Article

Yue Zou and Hao Lin*

Word order flexibility affects complementizer omission: a cross-linguistic investigation

https://doi.org/10.1515/cllt-2024-0121
Received November 18, 2024; accepted June 24, 2025; published online July 8, 2025

Abstract: It has been observed in many languages that the complementizer at the beginning of a complement clause (CC) can be optionally omitted. Several accounts of language processing have been proposed to explain the phenomenon. Specifically for the ambiguity avoidance account, however, mixed results have been reported in previous research. In our study, we investigate whether the cross-linguistic difference in omission rates can be explained by the ambiguity avoidance account at a general strategic level. Our hypothesis is that, in order to avoid ambiguities, users of languages with a more flexible word order should be more reluctant to omit the complementizer. To test it, we conducted an in-depth analysis in three languages that are different in the degree of word order flexibility – English, German, and Russian, and a broad analysis with 23 languages from 13 genera. The results provide complementary evidence that word order flexibility indeed emerges as an important predictor of complementizer omission.

Keywords: complementizer omission; cross-linguistic difference; ambiguity avoidance; word order flexibility

1 Introduction

Linguistic alternations have drawn the attention of researchers from various subfields in linguistics and have emerged as a topic that has been extensively studied over the past few decades (Gries 2017). Informally, a linguistic alternation can be defined as the phenomenon of language users choosing between "pairs of semantically more-or-less equivalent expressions" (Gries and Stefanowitsch 2004: 97). The

***Corresponding author: Hao Lin**, Institute of Language Sciences, Shanghai International Studies University, Shanghai, China, E-mail: linhao@shisu.edu.cn
Yue Zou, Institute of Language Sciences, Shanghai International Studies University, Shanghai, China

variety of alternations in languages reflects a remarkable degree of flexibility in how we can encode information with linguistic forms or signals.

One pervasive type of alternation is termed reduction, which refers to the choice of language users between the (a) full and (b) more or less reduced forms or signals (Jaeger and Buz 2018). At the phonetic level, for example, syllabic duration in English has been reported to be inversely related to language redundancy (Aylett and Turk 2004). Morphologically, optional case-marking in Japanese has been argued to be affected by communicative pressures (Kurumada and Jaeger 2015). In the present study, our focus is on a cross-linguistic syntactic phenomenon: complementizer omission. It has been observed in many languages that the complementizer at the beginning of a complement clause (CC) can be optionally omitted (English: Roland et al. 2006; Jaeger 2010; Spanish: Yoon 2015; Russian: Zou and Lin 2024). An example for English is provided in (1):

1. English
   He doesn't think (that) he did anything wrong.

A large number of studies have been conducted to investigate complementizer omission, and several accounts of language processing have been proposed to explain the phenomenon. Among the most influential accounts are the availability-based production account, the uniform information density account, and the ambiguity avoidance account. Apart from language processing mechanisms, complementizer omission has also been argued to be under the influence of grammaticalization.

The idea of availability-based production is based on the Principle of Immediate Mention, which states that "production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible" (Ferreira and Dell 2000: 299). Several variables have been proposed to be related to availability-based production, including (a) the coreferentiality between subjects in matrix and complement clauses and (b) the number of disfluencies at the CC onset. Specifically, a complementizer is more likely to be omitted when the matrix subject and the CC subject refer to the same entity, and when there is no sign of disfluency at the CC onset. In addition, an effect of the frequency of matrix verbs is also compatible with the availability account (Jaeger 2010): less frequent matrix verbs tend to have higher rates of using an overt complementizer. According to Jaeger, the high processing load associated with the production of the less frequent verbs may "spill over" to the CC onset, thereby leading to the production of an overt complementizer.

The uniform information density (UID) account has received wide support since its first proposal in Jaeger (2010: 25), where it is formulated as follows: "within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density)." Regarding the complementizer omission phenomenon, the UID account predicts that the higher the information

density at the CC onset, the more likely language users are to produce an overt complementizer. In Jaeger (2010), the information density at the CC onset is estimated with the subcategorization preference of matrix verbs. In Wulff et al. (2018), the same hypothesis was tested with a different variable (i.e., surprisal) that measures information density at a finer granularity. In either case, the UID account was strongly supported: speakers tend to lower the information density by producing an overt complementizer if the CC onset is highly surprising.

Unlike the above-mentioned two accounts, mixed results have been reported in previous research on the ambiguity avoidance mechanism. Based on previous observations on linguistic reductions, Frazier (1985) proposed the Impermissible Ambiguity Constraint: constructions that would lead to ambiguities or misanalyses on every occurrence tend to be prohibited. Specifically for complementizer omission, it has also been argued that the use of *that* is tailored to avoid temporary ambiguities (Temperley 2003). One example showing how an overt complementizer can avoid temporary ambiguities is shown in (2), where the matrix verb *know* allows both CCs and direct objects. The omission of the complementizer can thus lead to a garden path, in which the CC subject *the story* is incorrectly understood as a direct object of the verb.

2. She knew the story was true.
   She knew that the story was true.

A strong pattern of complementizer omission in English that supports the ambiguity avoidance account was found in Elsness (1984): an overt complementizer is less likely to be used when the CC subject is a pronoun. It was pointed out that in English, unlike nouns, some pronouns have distinct nominative and accusative forms. Therefore, a pronominal CC subject is less likely to be mistaken for a direct object of the matrix verb. At the same time, the ambiguity avoidance account also predicts a difference in omission rates between cases where the embedded subject is a case-distinguishing pronoun and cases where it is not. However, the predicted difference was observed neither in Elsness (1984) nor in Ferreira and Dell (2000). In other words, the evidence for and against the account is still inconclusive and further exploration is needed. In Temperley (2003), it is argued that the general tendency to omit complementizers in cases with pronominal CC subjects is itself an ambiguity avoidance mechanism at the syntactic level.

