GLOSSA psycholinguistics

Yadav, H., Frank, S. L., Futrell, R., & Husain, S. (2025). Revisiting processing complexity of nested and cross-serial dependencies. *Glossa Psycholinguistics*, 4(1): 8, pp. 1–28. DOI: https://doi.org/10.5070/G601120916

eScholarship University of California

Revisiting processing complexity of nested and cross-serial dependencies

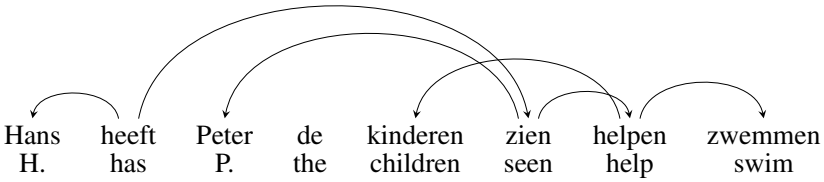
Himanshu Yadav, Indian Institute of Technology Kanpur, Kanpur, India, himanshu@iitk.ac.in
Stefan L. Frank, Radboud University, Nijmegen, the Netherlands, stefan.frank@ru.nl
Richard Futrell, University of California Irvine, Irvine (CA), USA, rfutrell@uci.edu
Samar Husain, Indian Institute of Technology Delhi, Delhi, India, samar@hss.iitd.ac.in

In two web-based experiments, we compare comprehension difficulty between Dutch and German sentences with clusters of two or three verbs. In Dutch, such sentences involve crossing dependencies, whereas these dependencies are nested in German. Replicating the seminal finding of Bach et al. (1986), we find that the crossing (Dutch) structure is easier to comprehend than the nested (German) structure, although we find a different pattern of results in terms of where this difficulty manifests. The results are in line with predictions from dependency locality theory.

1. Introduction

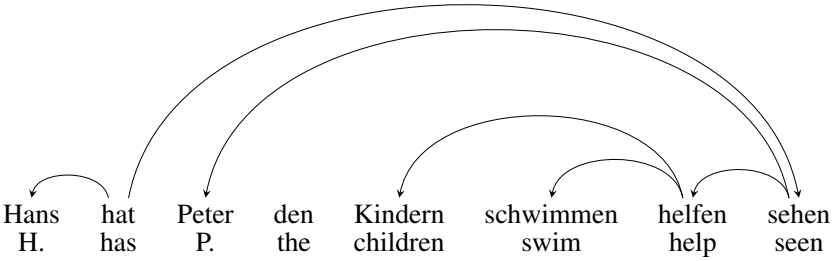
Languages differ in the kind of formal structures they use to encode grammatical relationships: for example, to encode embedded subject–verb relations, German uses nested structures while equivalent sentences in Dutch can use crossing structures, which correspond to discontinuous phrase structures: see (1a). Crossing structures such as these are at the heart of debates about the formal characterization of natural language grammar and parsing (Bresnan et al., 1982; Joshi et al., 1991; Kuhlmann, 2013; Shieber, 1985), both in terms of computational systems and human language processing (Levy et al., 2012).

(1) a. *Dutch*



'Hans has seen Peter help the children swim.'

b. *German*



'Hans has seen Peter help the children swim.'

Early research involving these structures established that natural language grammars are not context-free, both in terms of their weak generative capacity as well as their strong generative capacity (Bresnan et al., 1982; Shieber, 1985); also see Kobele (2006). Crossing dependency structures are precisely where we can identify deviations from context-free grammar. Any formal grammar that encompasses these crossing dependencies must be more complex than context-free: this additional complexity often takes the form of special mechanisms in the grammar which specifically handle discontinuous constituents (for example, the Move / Internal Merge operation in Minimalist grammars, or adjunction in Tree-Adjoining Grammars (TAG), or the multiple components of Multiple Context-Free Grammars; Seki et al., 1991). Grammars that minimally capture natural-language-like crossing dependencies are called **mildly context-sensitive** (Joshi, 1985; Joshi et al., 1991; Weir, 1988).

How do these findings from the formal language theory literature relate to processing of such structures? In a seminal study, Bach et al. (1986) investigated the processing difficulty

associated with crossing dependencies by comparing the comprehension difficulty of sentences containing crossing structures in Dutch with sentences containing nested structures in German. If there is inherent processing difficulty for crossing dependencies, Dutch sentences containing crossings should be more difficult than their German non-crossing counterparts. Surprisingly, the authors found that German sentences are more difficult to comprehend compared to Dutch sentences. This study inspired work in formal syntax, parsing, and psycholinguistics, because it suggested that crossing dependencies may not be a major determinant of processing difficulty (Graf et al. (2017); Joshi (1990); Rambow and Joshi (1994); Rambow and Satta (1994); among others).

However, this influential result has been found using only one methodology and set of materials, and the data is not available for more in-depth analysis using modern techniques and theories. Here we report two web-based replications and extensions of Bach et al. (1986), investigating if comprehension difficulty of German nested structures and equivalent Dutch crossing structures differs, and whether this depends on the depth of embedding. Consistent with the original results, we find that processing is subjectively more difficult with multiple embeddings in the German word order than the Dutch word order. We interpret these results in terms of dependency locality: when compared to the Dutch sentences, the equivalent German sentences have longer dependencies, which have independently been found to be associated with processing difficulty due to working memory constraints regardless of crossings (Bartek et al., 2011; Fedorenko et al., 2013; Gibson, 1998, 2000; Grodner and Gibson, 2005).

2. Crossing dependencies in psycholinguistics

2.1 Why crossing dependencies might be difficult

Since crossing dependencies can only be captured through more complex grammars, it would be reasonable to hypothesize that these dependencies also come with additional processing cost. The move from context-free to mildly-context-sensitive grammars comes with a definite cost in terms of the worst-case time complexity of exact parsing, from cubic time $O(n^3)$ for context-free grammars to $O(n^5)$, $O(n^7)$, etc. for various mildly context-sensitive grammars (Seki et al., 1991), up to $O(n^{28})$ for a wide-coverage Minimalist grammar (Torr et al., 2019). Furthermore, under an assumption of strong competence (Bresnan, 1982; Chomsky, 1965), the added complexity in formal grammars such as TAG, MG, etc., which accounts for non-context-freeness in natural language, could reflect special operations for crossing dependencies during human language processing (Levy et al., 2012).

In theories where the production system faithfully encodes syntactic dependencies in an utterance (Bock et al., 2002), it has been assumed that generating a context-free structure should be less costly than generating a non-context-free structure (involving a crossing dependency). In particular, theories of sentence production assume some degree of incrementality (Levelt, 1989),

but in some views, non-context-free dependencies such as filler–gap dependencies require advance planning (Momma, 2021). There is also evidence that sentences that are difficult to produce tend to also be difficult to comprehend (MacDonald, 2013; Scontras et al., 2015). On such accounts, the comprehension system should find the non-context-free structures involving crossing dependencies difficult to parse because it is rarely exposed to such configurations.

One implication is that the processing system would rather form a simpler context-free structure than a complex non-context-free structure. Evidence for this comes from comprehension of structures involving filler–gap dependencies in English where it has been found that native speakers tend to avoid forming filler–gap dependencies if possible (Staub et al., 2018), and that they tend to resolve the gap as early as possible (Clifton and Frazier, 1989; Frazier, 1985), limiting the range of the non-context-free dependency. Relatedly, recent results have shown that it is difficult to prime structures involving crossing dependencies (Husain and Yadav, 2020). Again, it is assumed that comprehension of such structures involves a (parsing) process that establishes all the required dependencies faithfully. Either way, it is quite reasonable to assume that the comprehension system should find parsing non-context-free structures to be difficult (cf. Frazier, 1987).

2.2 Why crossing dependencies might not be difficult

Despite the above, the most common theoretical and empirical stance in psycholinguistics has been that crossing dependencies do not pose special processing challenges, and the Bach et al. (1986) result that we revisit here was a key piece of evidence in establishing this idea.

