis scripts, as well as our experimental items, can be accessed at the following OSF repository: https://osf.io/xv97q/. The analysis scripts also document exploratory models in which Trial Order was included as a predictor for Experiment 1._

Ethics and consent

This study was conducted in accordance with the Declaration of Helsinki and approved by the University of Potsdam Ethics Committee (reference number 37/2011).

Acknowledgments

We thank Sina Bosch for contributing to the experiment design and materials creation, and Marina Fomenko for her support in setting up a pilot study and recruiting participants.

Funding information

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287.

Competing interests

The authors have no competing interests to declare.

Author contributions

Both authors were involved in all aspects of the study, including its conceptualization, data collection and analysis, and the writing of the manuscript. Both authors approved the final version of the manuscript.

ORCiD IDs

Álvaro Cortés Rodríguez: https://orcid.org/0000-0001-9727-5315
Claudia Felser: https://orcid.org/0000-0001-9424-0950

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(1), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Arnold, J. E., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Languages, 76*(1), 28–55. https://doi.org/10.1353/lan.2000.0045

Bader, M. (2014). Defining distance in language production: Extraposition of relative clauses in German. *Cognitive Processing, 15*(1), 81–84. https://doi.org/10.21500/20112084.807

Bader, M. (2024). Relative clause extraposition and information structure. In A. Himmelreich, D. Hole, & J. Mursell (Eds.), *To the left, to the right, and much in between: A Festschrift for Katharina Hartmann* (pp. 205–216). Goethe University Frankfurt. https://doi.org/10.17605/OSF.IO/3FX4M

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blackwell, A., Bates, E., & Fisher, D. (1996). The time course of grammaticality judgment. *Language and Cognitive Processes, 11*(4), 337–406. https://doi.org/10.1080/016909696387150

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Donkin, C., & Brown, S. D. (2018). Response times and decision-making. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience. Volume 5* (pp. 349–377). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119170174.epcn509

Francis, E. J. (2010). Grammatical weight and relative clause extraposition in English. *Cognitive Linguistics, 21*(1), 35–74. https://doi.org/10.1515/cogl.2010.002

Francis, E. J., & Michaelis, L. A. (2014). Why move? How weight and discourse factors combine to predict relative clause extraposition in English. In B. MacWhinney, A. Malchukov, & E. Moravcsik (Eds.), *Competing motivations in grammar and usage* (pp. 70–87). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198709848.003.0005

Francis, E. J., & Michaelis, L. A. (2017). When relative clause extraposition is the right choice, it's easier. *Language and Cognition, 9*(2), 332–370. https://doi.org/10.1017/langcog.2016.21

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language, 96*(2), 371–412. https://doi.org/10.1353/lan.2020.0024

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *PNAS, 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science, 34*, 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition, 37*(2), 269–297. https://doi.org/10.1017/S0272263114000850

Gosemann, L., Cortés Rodríguez, Á., & Felser, C. (in press). Dependency length minimization effects on clausal extraposition are modulated by clause type and data type. *Linguistische Berichte*.

Hawkins, J. A. (2004). *Efficiency and complexity in grammar*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199664993.001.0001

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research, 29*(6), 627–645. https://doi.org/10.1023/A:1026528912821

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition, 122*(1), 12–36. https://doi.org/10.1016/j.cognition.2011.07.012

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*, 226. https://doi.org/10.3389/fpsyg.2013.00226

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Psychology Software Tools, Inc. (2016). *E-Prime 3.0*. https://support.pstnet.com/

Psychology Software Tools, Inc. (2020). *E-Prime Go*. https://support.pstnet.com/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Schneider, W., & Raue, P. J. (2012). *Das neue Handbuch des Journalismus und des Online-Journalismus*. Rowohlt Verlag GmbH.

Shannon, T. F. (1992). Toward an adequate characterization of relative clause extraposition in modern German. In I. Rauch, G. F. Carr, & R. L. Kyes (Eds.), *On Germanic linguistics: Issues and methods* (pp. 253–281). De Gruyter. https://doi.org/10.1515/9783110856446.253

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language, 39*, 392–417. https://doi.org/10.1006/jmla.1998.2586

Strunk, J. (2014). A statistical model of competing motivations affecting relative clause extraposition in German. In B. MacWhinney, A. Malchukov, & E. Moravcsik (Eds.), *Competing motivations in grammar and usage* (pp. 88–106). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198709848.003.0006

Temperley, D., & Gildea, D. (2018). Minimizing syntactic dependency lengths: Cognitive/typological universal? *Annual Review of Linguistics, 4*, 67–80. https://doi.org/10.1146/annurev-linguistics-011817-045617

Uszkoreit, H., Brants, T., Duchier, D., Krenn, B., Konieczny, L., Oepen, S., & Skut, W. (1998). Studien zur performanzorientierten Linguistik: Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft, 7*(3), 129–133. https://doi.org/10.1007/s001970050065

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S: Statistics and computing*. Springer. https://doi.org/10.1007/978-0-387-21706-2

Wasow, T. (1997). End-weight from the speaker's perspective. *Journal of Psycholinguistic Research, 26*, 347–361. https://doi.org/10.1023/A:1025080709112

Wasow, T. (2002). *Postverbal behavior*. CSLI Publications.

Yadav, H., Mittal, S., & Husain, S. (2022). A reappraisal of dependency length minimization as a linguistic universal. *Open Mind, 6*, 147–168. https://doi.org/10.1162/opmi_a_00060

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832