Broadly, the three above-mentioned accounts all align with the principle of communicative efficiency. A large amount of evidence has been provided cross-linguistically for the argument that language users tend to act efficiently, saving effort for processing and articulation, and that language structure and use reflect this tendency (for review, see Gibson et al. 2019; Jaeger and Buz 2018; Levshina and Moran 2021). In Kachakeche et al. (2021), for instance, it was shown that communicative pressures influence the use of adjectives, such that prenominal languages use

adjectives at a higher rate compared to postnominal languages. In previous research on complementizer omission, however, only one specific language (primarily English) was studied at a time, and only language-internal factors have been explored. Here, "language-internal" factors contrast with "language-level" factors. Regarding complementizer omission, it has been shown in many studies that within a given language, various "language-internal" factors play a role in the omission of the complementizer (e.g., verb frequency, length of the CC). Cross-linguistically, although languages differ greatly in their overall omission rates, the question of which "language-level" properties may have contributed to the difference has rarely been explored. To the best of our knowledge, no cross-linguistic investigation on complementizer omission has been conducted. In the present study, we revisit the ambiguity avoidance account by analyzing complementizer omission cross-linguistically. Our study attempts to investigate the cross-linguistic difference in omission rates and to check whether it can be explained by the ambiguity avoidance account at a general strategic level.

Our study is motivated by the following hypothesis: word order flexibility of languages can influence the tendencies of their users to avoid ambiguities. Since human languages differ in the amount of word order flexibility they permit, we expect different languages to show different degrees of tendencies to avoid ambiguities. For instance, compared with English, German has a more flexible word order. In cases where an overt complementizer is omitted in German, the listener may have more difficulty in determining the grammatical role of the noun following the matrix verb. In German, the noun that follows the matrix verb can not only be its direct object and a CC subject but also a CC object, as shown in (3):

3. *Ich weiß, das Geheimnis hat Lucy dir schon verraten.*
   I know the secret have Lucy you already tell
   'I know Lucy has already told you the secret.'

Cacoullos and Walker (2009) offered a review of the treatment of the complementizer *that* in English in the prescriptive grammatical tradition and concluded that prescriptive grammarians' main argument for retaining *that* is to ensure clarity. More generally, it can be argued that the use of overt complementizers (in any case) is itself an ambiguity avoidance mechanism. We propose that when the complementizer is omitted, instances of complex clauses in languages with a more flexible word order should be potentially more ambiguous than those in languages with a more fixed word order. Therefore, users of languages with more flexible word order should be more reluctant to omit complementizers. Our prediction can be formulated as follows:

> Word order flexibility affects complementizer omission, such that languages with a less flexible word order have a higher tendency to omit complementizers.

In order to reliably test our prediction, we conducted two corpus-based analyses: an in-depth analysis of three languages, and a broad analysis of 23 languages from 13 genera (language branches). The first (in-depth) analysis includes three languages that clearly differ in their word order flexibility: English, German, and Russian. An example is provided for each of the latter two languages in (4)–(5). The in-depth analysis enables us to accurately extract relevant cases and to compare the omission rates of languages while controlling for other potential influencing factors.

4. German

Ich	*denke,*	*dass*	*das*	*funktionieren*	*kann.*
I	think	that	this	work	can.
Ich	*denke,*	*das*	*kann*	*funktionieren.*	
I	think	this	can	work.	

'I think (that) this can work.'

5. Russian

Ja	*dumaju,*	*(čto)*	*eto*	*vpolne*	*estestvenno.*
I	think	(that)	this	completely	natural.

'I think (that) this is perfectly natural.'

The broad analysis is conducted with 23 languages from the Universal Dependencies (UD) Treebank (version 2.14; Nivre et al. 2020). In the broad analysis, we quantitatively estimated the overall omission rates of languages and their word order flexibility. Results from the two parts of analysis complement each other and provide strong support for our hypothesis.

The remainder of this paper is structured as follows. Section 2 focuses on the dataset, method, and results of the in-depth analysis and ends with a few of its limitations. Section 3 reports the broad analysis that attempts to address the limitations. In Section 4, we summarize our findings and discuss a few general issues related to the results and the choices made in our study. Section 5 concludes the study.

2 In-depth analysis of three languages

2.1 Dataset

We downloaded corpora of the three languages (one for each language) from the Leipzig Corpora Collection (Goldhahn et al. 2012; https://wortschatz.uni-leipzig.de/en). The collection contains corpora that were constructed from different sources, with

different methods, and at different time points. For the sake of consistency, for each language, we chose a news corpus constructed through web-crawling in 2019 that contains one million sentences.

It has long been suggested that different lexical items often have different preferences for certain constructions (because of their semantic or information-structural characteristics) (Gries and Stefanowitsch 2004; Stefanowitsch and Gries 2003). The lexical effect on complementizer omission has been reported to be significant in many studies (Cacoullos and Walker 2009; Jaeger 2010). For instance, in Dor (2005), it was mentioned that the omission of the complementizer is unnatural under manner-of-speaking verbs in English. In the present study, we assume that lexical semantics have the same effect direction in different languages. More specifically, if manner-of-speaking verbs disprefer complementizer omission in English, then so do their translational equivalents in other languages. In order to minimize the effect of verb semantics, we first chose two English verbs for analysis, and then got their translational equivalents from the other two languages. For English, we chose *know* and *believe*, which are the third and sixth most frequent verbs in the sample used in Jaeger (2010). Compared with the top two most frequent verbs (i.e., *think* and *guess*), *know* and *believe* much more actively participate in the alternation. Their equivalents are *wissen* and *glauben* in German, and *znat'* and *verit'* in Russian. Importantly, all six verbs allow both CCs and direct objects (the Russian verb *verit'* takes nouns in the dative case, while the other five verbs take nouns in the accusative case). In other words, for all the chosen verbs, temporary ambiguities will arise when the complementizer is omitted.

For each language, to extract the target complex sentences with or without a complementizer, we first picked out sentences containing an inflected form of the target verbs. For instance, we considered 11 forms of the Russian verb *znat'* (10 inflected forms and one infinite form; a list of the considered forms is provided in the Appendix). Participle forms in Russian (e.g., *znajuščij*) are not considered due to their extremely low frequency.