While mildly context-sensitive grammars have worse time complexity than context-free grammars, the time complexity analysis is based on the worst case. For individual sentences, the time taken to parse may be higher or lower under context-free or mildly context-sensitive grammars. Indeed, although Torr et al. (2019) calculate a worst-case time complexity of $O(n^{28})$ for their parser, they find that its *average* parsing time seems to reflect cubic time complexity, similarly to context-free grammars. Most relevantly, inspired by the Bach et al. (1986) results, Joshi (1990) presents a parsing model for TAG based on an extended pushdown automaton in which the German-style nested dependencies require items to be stored on a (structured) stack for a longer time than the Dutch-style cross-serial dependencies; further automaton models along these lines include Rambow and Joshi (1994) and Kobele et al. (2013). If the number of items stored on the stack is a predictor of human processing difficulty (a common assumption: see for example Abney and Johnson, 1991; De Santo, 2020; Graf et al., 2017; Resnik, 1992; Stabler, 1994; Yngve, 1960), then the particular sentences with crossing dependencies would be easier than those with deep nested dependencies, even though the need to accommodate such dependencies results in worse worst-case behavior.

The preponderance of empirical psycholinguistic evidence also seems to suggest that it is long dependencies, not crossing dependencies, that cause processing difficulty under appropriate circumstances, because long dependencies tax working memory resources (Bartek et al., 2011; Ferrer-i-Cancho, 2006; Gibson, 1998, 2000; Grodner and Gibson, 2005; Levy, 2013).[1] For example, right-extraposition (which leads to a crossing dependency) is preferred over its embedded counterpart (which leads to a non-crossing dependency) when the length of the right-extraposed relative clause is longer (Francis, 2010; Hawkins, 1994). In other words, if the total dependency distance in the sentence with an extraposed relative clause is less than the sentence with an embedded relative clause, then the former configuration is preferred. Crossing constructions can also be preferred due to information structure (Huck and Na, 1990; Rochemont and Culicover, 1990). Second, there is evidence that parsing crossing dependencies is not difficult as long as the dependency is highly expected (e.g., Levy et al., 2012). Finally, although crossing dependencies are relatively rare in naturalistic corpora (Ferrer-i-Cancho et al., 2018; Havelka, 2007; Kuhlmann, 2013), this could be genre dependent: the dependency corpus data showing rarity of crossing dependencies mainly comes from news corpora in various languages. It could very well be the case that crossing dependencies are much more common in naturalistic dialogue settings where factors such as information structure and accessibility more strongly dictate sentence formulation rather than the presumed complexity of crossing dependencies.

The first and key empirical result suggesting that crossing dependencies are not especially difficult was Bach et al. (1986), which showed that crossing dependencies in Dutch are easier to comprehend than meaning-equivalent multiply nested structures in German. Bach et al. (1986) conducted a rating study using recorded spoken items similar to those in (1a) and (1b). The levels of nesting and crossing in German and Dutch respectively ranged from 1 to 4 (where Level 1 corresponded to no nesting/crossing). Each of these nested/crossing items had a corresponding paraphrase in order to control for semantic/propositional complexity. Unlike the nested/crossing items, the paraphrases used right-branching structures that were identical between the two languages. Participants were asked to rate the items on a 9-point scale from ‘easy’ to ‘difficult’. In addition, participants were asked comprehension questions that targeted specific NPs in the sentence (only at Levels 2 and 3, and corresponding Paraphrase items). The results showed that, relative to the paraphrases, crossing structures in Dutch were subjectively easier to understand and achieved higher comprehension accuracy than their German nested counterparts. Moreover, this difference between languages was larger at deeper levels of embedding.

[1] We use dependency locality theory (Gibson, 2000) to analyze the current results, but this theory is not in conflict with the perspective based on how long items must be stored on the stack of a parser: dependency length corresponds exactly to the minimal time that an item must be stored on the stack of a left-corner parser.

The dependency locality perspective predicts that crossing dependencies in Dutch should be easier than meaning-equivalent nested dependencies in German, because the multiple nesting of subject-verb dependencies in German creates a long dependency between the auxiliary and the verb (see **Figure 1**). The locality theory assumes that increased distance between a head and its dependent causes difficulty in maintaining the co-dependent in memory, predicting a higher cost when the long dependency must be integrated. For example, the AUX → V3 dependency in German has six intervening words (see **Figure 1**). In contrast to German, Dutch has relatively shorter dependencies (distance $\leq$ 4 for all dependencies) due to crossing structure. Dependency locality theory thus naturally predicts lower comprehension difficulty for Dutch crossing structures than German nested structures.

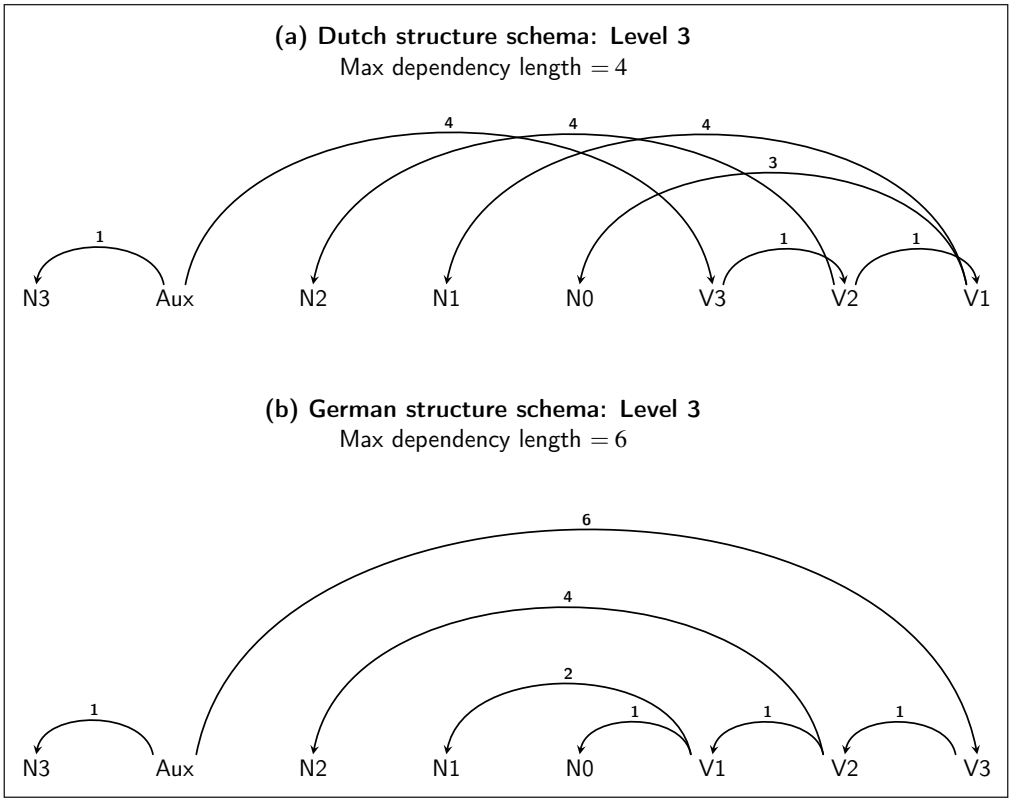


Figure 1: The schematic representation of dependency structure in German vs. Dutch sentences. The dependencies going from verbs (V1, V2, V3) to nouns (N1, N2, N3) are stacked over each other in German, but crossing each other in Dutch. The numbered labels on each dependency arc represent the length of the dependency.

Given the status of Bach et al. (1986) as a starting point for the formal and psycholinguistic work described above, we propose to replicate it. While the original study found robust results, they were obtained in only one modality (audio presentation) and the by-trial and by-participant

data are no longer available for analysis. In this article, we present two replications of Bach et al. (1986) using new materials in a written format and analyze the results from the perspective of dependency locality.

3. Experiment 1

3.1 Methods

We replicate Bach et al. (1986) in an online experiment in the written modality, presenting Dutch and German native speakers with sentences like those in (1a) and (1b), and collecting difficulty ratings and answers to comprehension questions.

3.1.1 Materials

3.1.1.1 Experimental stimuli

Exactly six verbs[2] can be used in these constructions in both German and Dutch: *sehen/zien* 'to see', *helfen/helpen* 'to help', *hören/horen* 'to hear', *lehren/leren* 'to teach', *fühlen/voelen* 'to feel', and *lassen/laten* 'to let'. For each of these verbs, we constructed four German and four Dutch translation-equivalent items. Each item comes in a two- and a three-levels-of-embedding version, that is, with two or three consecutive verbs. Furthermore, following Bach et al. (1986), each embedded item was paired with a non-embedded paraphrase to control for any non-syntactic differences between languages or between embedding levels. Only the German versions of these paraphrases contain commas, which are obligatory before complement clauses in German but not commonly included in Dutch.

Table 1 presents one item in all 2 (Language) $\times$ 2 (Level) $\times$ 2 (Paraphrase) $=$ 8 conditions. The 96 experimental sentences were divided over four lists such that, in each list, each condition occurred four times for each of the six verbs, and each item occurred in only one of the conditions. The item order was randomized per list.

3.1.1.2 Verb forms

In the no-paraphrase condition, all Dutch verbs (except for the auxiliary *heeft* 'has') were in the infinitive form. German speakers, however, disagree on whether the final verb of the cluster needs to be an infinitive or past participle. Bach et al. (1986) solved this by testing two groups of German participants, one on infinitives and one on participles. A potential issue with that approach is that it may reduce average comprehensibility ratings. Assuming that encountering the dispreferred form reduces comprehensibility, variance in preference between participants (or between verbs, for that matter) will lead to lower average ratings in German than in Dutch.

[2] Dutch has a seventh verb that may be used in this way, *doen* 'to make someone do'.

Table 1: Example item in all eight experimental conditions. English translations of the sentences are 'Timo saw the athlete run the marathon' (Level 2) and 'The binoculars helped Timo see the athlete run the marathon.' (Level 3)

Language	Level	Par.	Example
German	2	no	Timo hat den Athleten den Marathon laufen sehen.
		yes	Timo hat gesehen, dass der Athlet den Marathon lief.
	3	no	Das Fernglas hat Timo den Athleten den Marathon laufen sehen geholfen.
		yes	Das Fernglas hat Timo geholfen, um zu sehen, dass der Athlet den Marathon lief.
Dutch	2	no	Timo heeft de atleet de marathon zien lopen.
		yes	Timo heeft gezien dat de atleet de marathon liep.
	3	no	De verrekijker heeft Timo de atleet de marathon helpen zien lopen.
		yes	De verrekijker heeft Timo geholpen om te zien dat de atleet de marathon liep.

To tackle this issue, we ran a German verb-form preference pretest[3] in which 45 native German speakers indicated their preference for one of two sentences that differed only in the final verb form, for example:

(2) a. Timo hat den Athleten den Marathon laufen sehen. (infinitive)
b. Timo hat den Athleten den Marathon laufen gesehen. (participle)

Participants could also select an 'equally good'-option. One sentence pair was presented for each of the six verbs. We then selected for our experimental stimuli the most often preferred form of each verb. These were the infinitives *sehen* and *lassen*, and the past participles *geholfen* 'helped', *gehört* 'heard', *gelehrt* 'learned', and *gefühlt* 'felt'.

3.1.1.3 Fillers

Each of the four experimental lists included the same 52 filler sentences. Sixteen of the fillers, taken from Frank et al. (2016), had doubly nested center-embedded relative clauses, making them very difficult to understand. The other 36 fillers varied in the number of main verbs (one, two, or three; twelve sentences each) but did not have a purposefully difficult structure. German and Dutch fillers were translation-equivalents.

[3] Details of the pretest can be found in the supplementary materials.

3.1.1.4 Comprehension tests

In order to measure how well the experimental and filler sentences were understood, each sentence was paired with four statements, only one of which corresponded to the content of the sentence. The nouns and verbs of the three distractor statements also occurred in the corresponding sentence. For example, for the Level-2 sentences of **Table 1** these statements (translated to English) were:

(3) a. The athlete ran the marathon. (true)
b. The athlete saw the marathon. (false)
c. The athlete saw Timo. (false)
d. Timo ran the marathon. (false)

The four statements were identical for the paraphrase and non-paraphrase conditions, and translation-equivalent between German and Dutch. The statements were visible simultaneously, in random order and without the stimulus sentence present.

All experimental stimuli with their list assignment, filler sentences, and comprehension test items are available as supplementary materials.

3.1.2 Participants and procedure

Using the Prolific platform, we recruited 40 adult native German speakers living in Germany, and 40 adult native Dutch speakers living in the Netherlands. Participants of the verb form pretest were excluded from Experiment 1. The experiment was fully web-based and in the participant's native language; we use English translations in the description below.

After giving informed consent, participants read the instruction to rate sentences on how easy or difficult they are to understand using a slider response, and then choose one of four statements that is correct given the previously seen sentence. The experiment began with a single practice item, where feedback was provided on the comprehension test, followed by a reminder of the instructions. No feedback was provided on the other comprehension tests.

Each sentence was presented with the horizontal slider below. The slider was labeled (from left to right) 'very easy', 'easy', 'hard', and 'very hard'. The slider's midpoint (which was also its initial position) was indicated by a dot. Participants could move the slider using the mouse. Their difficulty ratings were then recorded on a continuous scale from 0 (very easy) to 100 (very hard). After at least clicking on the slider, participants could click the 'continue' button to replace the sentence with the comprehension question that asked which of the four statements shown on the screen is correct given the content of the sentence. They would then click what they considered to be the correct statement, and the next sentence item appeared directly.

After the last comprehension test, participants in the German condition received instructions about the following acceptability test in which they would select which of two sentences sounds

more natural, or click a button marked ‘Don’t know / equally good’. They were then shown pairs of new Level-2 No-paraphrase items (one pair for each of the six verbs) where one sentence used the infinitive verb form and the other the past participle. There was also one control trial with one clearly ungrammatical option.

Finally, both versions of the experiment asked for demographic information: Age, region in which the participant grew up (free text), region of current residence (free text), and highest level of education (multiple choice). German participants then indicated whether they were fluent in Dutch, while Dutch participants were asked about fluency in German.

Median completion time was 35 minutes for Dutch and 36 minutes for German. Participants were paid US$ 7.

3.1.3 Differences in methodology from Bach et al. (1986)

A careful reader will have noted that while we attempt a replication of Bach et al. (1986), there are some differences in our experimental setup compared to the original study; we list the key differences below.

i. The items used in the current study are not identical to the ones used in the original study. Also, we didn’t include Level 1 and Level 4 items. The original items are unavailable.

ii. The German items in the current study were selected through a norming study. Through this norming, it was ensured that the final verb cluster had the most preferred form (either infinitive or past participle). The original Bach et al. (1986) treated the verb form as a between-subjects factor, thereby testing two groups, one for infinitives and another for past participles. As stated above, a potential issue with that approach is that it may reduce average comprehensibility ratings if people’s actual preferences are graded and variable across verbs, presenting a potential confound for the results of Bach et al. (1986).

iii. The non-availability of the original items also meant that the filler items in the current study were different from those in Bach et al. (1986). However, similar to the original study, some filler items matched the critical items in terms of the number of nouns and verb. Unlike the original study, the number of filler items in the current study was higher (52 vs. 36). The current study also had some difficult fillers to obscure the difficulty associated exclusively with the critical items.

iv. The comprehension questions in the original study were open-ended: “Was tat NP?”/“Wat deed NP?”. Unlike those, the comprehension test in the current study comprised of a forced-choice task where participants had to choose a correct statement out of four options. These options targeted specific dependencies between nouns and verbs.

v. The items in the original study were presented auditorily. The current replication used visual (written) presentation.

vi. The rating in the original was done on a discrete scale of 1–9 (1 was labeled as 'easy', 9 was labeled as 'difficult'). In the current study, ratings were given using a slider that was labeled as 'very easy', 'easy', 'hard', 'very hard'; the ratings were recorded on a continuous scale of 0 (very easy) to 100 (very hard).

vii. Apart from the rating and the comprehension tasks (as in Bach et al., 1986), the current study had an additional task that was presented at the end of the German experiment. In this task, the participants chose their preference of the infinitive vs. past participle ending for the 6 critical verbs used in the study.

viii. Finally the original Bach et al. (1986) study was conducted in a lab setting. The data for the current study was collected online.