We then conducted part-of-speech (POS) tagging and dependency parsing on the picked-out sentences with the Stanza package (Qi et al. 2020) in Python. For simplicity, we only included declarative sentences (ending with a full stop) where the optional complementizer is used directly after the target verbs. As reported in Jaeger (2010), the matrix verb directly precedes the CC in 93.5 % of cases in his sample. Therefore, it can be estimated that only a small proportion of cases would be excluded from our analysis. An example of the English annotated target sentences is provided in Figure 1:

In our analysis, a complex sentence is tagged as having a complementizer if it satisfies the following conditions: (a) one of the target verbs is the root of the sentence; (b) the target verb is the head of the dependency relation *ccomp*; (c) the sentence has at least two subjects (one occurring before the matrix verb and in the

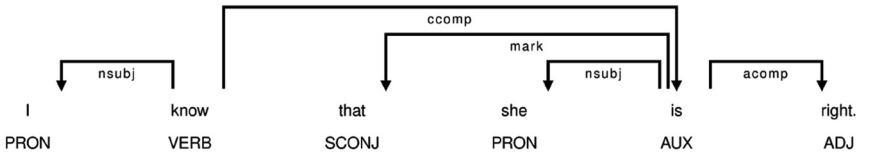

Figure 1: An English complex sentence with an overt complementizer, annotated with POS and dependency information.

matrix clause (MC), and the other occurring after the matrix verb and in the CC; the dependency relation with the main verbs in the respective clauses being *nsubj*); (d) the word following the target verb is a subordinating conjunction. On the other hand, a complex sentence is tagged as having no overt complementizer if it satisfies conditions (a), (b), and (c), but does not contain a subordinating conjunction. Items from the three languages were extracted with the same rules. For condition (d), to reduce the number of false hits, we specified the exact complementizer (subordinating conjunction) in the three languages: *that* in English, *dass* in German, and *čto* in Russian. Finally, we excluded sentences with interrogative pronouns as CC subjects. In total, our dataset contains 5,951 items. For each language, the number of items with and without a complementizer is presented in Table 1. The number of cases with each verb is shown in Table 2. It can be seen that the subcategorization bias of the verbs,

Table 1: The number of extracted items for each language. The first and second columns indicate the number of items with and without an overt complementizer, respectively. The last column provides the omission rate.

Language	Present	Absent	Sum	Omission rate
English	1,290	2,835	4,125	0.687
German	569	411	980	0.416
Russian	609	237	846	0.280

Table 2: Raw frequency, number of usages as matrix verbs, subcategorization preference, and estimated tendency for complementizer omission of the target verbs.

Verb	Raw frequency	# of target cases	Subcategorization
know	17,065	2,269	0.133
believe	7,248	1,856	0.256
wissen	5,028	521	0.104
glauben	2,819	459	0.163
znat’	6,610	679	0.103
verit’	1,264	167	0.132

which is calculated by dividing the number of target cases by their raw lemma frequency, is consistently low.

2.2 Methods

2.2.1 Main predictor of interest

In the first analysis, the main predictor of interest in our study is word order flexibility at the language level. Each language allows a certain degree of flexibility. Since only three languages are included in this analysis, their relative flexibility can be determined qualitatively from previous studies. First, in the marking of agent-patient relations, English employs a relatively fixed word order (i.e., subject-verb-object; Matthews et al. 2005). In contrast to English, German allows a relatively free word order (Bornkessel et al. 2002). An example is provided below in (6):

6. *Ich glaube, dass den Jäger der Gärtner beruhigte.*
   I think that the hunter the gardener calmed.
   'I think that the gardener calmed the hunter (Bornkessel et al. 2002: B22).'

As can be seen, although German also has a default subject-before-object order, the reversed order is also acceptable because the subject and object have already been marked with morphological cases (nominative and accusative, respectively). Despite the relative free order of subject and object, German still needs to follow the general rule about the verb position: when the complementizer is present, the finite verb has to be in the clause-final position (as in Example 6); when the complementizer is absent, the finite verb should take the second position (Brandt et al. 2010).

Russian belongs to the eastern branch of Slavic languages (Siewierska and Uhlířová 1998). Compared with German, it exhibits an even higher level of flexibility (Mykhaylyk et al. 2013): not only can the order between subject and object be freely reversed, the position of the finite verb is also not restricted. Therefore, the relative order of flexibility of the three languages is as follows: Russian > German > English. The prediction of our hypothesis can be reformulated as follows:

> According to our hypothesis, Russian should omit the complementizer least frequently, English should omit the complementizer most frequently, while German should be somewhere in the middle.

2.2.2 Controls

In order to rigorously test our hypothesis, it is necessary to control for other effects known to affect complementizer omission (Jaeger 2010). The control variables considered in our studies are briefly introduced below.

TYPE OF MATRIX SUBJECT. In analyzing complementizer omission in English, Thompson and Mulac (1991) found that the first and second person pronouns (i.e., *I* and *you*) disfavor the presence of complementizer more than other matrix subjects, which they explained by the higher frequency of *I* and *you* in discourse and their capacity to express epistemicity or subjectivity. As one of the most often considered variables, its effect on complementizer omission has been empirically supported in many languages (French: Liang et al. 2021; Danish: Boye and Poulsen 2011; English: Cacoullos and Walker 2009; Jaeger 2010). In our analysis, TYPE OF MATRIX SUBJECT is coded as a categorical variable with 5 levels: (1) first person singular pronoun, (2) first person plural pronoun, (3) second person pronoun, (4) third person pronouns, and (5) nouns. The 2nd to 4th levels are grouped into one level ("other pronouns") because of their relatively small number of cases.

TYPE OF CC SUBJECT. As mentioned above, Elsness (1984) found a strong pattern of complementizer omission in English that *that* is more likely to be omitted when the CC subject is a pronoun. At the same time, the pattern is also compatible with the availability account. The use of nouns rather than pronouns indicates that the corresponding referent is newly introduced and not readily available for production. Therefore, the availability account also predicts a higher omission rate when the CC subject is a pronoun. In our analysis, TYPE OF CC SUBJECT is coded as a binary variable with two levels by looking at the POS tag of the CC subject: pronominal and nominal.