3.1.4 Data analysis

Participants who reported fluency in the other language or who had less than 40% accuracy (chance accuracy is 25%) on comprehension questions were rejected from the analysis. For German, individual trials were rejected for participants who indicated a preference for the non-presented verb form, so that only items with the subject's preferred verb form were included in the analysis. The final data used for analysis consisted of 5,616 observations (36 participants) from Dutch and 6,314 observations (40 participants) from German.

We want to infer whether the difference in comprehension difficulty between embedded and paraphrase sentences is significantly higher for German sentences compared to Dutch, and whether this difference in relative difficulty in German vs. Dutch is higher in double-embedded (Level 3) sentences compared to single-embedded (Level 2) sentences.

To draw these inferences from the data, we fit a linear mixed-effects model (Baayen et al., 2008; Gelman and Hill, 2007) with varying intercepts and slope adjustments for subjects and items using the lme4 package (Bates et al., 2014) in R. We tested the effect of language and embedding level on the mean difficulty ratings using the following formula:

$$\text{difficulty} \sim \text{language} * \text{level} * \text{paraphrase} + (\text{level} + \text{paraphrase} \mid \text{subject}) + (1 \mid \text{item}),$$

which is the maximal converging model structure (Barr et al., 2013; Bates et al., 2015).

To test the effect of language and embedding levels on the accuracy of comprehension question responses, we used a logistic regression model with the same fixed- and random-effects structure as the previous model, predicting whether a participant's response is correct (1) or not (0).

However, the mixed-effects linear regression for the difficulty rating data assumes that the residuals of the mixed-effects model are normally distributed. We need to verify whether this

assumption holds. **Figure 2** shows the distribution of residuals obtained from the mixed effect model fitted to the ratings data. We observe that the residuals are approximately normally distributed. However, the values of the rating scores are slightly inflated at the boundaries. An alternative way of analyzing these data is to use a Beta regression which would allow us to model the truncated data with inflation at the boundaries. The details and the results of the Beta regression are shown in supplementary materials Section 3. We find that our main conclusion does not change with the use of a Beta regression. We stick to linear regression analysis in our main text.

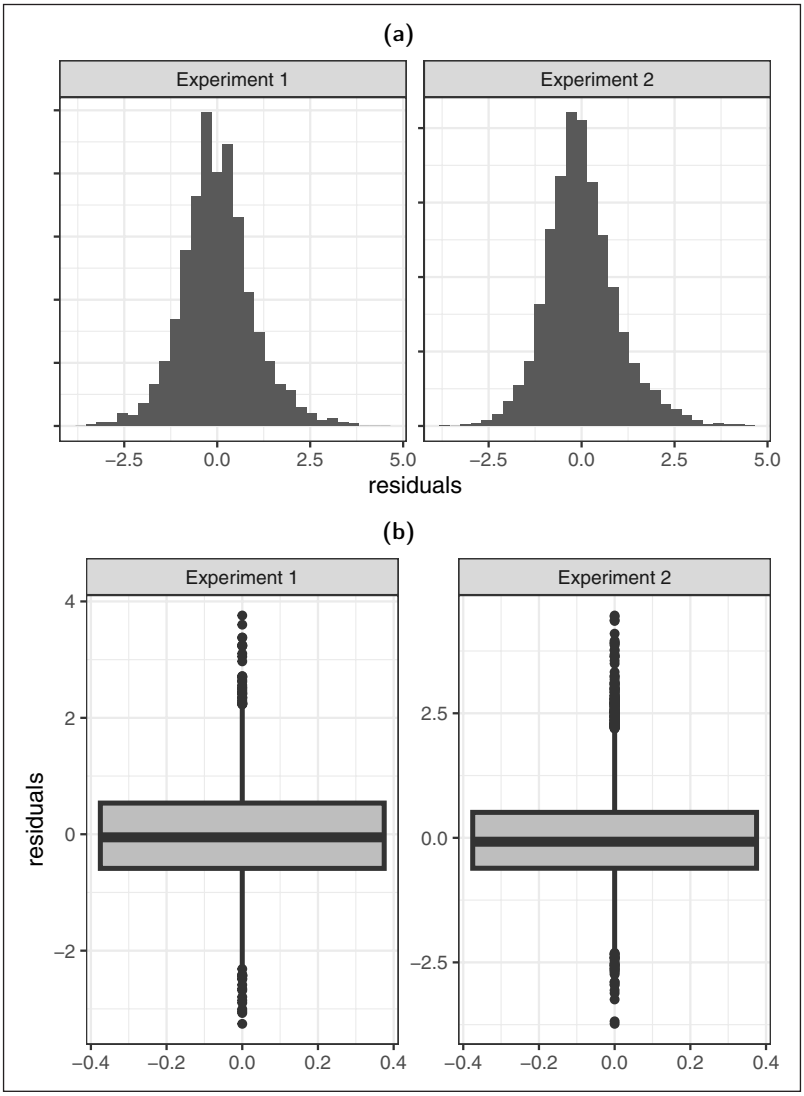


Figure 2: The distributions of residuals obtained from the linear mixed models fitted to difficulty ratings data from Experiment 1 (left panel) and Experiments 2a–b (right panel). Plot (a) shows the histograms of the residuals and Plot (b) shows the boxplots.

3.1.5 Predictions

The effects of interest for our analysis are the interaction effect of sentence type (paraphrase or embedded) and language, and the three-way interaction of language, sentence type, and embedding level. If these interaction effects have (significantly) positive estimates, Bach et al. (1986)'s results are successfully replicated.

The locality theory predicts that German embedded sentences should be more difficult to process than Dutch ones, i.e., the interaction of language and sentence type should be positive. Moreover, the increase in difficulty with respect to embedding level should be larger in German, that is, the three-way interaction of language, level, and paraphrase should be positive. Thus, both Bach et al. (1986)'s claim—that crossing dependencies are easier to comprehend than meaning-equivalent multiple-nested structures—and the locality account predict a positive estimate for the effect of Paraphrase = No × Language = German and for the effect of Level 3 × Paraphrase = No × Language = German.

3.2 Results and discussion

3.2.1 Comprehension difficulty ratings

Figure 3 shows the average difficulty rating in each of the eight conditions. Unsurprisingly, Level-3 sentences are rated as more difficult to understand than Level-2 sentences, and paraphrases are rated as less difficult than non-paraphrases. Critically, the difference in difficulty between

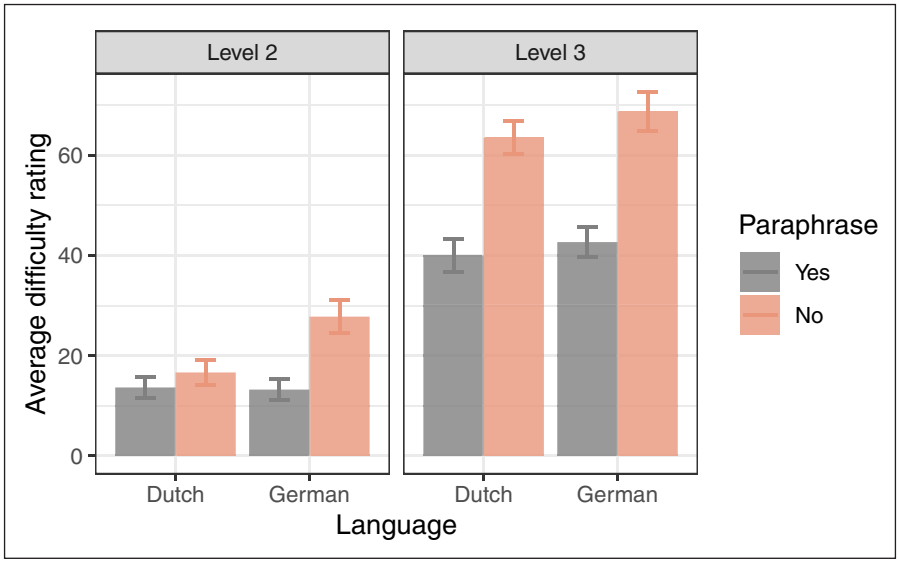


Figure 3: Average comprehension difficulty in each condition of Experiment 1. The difference between the orange (No-paraphrase) and gray (Paraphrase) bars indicates the comprehension difficulty caused by the use of an embedded structure, which is the effect of interest. The error bar shows the standard error of the mean difficulty rating.

paraphrase and non-paraphrase sentences appears to be larger in German than in Dutch, which is confirmed by a significant positive interaction of Language and Paraphrase in the regression analysis results presented in **Table 2**.