LENGTH OF CC. In English, sentences where grammarians allow complementizer omission are relatively simple syntactically (Cacoullos and Walker 2009). In previous research, syntactic complexity has been operationalized with variables that measure the length of different parts of a complex clause. LENGTH OF CC is one of the most straightforward ways of encoding complexity and its effect on complementizer omission has been reported to be highly significant: an overt complementizer is more likely to be present when the CC is longer (Gries 2021; Jaeger 2010). In our study, LENGTH OF CC is a numeric variable and is measured by counting the number of words in the CC.

SURPRISAL. As an information-theoretic notion, SURPRISAL quantifies "how uncertain one would be about observing some event – how 'surprising' that event would be – given a known probability distribution of related events" (Wulff et al. 2018: 107). In analyzing complementizer omission, it was included in Wulff et al. (2018) and Gries (2021) to measure how surprising the transition from the matrix to complement clause would be if no complementizer had been used. In Wulff et al. (2018) and Gries

(2021), conditional surprisal of the first word in the CC was measured by considering "the last word in the MC prior to the clause juncture, regardless of whether the complementizer separates the words or not." The operationalization of conditional surprisal is based on Equation (1), where *x* is the first word in the CC and *y* is the last word in the MC. Since both SURPRISAL and SUBCATEGORIZATION PREFERENCE measure information density at the CC onset, and SURPRISAL is at a finer level of granularity, we decide to only include SURPRISAL in our analysis.

$$S_c(x|y) = -\log_2 P(x|y) \quad (1)$$

FREQUENCY. The role of frequency in linguistic variation and change has attracted the attention of researchers for decades (Bybee 2003; Cacoullos and Walker 2009; Fenk-Oczlon 2001; Zipf 1949). In English, frequency has been argued to propel the reduction of subject-verb combinations to discourse formulas such as "I think" and "I guess" (Thompson 2002). Generally, verbs with higher frequency have been reported to have a higher tendency to omit the complementizer. In our study, the raw frequency of our target verbs is calculated by summing up the frequencies of their various forms. The raw frequency of the target verbs is shown in the first numeric column of Table 2.

Apart from the above-mentioned variables, a few others have also been proposed to have a potential effect on complementizer omission. However, their effects are relatively small and have not always been successfully replicated. For instance, Ferreira and Dell (2000) reported that coreferentiality of the matrix and CC subject correlated with a higher omission rate, but the effect was found to be only marginally significant in Jaeger (2010) and was not replicated in Cacoullos and Walker (2009). Bolinger (1972) suggested that the omission of the complementizer is more likely when the matrix and complement clauses agree in polarity (negative or affirmative). The effect was also not replicated in Cacoullos and Walker (2009). Since their coding requires much manual work and their effects are not the primary focus of our study, we decide to not include them in this analysis.

2.3 Results

We used a mixed-effects logistic regression model to test the effects of our predictors with the lme4 package (Bates et al. 2015) in R (R Core Team 2023). The response is a binary variable with two levels: present and absent. The fixed-effects structure (FES) of the model contains our main predictor (i.e., language) and five controls, as introduced above. LENGTH OF CC and FREQUENCY are logged, and all the continuous variables are centered and scaled before being included in the model. Additionally, we included VERB, which has 6 levels, into the random-effects structure (RES) of our

model by adding by-verb random intercepts. Issues of multicollinearity and overdispersion are checked with helper functions provided in Gries (2013).

Results of our regression analysis indicate that all the predictors have a significant effect on complementizer omission. No issue of multicollinearity or overdispersion was found. Coefficients and results of significance tests for predictors are presented in Table 3. Marginal R^2 and Conditional R^2 of the model are 0.236 and 0.251, respectively.

As predicted by the ambiguity avoidance account, there was a significant effect of LANGUAGE on complementizer omission. Specifically, the predicted probability of complementizer omission is the lowest in English and is significantly lower than that in German ($\widehat{\beta}$ = 0.770, z = 2.115, p = 0.034). The probability of omission in German is significantly lower than that in Russian ($\widehat{\beta}$ = 0.751, z = 2.536, p = 0.011). The effect holds even while other variables are controlled for. Figure 2 illustrates the effect of LANGUAGE. In other words, the language with the least flexible word order omits the complementizer most frequently, and the language with the most flexible word order omits the complementizer least frequently.

We then used likelihood ratio tests (LRTs) to check the contributions of predictors in our model. Since the variable VERB in the RES contains language-related information, in order to compare the relative contribution of LANGUAGE to that of the other controls, we needed to refit a model with no RES. Results of LRTs indicate that LANGUAGE is the strongest predictor in terms of its contribution to the likelihood of the model ($\chi^2(2) = 275.600$, $p < 0.001$). Moreover, the improvement in model quality by including LANGUAGE is much higher than the improvement by including any of the controls. In other words, there exist large cross-linguistic differences in terms of complementizer omission, and their effects are much larger than those of language-internal

Table 3: Summary of results. Coefficient estimates, standard errors, *z* scores, and *p* values for all the predictors in the analysis.

	Estimate	SE	*z* score	*p* value
(Intercept)	−1.162	0.202	−5.751	<0.001
length	0.309	0.032	9.783	<0.001
surprisal	0.283	0.030	9.490	<0.001
frequency	−0.477	0.147	−3.234	0.001
language (en → de)	0.770	0.364	2.115	0.034
language (en → ru)	1.520	0.375	4.051	<0.001
CC subject (pronouns → nouns)	0.346	0.064	5.423	<0.001
matrix subject (first singular → other pronouns)	0.527	0.075	7.010	<0.001
matrix subject (first singular → nouns)	0.216	0.075	2.893	0.004

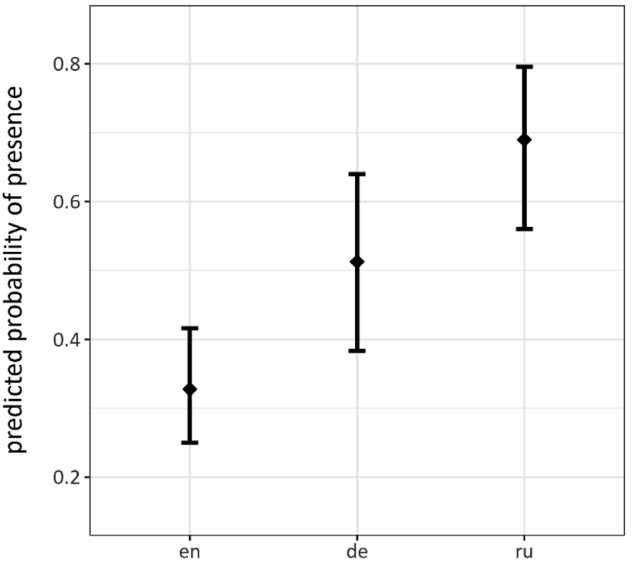

Figure 2: Effect of the predictor LANGUAGE on the predicted probability of using an overt complementizer. Error bars denote 95 % CIs.

factors. Word order flexibility at the cross-linguistic level emerges as the most important predictor of complementizer omission.

Next, we briefly discuss effects of the control variables in the model. As shown in Table 2, all the control variables have a significant effect on complementizer omission in the expected direction. First, the predicted probability of complementizer omission is significantly higher when the matrix subject is a first-person singular pronoun (*I* in English, *ich* in German, and *ja* in Russian) than when it is some other pronoun ($\widehat{\beta}$ = 0.527, z = 7.010, p < 0.001) or noun ($\widehat{\beta}$ = 0.216, z = 2.893, p = 0.004). The observed effect aligns with the grammaticalization account, according to which the frequent collocation of first person singular pronominal subject and complement-taking verb has been reanalyzed as an epistemic phrase that does not take subordinate CCs (Thompson and Mulac 1991). Second, the length of the CCs is strongly correlated with lower omission rates ($\widehat{\beta}$ = 0.309, z = 9.783, p < 0.001). Longer CCs prefer the use of an overt complementizer. Similar results have also been reported in Jaeger (2010) and Gries (2021). As mentioned in Jaeger (2010), the significant effect of LENGTH OF CC is surprising because it is not directly predicted by accounts of language processing. Moreover, it seems to contradict with the experimental evidence that language production is radically incremental (Brown-Schmidt and Konopka 2008). The observed effect of LENGTH OF CC in our study provides further support for the

explanation in Jaeger (2010: 42) that language users seem to have "at least heuristic weight or complexity estimates of material that is not yet phonologically encoded." Third, the effect of SURPRISAL ($\widehat{\beta}$ = 0.283, z = 9.490, p < 0.001) indicates that the predicted probability of using an overt complementizer increases as the degree of surprisal at the CC onset increases, thereby providing support for the UID account. The effect of FREQUENCY ($\widehat{\beta}$ = −0.477, z = −3.234, p < 0.001) shows that the higher the lemma frequency of the matrix verb, the more likely the complementizer is to be omitted, which is also in line with the grammaticalization account. Finally, the effect of the CC subject ($\widehat{\beta}$ = 0.346, z = 5.423, p < 0.001) is compatible with both the ambiguity avoidance account and the availability account.

In sum, the comparison of omission rates across three languages provides initial support for our hypothesis. However, it has a few obvious limitations. First, although the choice of semantically equivalent verbs across languages avoids the potential influence of lexical semantics, it runs the risk of misrepresenting the language-level overall omission rates. For instance, it is potentially possible that the two Russian verbs chosen in the analysis happen to have the lowest omission rates in the language. In other words, the observed difference in omission rates could be an artefact of our choice of verbs. Second, the number of languages included in the analysis is small. It could be the case that the three languages just happen to differ in their omission rates in the expected direction. Third, there is a difference between potential and effective flexibility. Although Russian is theoretically more flexible in word order than German, their actual flexibility still needs to be determined empirically. In any event, in order to gather more robust evidence for our hypothesis, we think it is important to include more languages, to compare their overall omission rates, and to quantitatively correlate omission rates with word order flexibility. In the following section, we attempt to address these concerns through our analysis of 23 languages from the UD Treebank.

3 Broad analysis of the UD Treebank

3.1 Dataset

We used the UD Treebank, which contains 283 treebanks from 161 languages to test our hypothesis at a broad level. If a certain language has several treebanks, we group them into a larger one. The treebank is developed with cross-linguistically consistent POS and dependency annotations, so the target structures can be easily extracted with a Python script.

A few steps need to be taken to decide on the exact structures to analyze and the languages to include. First, we found that we cannot rely on dependency structures and POS tags to exclusively extract all the sentences containing a CC without an overt complementizer when the CC object is a pronoun. It can be seen in Figure 3 that, when the CC starts with an interrogative pronoun *who*, the entire complex sentence has the exact same annotations as the one in Figure 4. Therefore, to circumvent this issue completely, we decided to only include in our broad analysis sentences where the CC subject is a noun. Second, we wanted to exclude the languages that do not allow complementizer omission. Ideally, languages with a 0 or 100 percent omission rate should be excluded. However, our extraction method, which is described below, still returns some false hits. For example, our analysis shows that Cantonese, which does not have a complementizer similar to *that* in English, has an omission rate of 0.96. French, which does not allow complementizer omission (in its standard variety), has an omission rate of 0.06. Therefore, we only included in our analysis languages with an omission rate between 0.1 and 0.9. In addition, Arabic, which also does not allow omission, is excluded. Finally, we excluded languages containing too few target structures (less than 20) and ancient languages (Berdicevskis et al. 2018). After taking the above-mentioned steps, our sample contains 23 languages from 13 genera.

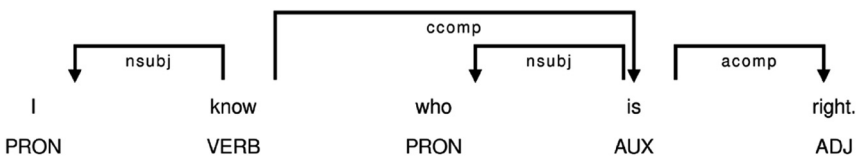

Figure 3: An English complex sentence with an interrogative pronoun as the CC subject, annotated with POS and dependency information.