Table 2: The estimated effects of embedding level, language, and paraphrase condition on comprehension difficulty ratings in Experiment 1. The estimates were obtained from a linear mixed-effects regression model fitted to the difficulty rating data. The two theoretically important interaction effects are highlighted with a gray shade; significant effects are marked with an asterisk *.

Effect	Estimate	Std.Err.	*t*-value	
Intercept	13.70	2.56	5.36	*
Level 3	26.47	2.19	12.10	*
Paraphrase = No	2.82	1.94	1.45	
Language = German	–0.43	3.10	–0.14	
Level 3 × Paraphrase = No	20.44	2.39	8.55	*
Level 3 × Language = German	2.87	3.02	0.95	
Paraphrase = No × Language = German	12.45	2.74	4.55	*
Level 3 × Paraphrase = No × Language = German	–8.87	3.46	–2.56	*

This result implies that nested structures in German are more difficult than crossing structures in Dutch. The result supports the locality theory and confirms a successful replication of Bach et al.,'s (1986) general finding.

However, unexpectedly, the three-way interaction of Language, Level, and Paraphrase is negative. This negative interaction means that the relative difficulty of German decreases, rather than increases, with the extra level of embedding. This negative interaction is unexpected both from the perspective of the results of Bach et al. (1986) and from a locality theory of processing difficulty. The negative interaction is unlikely to be a ceiling effect—notice that the maximal difficulty rating is 100, but the average difficulty rating for level-3 non-paraphrase sentences in German is around 70. A possible explanation for this effect is that while the processing difficulty in German is higher than Dutch for level-2 sentences, the subjective difficulty saturates at level-3 embedding across both languages.

3.2.2 Comprehension question accuracy

Comprehension accuracy results are shown in **Figure 4**. The logistic regression in **Table 3** shows no significant effects of interest on question response accuracy. There is a significant general effect whereby level-3 stimuli, paraphrase or not, have lower comprehension accuracy.

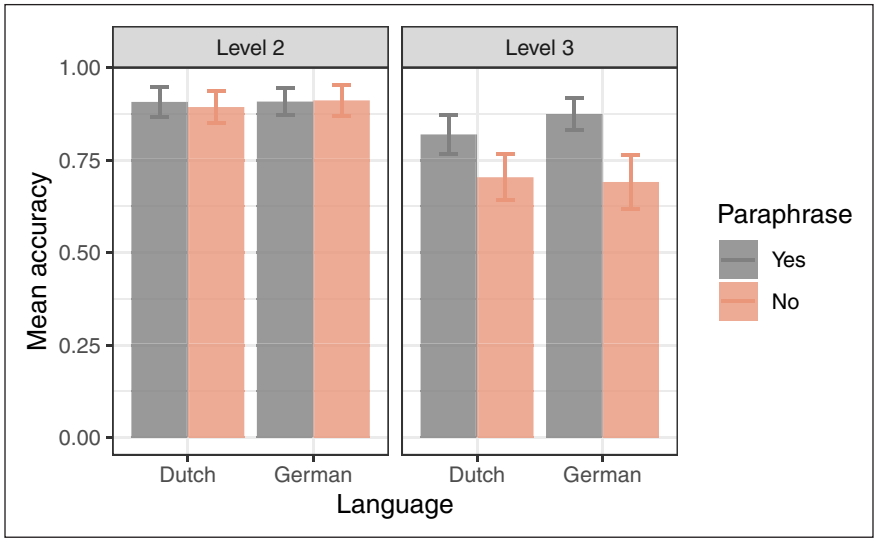


Figure 4: Average question response accuracy in each condition of Experiment 1. The difference between the orange (No-paraphrase) and gray (Paraphrase) bars indicates the decrease in question response accuracy caused by the use of an embedded structure, which is the effect of interest. The error bar shows the standard error of the mean response accuracy.

Table 3: The estimated effects of embedding level, language, and paraphrase condition on question response accuracy in Experiment 1. The estimates were obtained from a mixed-effects logistic regression fitted to the question response data. The two theoretically important interaction effects are highlighted with a gray shade.

Effect	Estimate	Std.Err.	*z*-value	*p*-value
Intercept	2.49	0.31	8.09	<0.01
Level 3	–1.43	0.28	–5.10	<0.01
Paraphrase = No	0.20	0.33	0.59	0.55
Language = German	0.26	0.37	0.71	0.48
Level 3 × Paraphrase = No	0.54	0.41	1.32	0.19
Level 3 × Language = German	–0.35	0.43	–0.81	0.42
Paraphrase = No × Language = German	–0.30	0.48	–0.64	0.53
Level 3 × Paraphrase = No × Language = German	0.91	0.61	1.50	0.13

4. Experiments 2a and 2b

Although the results from Experiment 1 are partially consistent with those from Bach et al. (1986), there are three major concerns. First, the study might be underpowered. Second, because participants can take as long as they wanted reading each sentence (unlike in the auditory

presentation from Bach et al.,'s original study) higher difficulty in one language could have been compensated by spending more time on the sentence—this may explain the lack of a crosslinguistic difference in comprehension accuracy, as German speakers may be taking more time to achieve the same level of accuracy as Dutch speakers. Third, the very high comprehension accuracy rates in the level 2 condition can be indicative of a ceiling effect.

All three concerns are dealt with in Experiment 2, where we control reading time by using a Rapid Serial Visual Presentation (RSVP) format to present the stimuli. Experiment 2a was an exploratory study, the results of which were used in a power analysis to determine the number of participants for a preregistered confirmatory study (Experiment 2b).[4]

4.1 Methods

4.1.1 Materials

Stimuli and comprehension tests were identical to those of Experiment 1 with three minor changes. First, a small number of the verb forms in the German no-paraphrase items of Experiment 1 turned out not to match the outcome of the verb-form preference pretest, which was corrected. Second, semantically incongruous answer options in the comprehension test were replaced by semantically meaningful alternatives in order to reduce the probability of guessing correctly. Third, minor changes were applied to the fillers to increase parallelism between languages, for example by using German-Dutch cognate words whenever possible.

All experimental stimuli with their list assignment, filler sentences, and comprehension test items are available as supplementary materials.

4.1.2 Participants

All participants were adult native German speakers living in Germany and adult native Dutch speakers living in the Netherlands, recruited on the Prolific platform. Participants of Experiment 1 and the verb form pre-test were excluded from taking part in Experiments 2a and 2b.

Forty-four Dutch participants and forty-one German participants were recruited for Experiment 2a. The power analysis based on effect estimates from Experiment 2a revealed that 96 participants per language would provide 80% power after taking into account likely data loss (see Supplementary Materials Section 3 for the power analysis details). After testing these additional participants, twelve additional Dutch speakers and one additional German speaker were recruited because a higher-than-expected number of Dutch speakers indicated fluency in German. Participants from Experiment 2a were excluded from taking part in Experiment 2b. Participants were paid £7.20.

[4] The preregistration can be found at https://aspredicted.org/blind.php?x=3XC_VMS.