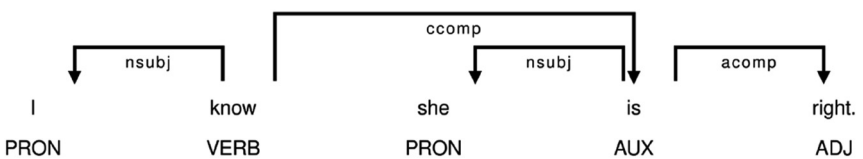

Figure 4: An English complex sentence without an overt complementizer, annotated with POS and dependency information.

3.2 Methods

In the broad analysis, we want to see whether word order flexibility can predict omission rates cross-linguistically, such that languages with more flexible word order will tend to omit the complementizer less frequently. The dependent variable is the language-level omission rate and is calculated by dividing the number of cases without an overt complementizer by the total number of target cases (sentences containing a CC with and without an overt complementizer). The target cases are extracted with the following three steps. First, a target sentence should have a verb token that is in the head position of the *ccomp* dependency relation. That verb token is identified as the matrix verb of the sentence. Second, between the matrix verb and its dependent, there should be a noun token that is in the dependent position of the *nsubj* dependency relation. It should be noted that, as mentioned above, pronoun tokens are excluded altogether. That noun token is identified as the CC subject. Finally, no adv token should lie between the matrix verb and CC subject. This step is taken to exclude cases like *I know how she left home*. To see if a target sentence contains an overt complementizer, we checked if there is a subordinating conjunction marker between the matrix subject and the CC subject.

The independent variable is word order flexibility, which is measured in our study by looking at the codependencies between subject and object of the same predicate (Levshina 2019). Specifically, we used Shannon entropy (Shannon 1948) to represent variation of word order in the subject-object codependencies. The entropy can be calculated with Equation (2), where X is a binary variable representing two possible word orders (i.e., SO and OS), and $p(x)$ represents the probability of one of the orders in a given language, which can be approximated by looking at the proportion of that order in the corpus. For instance, if in a certain language, the object precedes the subject in 10 % of cases, and the subject precedes the object in 90 % of cases, then its word order entropy is –(0.1 * log2(0.1) + 0.9 * log2(0.9)) = 0.469.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{2}$$

3.3 Results

To test our prediction across 23 languages, we fitted a linear mixed-effects regression model predicting the omission rate by word order entropy, with random intercepts by language family and genus. Information about language families and genera was taken from WALS Online (https://wals.info). As predicted and shown in Figure 5, we find a significant effect of word order entropy ($\widehat{\beta}$ = –0.457, p = 0.010). Cross-linguistically, the

predicted probability of the omission rate of a language tends to be lower if its word order entropy is higher. Importantly, the observed effect remains robust when we controlled for genealogical relations across languages and, therefore, provides strong support for our hypothesis.

4 General discussion

The primary purpose of this study is to investigate cross-linguistic differences in complementizer omission and to test if the differences across languages can be explained in terms of their word order flexibility. Our hypothesis of ambiguity avoidance predicts that language users should show a higher preference for omission if the word order in their languages is more fixed and is thus less likely to cause ambiguities. As shown in our two analyses, the effect is indeed observed.

In previous studies on complementizer omission, only the effects of language-internal factors have been investigated. The present study goes beyond them by expanding the empirical base and exploring cross-linguistic differences. Each of the two analyses conducted in our study has its advantages and provides complementary evidence for our hypothesis. The in-depth analysis is based on large corpora of three

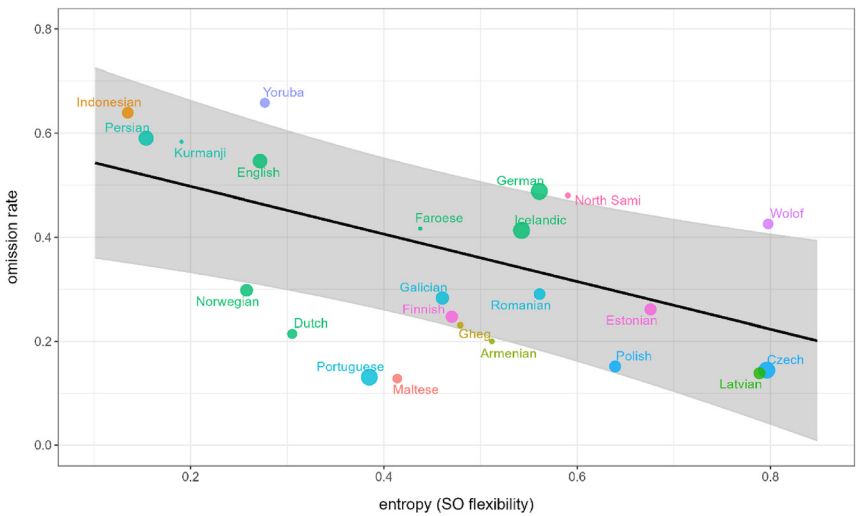

Figure 5: Relationship between word order entropy and omission rate across 23 languages in the UD Treebank. Each point represents a language, with point size denoting the number of target cases and point color denoting language genus. The correlation line and the confidence interval take into consideration genealogical relations between languages.

languages and thus offers accurate estimates of the omission rates of the target verbs from the three languages. In extracting target cases, we relied on our knowledge of these languages, in addition to POS and dependency annotations, to improve our accuracy. For instance, we stipulated that the complementizer in Russian has to be *čto* (rather than other complementizers that cannot be omitted, like *čtoby*). Moreover, the analysis controlled for factors that have been reported to have an effect on complementizer omission, so that we can ensure that the language-level difference we observed was not due to their effects. In the in-depth analysis, we used language itself as a proxy for word order flexibility in the clause and found that the language with the least flexible order (English) has the highest omission rate, and that the language with the most flexible order (Russian) has the lowest omission rate. The results, therefore, constitute initial evidence that complementizer omission is affected by language users' syntactic strategies to avoid ambiguities.

In our broad analysis, we addressed a few limitations of the first analysis by quantitatively estimating the word order flexibility and the complementizer omission rate of 23 languages. As mentioned above, the POS and dependency annotations do not allow us to filter out all the unwanted cases, so the accuracy of estimated omission rates would be lower. However, it has the advantages of including languages from various families and quantitatively estimating their word order flexibility with an entropy measure. As predicted by the ambiguity avoidance hypothesis, we found a significant negative correlation between word order flexibility and complementizer omission rate even after we controlled for genealogical relations between languages. Combining results from the two analyses, we, therefore, find strong support for the hypothesis.

According to Temperley (2003), ambiguity avoidance can be observed from three aspects. First, ambiguity avoidance could play a role in the formation of general syntactic principles. For instance, it was suggested in Bever (1970) that the (mandatory) requirement of a relative pronoun in subject relative clauses is due to the fact that such clauses would be highly ambiguous without an overt relative pronoun. See the contrast in (7) (Temperley 2003):

7. The man who hired me was very tall.
   * The man hired me was very tall.

Second, ambiguity avoidance could influence syntactic strategies. One example of such strategies regarding complementizer omission is that language users show a preference for an overt complementizer when the CC subject is a pronoun (Elsness 1984). Third, ambiguity avoidance could come into play in a highly situation-specific, ad-hoc fashion. At the third level, whether or not a complementizer is needed to avoid ambiguity should be discussed case by case. One makes a decision by taking

into account various (syntactic, semantic, and pragmatic) factors in a specific context.

In our study, the observed difference in omission rates across languages is attributed to the effect of word order flexibility and is in line with the ambiguity avoidance account at the second level (strategic ambiguity avoidance). Although some previous studies seemed to have failed in identifying the expected effects of ambiguity avoidance, our finding that word order flexibility can affect complementizer omission does not contradict with their results. In Elsness (1984), a prediction was made on the basis of the ambiguity avoidance account that the English complementizer *that* would be used more frequently with *you* as the subject of the CC than with other case-distinguishing pronouns. Yet no significant difference was detected. Similarly in Ferreira and Dell (2000), no specific preference for an overt complementizer was found in potentially ambiguous sentences, again casting doubt on the ambiguity avoidance account. However, as also pointed out in Temperley (2003), their studies mainly focused on specific lexical syntactic strategies (*you* vs. *I* and *she*), and the general tendency to omit complementizer in cases of a pronominal subject of the CC reported in Elsness (1984) can already be seen as an ambiguity avoidance mechanism. Further supporting evidence for the ambiguity avoidance account at the general syntactic level was provided in the analysis of the use of relative pronouns in Temperley (2003). In comparison, the effect of word order flexibility detected in our study is at an even higher and cross-linguistic level, which is in line with the claim that it is more likely to observe ambiguity avoidance mechanisms at general strategic levels than at specific and situational levels. The use of ambiguity avoidance strategies is much more feasible at the general syntactic level, especially if "their conditions for their application are defined in a fairly simple way (Temperley 2003: 482)."

Broadly, our findings are compatible with the communicative efficiency hypothesis, according to which languages are efficient in both production and perception to meet the communicative needs of their users (Jaeger and Buz 2018; Levshina and Moran 2021). The availability account was proposed as an account for efficient production. Previous research has argued that the effects of several language-internal factors, including the frequency of matrix verbs (Jaeger 2010), the type of subjects in complement clauses (Cacoullos and Walker 2009), and the coreferentiality between subjects in matrix and complement clauses (Elsness 1984), can be explained with the availability account. In our study, we attributed the effect of word order flexibility to the ambiguity avoidance account, which is related to efficient communication on the perception side. Ambiguity avoidance is regarded as a type of audience design, which refers to "when speakers fashion their utterances so as to cater to the needs of their addressees" (Ferreira 2019: 29). It should be noted that the observed effect in our study does not indicate that users of a certain language will

choose to use an overt complementizer in a specific ambiguous situation. Rather, our results indicate that in order to avoid ambiguities, a general tendency (or strategy) is developed at the language level. Different languages thus show different tendencies. In brief, our results align well with the hypothesis that complementizer omission is not only conditioned by various factors to facilitate the production of language users but also to ensure the understanding of their interlocutors (or readers in the context of writing).

The idea that language-level properties can influence language use has been well documented in previous research (Jaeger and Buz 2018; Levshina and Moran 2021). For example, in Rubio-Fernandez et al. (2021), it is reported that speakers of English, which is a prenominal language (adjectives come before nouns), use more redundant adjectives than Spanish, which is a postnominal language. Kachakeche et al. (2021) quantified the propensity of languages to use adjectives prenominally and found that, across 74 languages, the ones that favor prenominal adjectives indeed exhibit higher rates of adjectival modification. Through miniature artificial language learning experiments, Fedzechkina et al. (2017) and Fedzechkina and Jaeger (2020) showed that learners tend to drop optional case markings in the fixed order language but retain them in the flexible order language. However, in previous studies on linguistic alternations, the direction of alternations has rarely been associated with language-level properties. In future research, it would be interesting to investigate other alternations cross-linguistically and to see if there is any systematic difference that can be explained with language-level characteristics.

5 Conclusions

In the present study, we proposed an account of ambiguity avoidance and made a prediction about complementizer omission at the cross-linguistic level: in order to avoid ambiguities, users of languages with more flexible word order should be more reluctant to omit the complementizer. The prediction was confirmed with an in-depth analysis of three languages (English, German, and Russian) and a broad analysis of 23 languages from 13 genera. To the best of our knowledge, our study is the first to explore cross-linguistic differences in complementizer omission. Our findings support the claim that language users have a strategic (cross-linguistic) tendency to produce optional linguistic forms when it helps to avoid ambiguities.

In conclusion, our study provides replicating and novel supporting evidence that language-processing mechanisms have a major influence on syntactic reduction. The observed effect of word order flexibility adds to the literature on ambiguity avoidance and is compatible with the hypothesis that language users consider the knowledge and processor state of their interlocutors (or readers in the context of

writing) and adjust their language production accordingly in an attempt to successfully transfer the intended information. Ultimately, our findings on complementizer omission align well with the general hypothesis that language production is organized to transfer information efficiently.

Appendix

English

KNOW: know, knows, knew, known, knowing.
BELIEVE: believe, believes, believed, believing.