4.1.3 Procedure

The procedure was identical to that of Experiment 1 with the exception of the sentence presentation method. Each sentence was preceded by a centrally presented fixation cross above a button labeled 'Start'. Upon clicking the button, the fixation cross was replaced by the sentence's first word which was then automatically replaced by each following word until completion of the sentence. To ensure that the number of presented tokens per sentence was identical between languages, it was occasionally necessary to display two words at a time for one of the languages, as indicated in the supplementary materials.

Following the EEG study by Frank et al. (2015), tokens were visible for a length-dependent duration of $190 + 20\max\{n_{\text{German}}, n_{\text{Dutch}}\}$ ms, where n is the number of characters in the token, including any punctuation or space. This was followed by a 390 ms interval before the next token appeared. Taking the maximum length of the German and Dutch token ensures that the total time for each sentence is identical between languages.

After the offset of the sentence-final word, the comprehensibility rating slider would appear, followed by the comprehension test, as in Experiment 1. The German rating experiment was followed by the verb-form preference test. The same demographic information was collected as in Experiment 1, except that the German study also asked about fluency in Swiss German and the Dutch study also asked about fluency in Frisian. This is because Swiss German, like Dutch, has crossing dependencies, and Frisian, like standard German, has nested dependencies in these verb clusters.

Median completion times was 34 and 31 minutes for German and Dutch, respectively, in Experiment 2a; and 36 and 35 minutes in Experiment 2b.

4.1.4 Data analysis and predictions

Participants who reported fluency in the other language, in Swiss German, or in Frisian were rejected from the analysis. For the German experiment, individual trials were rejected for participants who indicated a preference for the non-presented verb form so that only items with the preferred verb form were included in the analysis. We analyze the data from Experiments 2a and 2b together. The data consisted of 10,184 observations (134 participants) from Dutch and 9,572 observations (132 participants) from German.

We want to infer whether the comprehension difficulty and accuracy are significantly higher for German sentences compared to Dutch and whether this difference in difficulty in German vs. Dutch is higher in double-embedded (level-3) sentences compared to single-embedded (level-2) sentences. The regression models and the predictions were the same as described in Sections 3.1.4 and 3.1.5, respectively. Bach and colleagues' claim and the locality hypothesis would predict a (significantly) positive estimate for the interaction effect of language (= German) and

sentence type (= no-paraphrase) and also a positive three-way interaction of language, sentence type, and level of embedding. The positive estimate for these two interaction effects would imply a successful replication of Bach et al.'s main findings.

4.2 Results and Discussion

4.2.1 Comprehension difficulty ratings

Figure 5 shows the average difficulty rating in each of the eight conditions. We find that German sentences have higher comprehension difficulty compared to Dutch sentences, as confirmed by the significant interaction of language and paraphrase in **Table 4**. We also find a significant three-way interaction of language, paraphrase, and level, this time positive, implying that the difference in comprehension difficulty of German vs Dutch is higher for level-3 sentences compared to level-2 sentences.[5] The results collectively indicate that German sentences (containing nested structures) are more difficult to comprehend than Dutch sentences (with crossed structures), and this comprehension difficulty in nested structures gets significantly increased in double-embedding compared to single-embedding structures.

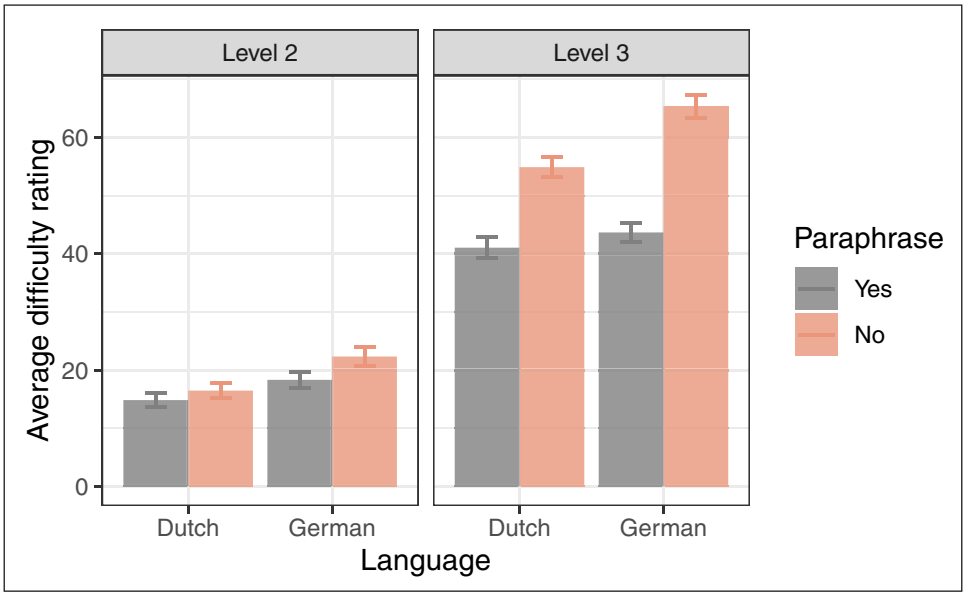


Figure 5: Average comprehension difficulty in each condition of Experiment 2. The difference between the orange (No-paraphrase) and gray (Paraphrase) bars indicates the comprehension difficulty caused by the use of an embedded structure, which is the effect of interest. The error bar shows the standard error of the mean difficulty rating.

[5] The RSVP presentation seems to eliminate the unexpected negative three-way interaction found in Experiment 1, although it was not intended to do so.

Table 4: The estimated effects of embedding level, language, and paraphrase condition on comprehension difficulty ratings in Experiment 2. The estimates were obtained from a linear mixed-effects regression model fitted to the difficulty rating data. The two theoretically important interaction effects are highlighted with a gray shade; significant effects are marked with an asterisk *.

Effect	Estimate	Std.Err.	*t*-value	
Intercept	15.22	1.66	9.15	*
Level 3	25.82	1.24	20.90	*
Paraphrase = No	1.44	1.05	1.36	
Language = German	2.71	1.70	1.59	
Level 3 × Paraphrase = No	12.12	1.31	9.24	*
Level 3 × Language = German	0.59	1.76	0.34	
Paraphrase = No × Language = German	3.11	1.53	2.03	*
Level 3 × Paraphrase = No × Language = German	4.86	1.97	2.47	*

In alternative Beta regressions shown in Supplementary Materials Section 3, we find a significant negative language by paraphrase interaction, but the three-way interaction is not supported. We note that there may be a power issue here; our power analysis was designed around detecting the two-way interaction of language and paraphrase.

4.2.2 Comprehension question accuracy

Despite the RSVP method, we do not find significant crosslinguistic differences in comprehension accuracy (**Figure 6**), although accuracy is now overall lower than in Experiment 1. Logistic regression results predicting accuracy are shown in **Table 5**: the only significant effect interacting with the Paraphrase factor is that Level-3 embeddings are more difficult. A $p = .03$ negative interaction between Level 3 and Language = German is likely not meaningful, as it applies to the Paraphrase = Yes condition.

5. General discussion

5.1 Replication status of Bach et al. (1986)

In a broad sense, the above results replicate the findings of Bach et al. (1986) in a different modality: we find greater difficulty in German than in Dutch. However, the particulars of the results diverge from Bach et al. (1986). The cross-linguistic difference in difficulty emerges at embedding level 2 in our results, whereas it emerged at level 3 in the original results. We find only mixed evidence that the cross-linguistic difference becomes stronger going from level 2 to level 3 embeddings.

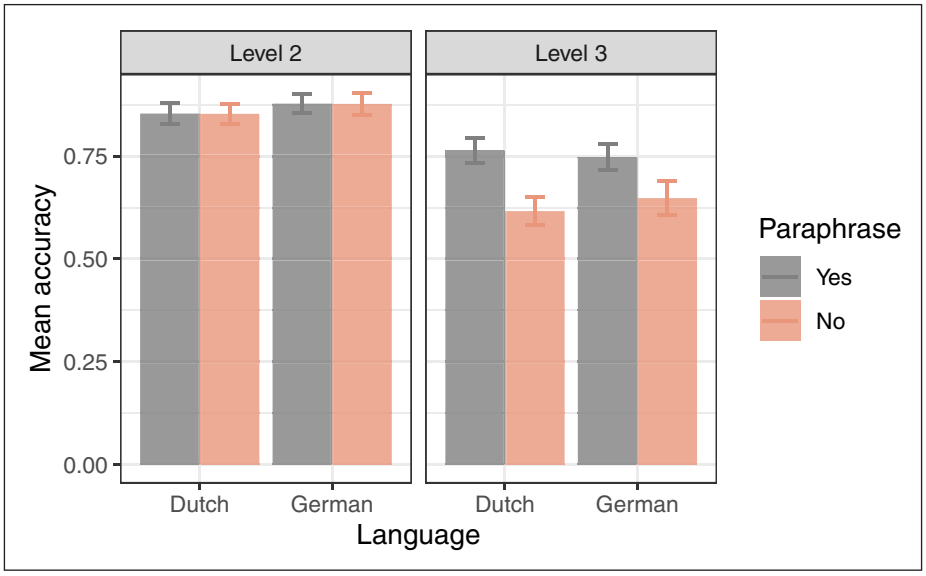


Figure 6: Average question response accuracy in each condition of Experiment 2. The difference between the orange (No-paraphrase) and gray (Paraphrase) bars indicates the decrease in response accuracy caused by the use of an embedded structure, which is the effect of interest. The error bar shows the standard error of the mean response accuracy.

Table 5: The estimated effects of embedding level, language, and paraphrase condition on question response accuracy. The estimates were obtained from a mixed-effects logistic regression fitted to the question response data. The two theoretically important interaction effects are highlighted with a gray shade.

Effect	**Estimate**	**Std.Err.**	***z*-value**	***p*-value**
Intercept	2.16	0.25	8.64	<0.01
Level 3	–0.61	0.15	–4.10	<0.01
Paraphrase = No	0.13	0.16	0.83	0.41
Language = German	0.35	0.19	1.83	0.07
Level 3 × Paraphrase = No	–1.08	0.20	–5.29	<0.01
Level 3 × Language = German	–0.47	0.21	–2.21	0.03
Paraphrase = No × Language = German	–0.10	0.24	–0.41	0.68
Level 3 × Paraphrase = No × Language = German	0.33	0.31	1.08	0.28

The most noticeable difference is that we do not find any interesting effects in the comprehension question accuracy rates. Bach et al. (1986) found a difference in comprehension accuracy between languages for 2 and 3 levels of embedding, although only for the infinitive form of the German materials. It is possible that the difference found by Bach et al. (1986) is due to the use of infinitive verb forms, which may be dispreferred by speakers; in contrast, we only

analyze data from the preferred verb forms by participant. The difference may also arise due to the use of open-ended questions in the original work, whereas we used multiple-choice questions probing different dependencies.

5.2 Frequency-based explanation?

One natural question is whether the difference in processing between German and Dutch arises due to the differences in frequency of these verb sequences in the two languages. Indeed, a large number of processing phenomena across languages seem to be reducible to effects of surprisal (Frank and Bod, 2011; Futrell et al., 2020a; Hale, 2001; Levy, 2008), which reflects language-internal statistics (but see Huang et al., 2024; van Schijndel and Linzen, 2021). In the case of crossing dependencies, Levy et al. (2012) find that the difficulty of right-extraposed relative clauses is a function of their surprisal given the main verb.

To investigate this possibility for the Dutch and German structures under study, we collected corpus counts from web-based corpora (Schäfer, 2015; Schäfer and Bildhauer, 2012) for the sequences of verbs found in our critical experimental items. We used the actual lexical items found in our experiments in order to test the simplest frequency-based account: that the acceptability difference is due to frequency of exposure alone. In Dutch and German web text (252.8 million sentences of Dutch and 607.7 million sentences of German), we find only 3 instances of the critical verb trigrams in Dutch, and 0 in German, suggesting extreme rarity for the level-3 embedded structures. For verb bigrams, reflecting the frequency of level-2 embedded structures, we find 13,077 in Dutch and 47,326 in German, suggesting that the level-2 structures are significantly *more* frequent in German than Dutch (in a χ-squared test on the proportion of verb bigrams per sentence in the Dutch vs. German corpora, we find $\chi^2 = 1738.9, p < .001$). The lower frequency of verb trigram sequences in German may explain the higher difficulty ratings for level-3 embeddings, but the higher frequency of verb bigrams in German is not reflected in lower difficulty for level-2 embeddings.

In any case, a frequency-based explanation of the processing difference between Dutch and German cannot provide a complete account, because it leaves open the question of *why* some structures are more frequent than others to begin with. The low frequency of multiple embedded verb phrases may be a result of processing difficulty, rather than a cause of it, if speakers avoid using constructions that engender difficulty.

5.3 Consequences for psycholinguistic theories

Our results compound the evidence that there is no particular processing difficulty in comprehension associated with crossing dependencies. Although this result has been found before, it is still somewhat remarkable given the formal complexity associated with such dependencies.

The results are, however, compatible with accounts of processing difficulty based on dependency locality (Gibson, 1998, 2000), where difficulty occurs when a word must be integrated

with a head or dependent far away from it in the linear order of words, and with automaton-based theories where processing difficulty is associated with the amount of time an item must be stored on a suitably structured stack. As shown in **Figure 1**, the nested German structures involve longer (maximum) dependency lengths than the cross-serial Dutch structures. Note that it is not always the case that crossing dependencies create shorter dependency lengths, as they do here: for example, topicalization (a form of *wh*-movement) often increases dependency length.

The generality of our results needs some qualification, however, because we studied only one construction, and we were comparing across two languages. Within Dutch, there is no alternative word order available to express exactly the dependency structure in example (1a). It is thus possible that the processing system for Dutch is 'fine-tuned' toward such word orders because they occur inevitably within the language, and this fine-tuning masks any underlying difficulty associated with the crossing dependencies. We note however, that the German order is equally inevitable given the dependency structure, and the frequency of the relevant verb trigrams is vanishingly low in both languages.[6]

5.4 Consequences for typology: Why are crossing dependencies rare?

While crossing dependencies are widely attested across languages, they are rare within languages. The underlying explanation for this rarity remains unknown. A common view is that there are hard formal restrictions on syntactic patterns that humans can learn: that is, a universal constraint on mental representations of grammars limits the occurrence of crossing structures in a sentence (Chomsky, 1965; Joshi et al., 1991; Silva et al., 2022). However, artificial language learning experiments do not support a preference for context-free structures (Öttl et al., 2015). Our results argue against a view where crossing dependencies are avoided because they are associated with online comprehension difficulty. While avoidance of long dependencies (Ferrer-i-Cancho, 2004; Futrell et al., 2020b; Hawkins, 1994; Liu et al., 2017) does reduce the rate of crossing dependencies on average (Ferrer-i-Cancho, 2006), it does not fully explain their observed rarity (Yadav et al., 2021). Nevertheless, processing-based explanations are not yet fully ruled out: there may still be processing difficulty associated with crossing dependencies during production.

6. Conclusion

We have revisited the classic psycholinguistic results on cross-serial versus nested dependencies from Bach et al. (1986). Our findings broadly support the conclusion that long nested dependencies in German engender more processing difficulty than the equivalent cross-serial dependencies in Dutch. The results support a dependency locality account for the processing difficulty of crossing dependencies.

[6] Interestingly, in less standardized varieties of German and in other verb clusters in Dutch, there is variation in word order for these kinds of structures (Shieber, 1985; Barbiers, 2005; Barbiers et al., 2018).

Data accessibility statement

All code and data are publicly available at https://osf.io/qbc5f/. Supplementary files with the German verb-form pretest, regression modeling, and power analysis details are available at https://osf.io/93vq6.

Ethics and consent

Experiments were performed following UC Irvine IRB protocols. Participants gave informed consent for participation.