German

WISSEN: wissen, weiß, weißt, wisst, wusste, wusstest, wussten, wusstet, gewusst.
GLAUBEN: glauben, glaube, glaubst, glaubt, glaubte, glaubtest, glaubten, glaubtet, geglaubt.

Russian

ZNAT′: знать, знаю, знает, знаешь, знают, знаем, знаете, знал, знало, знала, знали.
VERIT′: верить, верю, верит, веришь, верят, верим, верите, верил, верило, верила, верили.

References

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Berdicevskis, Aleksandrs, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama & Christian Bentz. 2018. Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the second workshop on universal dependencies*, 8–17. Brussels, Belgium: Association for Computational Linguistics.

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. In John R. Hayes (ed.), *Cognition and the development of language*, 279–362. New York: Wiley.

Bolinger, Dwight. 1972. *That's that*. The Hague: Mouton.

Bornkessel, Ina, Matthias Schlesewsky & Angela D. Friederici. 2002. Grammar overrides frequency: Evidence from the online processing of flexible word order. *Cognition* 85(2). B21–B30.

Boye, Kasper & Poulsen Mads. 2011. Complementizer deletion in spoken Danish. *Paper presented at the 44th annual meeting of the Societas Linguistica Europaea,* Logroño, September 2011.

Brandt, Silke, Elena Lieven & Michael Tomasello. 2010. Development of word order in German complement-clause constructions: Effects of input frequencies, lexical items, and discourse function. *Language* 86(3). 583–610.

Brown-Schmidt, Sarah & Agnieszka E. Konopka. 2008. Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition* 109(2). 274–280.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 602–623. Oxford: Blackwell.

Cacoullos, Rena Torres & James A. Walker. 2009. On the persistence of grammar in discourse formulas: A variationist study of that. *Linguistics* 47(1). 1–43.

Dor, Daniel. 2005. Toward a semantic account of that-deletion in English. *Linguistics* 43(2). 345–382.

Elsness, Johan. 1984. That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65(6). 519–533.

Fedzechkina, Maryia & T. Florian Jaeger. 2020. Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition* 196. 104115.

Fedzechkina, Maryia, Elissa L. Newport & T. Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science* 41(2). 416–446.

Fenk-Oczlon, Gertraud. 2001. Familiarity, information flow, and linguistic form. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 431–448. Amsterdam: John Benjamins.

Ferreira, Victor S. 2019. A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology* 70(1). 29–51.

Ferreira, Victor S. & Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40(4). 296–340.

Frazier, Lyn. 1985. Syntactic complexity. In David R. Dowty, Lauri Karttunen & Arnold M. Zwicky (eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives*, 129–189. Cambridge: Cambridge University Press.

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(5). 389–407.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the international conference on language resources and evaluation*, 31–43. Istanbul, Turkey: European Language Resources Association.

Gries, Stefan Th. 2013. *Statistics for linguistics with R: A practical introduction*. Berlin, New York: De Gruyter Mouton.

Gries, Stefan Th. 2017. Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics* 31(1). 8–29.

Gries, Stefan Th. 2021. (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning* 71(3). 757–798.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1). 97–129.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.

Jaeger, T. Florian & Esteban Buz. 2018. Signal reduction and linguistic encoding. In Eva M. Fernández & Helen Smith Cairns (eds.), *The handbook of psycholinguistics*, 38–81. Hoboken: John Wiley & Sons.

Kachakeche, Zeinab, Richard Futrell & Gregory Scontras. 2021. Word order affects the frequency of adjective use across languages. In *Proceedings of the annual meeting of the cognitive science society*, 3006–3012. Vienna, Austria: Cognitive Science Society.

Kurumada, Chigusa & T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.

Levshina, Natalia & Steven Moran. 2021. Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard* 7(s3). 20200081.

Liang, Yiming, Pascal Amsili & Heather Burnett. 2021. New ways of analyzing complementizer drop in Montréal French: Exploration of cognitive factors. *Language Variation and Change* 333. 359–385.

Matthews, Danielle, Elena Lieven, Anna Theakston & Michael Tomasello. 2005. The role of frequency in the acquisition of English word order. *Cognitive Development* 20(1). 121–136.

Mykhaylyk, Roksolana, Yulia Rodina & Merete Anderssen. 2013. Ditransitive constructions in Russian and Ukrainian: Effect of givenness on word order. *Lingua* 137. 271–289.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal dependencies v2: An ever-growing multilingual treebank collection. In *Proceedings of the international conference on language resources and evaluation*, 4034–4043. Marseille, France: European Language Resources Association.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the annual meeting of the association for computational linguistics: System demonstrations*, 101–108. Association for Computational Linguistics.

R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roland, Douglas, Jeffrey L. Elman & Victor S. Ferreira. 2006. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* 98(3). 245–272.

Rubio-Fernandez, Paula, Francis Mollica & Julian Jara-Ettinger. 2021. Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General* 150(3). 583–594.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Siewierska, Anna & Ludmila Uhlirova. 1998. An overview of word order in Slavic languages. In Anna Siewierska (ed.), *Constituent order in the languages of Europe*, 105–150. Berlin, New York: De Gruyter Mouton.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Temperley, David. 2003. Ambiguity avoidance in English relative clauses. *Language* 79(3). 464–484.

Thompson, Sandra A. 2002. "Object complements" and conversation towards a realistic account. *Studies in Language* 26(1). 125–163.

Thompson, Sandra A. & Anthony Mulac. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, 313–329. Amsterdam: John Benjamins.

Wulff, Stefanie, Stefan Th. Gries & Nicholas Lester. 2018. Optional that in complementation by German and Spanish learners. In Andrea Tyler, Lihong Huang & Hana Jan (eds.), *What is applied cognitive linguistics*, 99–120. Berlin, Boston: De Gruyter Mouton.

Yoon, Jiyoung. 2015. The grammaticalization of the Spanish complement-taking verb without a complementizer. *Journal of Social Sciences* 11(3). 338–351.

Zipf, George K. 1949. *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.

Zou, Yue & Hao Lin. 2024. The optionality of complementizer čto in Russian – a multi-factorial analysis. In *Proceedings of the annual meeting of the cognitive science society*, 1902–1908. Rotterdam, The Netherlands: Cognitive Science Society.