Acknowledgments

We thank Walter Haesereyn for his input on Dutch syntax and Volker Struckmeier for his input on German syntax during stimuli preparation. We also thank Inés Schönmann for checking the German sentence stimuli. We are thankful to three anonymous reviewers as well whose comments considerably improved the quality of this manuscript.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

HY, SF, RF, and SH conceived the study and its design. SF and RF wrote items with input from all authors. HY conducted data analysis with input from all authors. All authors wrote and edited the manuscript.

References

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250. https://doi.org/10.1007/BF01067217

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bach, E., Brown, C., & Marslen-Wilson, W. D. (1986). Cross and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, *1*(4), 249–262. https://doi.org/10.1080/01690968608404677

Barbiers, S. (2005). Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics. In L. Cornips & K. Corrigan (Eds.), *Syntax and variation. Reconciling the biological and the social* (pp. 233–264). Benjamins. https://doi.org/10.1075/cilt.265.14bar

Barbiers, S., Bennis, H., & Dros-Hendriks, L. (2018). Merging verb cluster variation. *Linguistic Variation*, *18*(1), 144–196. https://doi.org/10.1075/lv.00008.bar

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

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1178–1198. https://doi.org/10.1037/a0024194

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Bock, K., Levelt, W., & Gernsbacher, M. A. (2002). Language production: Grammatical encoding. In *Psycholinguistics: Critical concepts in psychology* (pp. 405–452). Routledge.

Bresnan, J. W. (1982). *The mental representation of grammatical relations*. MIT Press.

Bresnan, J. W., Kaplan, R., Peters, S., & Zaenen, A. (1982). Cross-serial dependencies in Dutch. *Linguistic Inquiry*, *13*, 613–635. https://doi.org/10.1007/978-94-009-3401-6_11

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. https://doi.org/10.21236/AD0616323

Clifton, C., & Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In G. N. Carlson & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 273–317). Springer. https://doi.org/10.1007/978-94-009-2729-2_8

De Santo, A. (2020). MG parsing as a model of gradient acceptability in syntactic islands. In *Proceedings of the Society for Computation in Linguistics* (pp. 59–69).

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, *37*, 378–394. https://doi.org/10.1111/cogs.12021

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*, 056135. https://doi.org/10.1103/PhysRevE.70.056135

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters*, *76*(6), 1228. https://doi.org/10.1209/epl/i2006-10406-0

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., & Esteban, J. L. (2018). Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, *493*, 311–329. https://doi.org/10.1016/j.physa.2017.10.048

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

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. https://doi.org/10.1177/0956797611409589

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., Trompenaars, T., Lewis, R. L., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578. https://doi.org/10.1111/cogs.12247

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.005

Frazier, L. (1987). Sentence processing: A tutorial review. In Coltheart, M., editor, *Attention and performance 12: The psychology of reading* (pp. 559–586). Lawrence Erlbaum Associates, Inc.

Futrell, R., Gibson, E., & Levy, R. P. (2020a). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*, e12814. https://doi.org/10.1111/cogs.12814

Futrell, R., Levy, R. P., & Gibson, E. (2020b). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–413. https://doi.org/10.1353/lan.2020.0024

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.32614/CRAN.package.arm

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (pp. 95–126). MIT Press. https://doi.org/10.7551/mitpress/3654.003.0008

Graf, T., Monette, J., & Zhang, C. (2017). Relative clauses as a benchmark for minimalist parsing. *Journal of Language Modelling*, *5*(1), 57–106. https://doi.org/10.15398/jlm.v5i1.157

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies* (pp. 1–8). https://doi.org/10.3115/1073336.1073357

Havelka, J. (2007). Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 608–615). Association for Computational Linguistics.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press. https://doi.org/10.1017/CBO9780511554285

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic

disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

Huck, G. J., & Na, Y. (1990). Extraposition and focus. *Language*, *66*, 51–77. https://doi.org/10.1353/lan.1990.0023

Husain, S., & Yadav, H. (2020). Target complexity modulates syntactic priming during comprehension. *Frontiers in Psychology*, *11*, 454. https://doi.org/10.3389/fpsyg.2020.00454

Joshi, A. K. (1985). Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical Perspectives* (pp. 190–205). Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.007

Joshi, A. K. (1990). Processing crossed and nested dependencies: An automaton perspective on the psycholinguistic results. *Language and Cognitive Processes*, *5*, 1–27. https://doi.org/10.1080/01690969008402095

Joshi, A. K., Vijay-Shanker, K., & Weir, D. J. (1991). The convergence of mildly context-sensitive grammar formalisms. In P. Sells, S. Shieber, & T. Wasow (Eds.), *Foundational issues in natural language processing* (pp. 31–81). MIT Press.

Kobele, G. M. (2006). *Generating copies: An investigation into structural identity in language and grammar* [Doctoral dissertation]. University of California Los Angeles.

Kobele, G. M., Gerth, S., & Hale, J. (2013). Memory resource allocation in top-down minimalist parsing. In G. Morrill & M.-J. Nederhof (Eds.), *Formal grammar* (pp. 32–51). Springer. https://doi.org/10.1007/978-3-642-39998-5_3

Kuhlmann, M. (2013). Mildly non-projective dependency grammar. *Computational Linguistics*, *39*(2), 355–387. https://doi.org/10.1162/COLI_a_00125

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

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

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence Processing* (pp. 78–114). Hove: Psychology Press.

Levy, R., Fedorenko, E., Breen, M., & Gibson, T. (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

Momma, S. (2021). Filling the gap in gap-filling: Long-distance dependency formation in sentence production. *Cognitive Psychology*, *129*, 101411. https://doi.org/10.1016/j.cogpsych.2021.101411

Öttl, B., Jäger, G., & Kaup, B. (2015). Does formal complexity reflect cognitive complexity? Investigating aspects of the Chomsky hierarchy in an artificial language learning study. *PloS ONE*, *10*(4), e0123059. https://doi.org/10.1371/journal.pone.0123059

Rambow, O., & Joshi, A. K. (1994). A processing model for free word-order languages. In J. Charles Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing*. Psychology Press.

Rambow, O., & Satta, G. (1994). A rewriting system for natural language syntax that is non-local and mildly context sensitive. In *North-Holland Linguistic Series: Linguistic Variations* (volume 56, pp. 121–130). Elsevier.

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *Proceedings of the 14th International Conference on Computational Linguistics* (pp. 191–197). https://doi.org/10.3115/992066.992098

Rochemont, M. S., & Culicover, P. W. (1990). *English focus constructions and the theory of grammar*. Cambridge University Press.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* (pp. 28–34). Institut für Deutsche Sprache.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 486–493).

Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2015). Syntactic complexity effects in sentence production. *Cognitive Science*, *39*(3), 559–583. https://doi.org/10.1111/cogs.12168

Seki, H., Matsumara, T., Fujii, M., & Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, *88*(2), 191–229. https://doi.org/10.1016/0304-3975(91)90374-B

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In *The formal complexity of natural language* (pp. 320–334). Springer. https://doi.org/10.1007/978-94-009-3401-6_12

Silva, S., Inácio, F., Rocha e Sousa, D., Gaspar, N., Folia, V., & Petersson, K. M. (2022). Formal language hierarchy reflects different levels of cognitive complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*, 642–660. https://doi.org/10.1037/xlm0001182

Stabler, E. P. (1994). The finite connectivity of linguistic structure. *Perspectives on sentence processing* (pp. 303–336).

Staub, A., Foppolo, F., Donati, C., & Cecchetto, C. (2018). Relative clause avoidance: Evidence for a structural parsing principle. *Journal of Memory and Language*, *98*, 26–44. https://doi.org/10.1016/j.jml.2017.09.003

Torr, J., Stanojević, M., Steedman, M., & Cohen, S. B. (2019). Wide-coverage neural A* parsing for Minimalist Grammars. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2486–2505). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1238

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988. https://doi.org/10.1111/cogs.12988

Weir, D. J. (1988). *Characterizing mildly context-sensitive grammar formalisms* [Doctoral dissertation]. University of Pennsylvania.

Yadav, H., Husain, S., & Futrell, R. (2021). Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, *7*(s3). https://doi.org/10.1515/lingvan-2019-0070

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466.