The optionality of complementizer čto in Russian — a multifactorial analysis

Yue Zou (yuezou@shisu.edu.cn)
Institute of Linguistics, Shanghai International Studies University
Shanghai, 201600 China

Hao Lin (linhao@shisu.edu.cn)
Institute of Linguistics, Shanghai International Studies University
Shanghai, 201600 China

Abstract

The present study focuses on a type of seemingly arbitrary alternation in modern Russian. Specifically, we investigate the phenomenon of Complementizer Omission, i.e. the alternation between the presence and absence of complementizer with regard to the factors that potentially exert an influence on the alternation in Russian. The choice of alternating pairs is statistically modeled with mixed-effects logistic regression. We find that the complementizer is more likely to be absent in Russian when the matrix subject is a first-or-second-person pronoun, the matrix predicate has a high frequency and the onset of the complement clause is non-ambiguous and non-informational. The findings align well with the Grammaticalization theory, according to which the distribution of complementizer is partially driven by certain types of combinations of matrix subjects and verbs that have become grammaticalized as epistemic markers. Moreover, we argue that the results provide weak support for ambiguity avoidance at the general syntactic level and that the Uniform Information Density account more fully explains the alternation than the Availability account. As in Jaeger and Norcliffe (2009), we propose that more cross-linguistic research should be done on syntactic alternations as "even similar constructions may be processed differently in different languages".

Keywords: Complementizer Omission; Russian; Multifactorial Analysis; Corpus Study

Introduction

Syntactic alternations have captured the attention and imagination of researchers from different sub-fields in linguistics and according to Gries (2017), have become "one of the most thoroughly researched kinds of topics during the past fifty years".

In the present study, we investigate a particular type of syntactic alternation in Russian, namely Complementizer Omission (C-omission), which has already been observed and analyzed in several other languages (see Bolinger, 1972; Ferreira and Dell, 2000; Finegan and Biber, 2001; Roland et al., 2005; Cacoullos and Walker, 2009; Jaeger, 2010 for English; Boye and Poulsen, 2011 for Danish; Liang et al., 2021 for Montréal French; Yoon, 2015 for Spanish; Poletto, 1995 for Italian). An example of C-omission in Russian is provided in (1):

(1) Ja znaju, Ø on rabotaet na zavode.
I know Ø he works at a factory.

C-omission has received wide attention in studies within different frameworks and from different approaches, which have uncovered a number of factors conditioning the variation (Cacoullos and Walker, 2009). However, unlike in English, the optional realization of complementizer in Russian has barely been studied. As pointed out in Morgunova (2021), the conditions under which the complementizer čto might be phonologically null are not clear and are in need of a careful examination.

The focus of our study is the variable use of complementizer čto in Russian to introduce a subordinate clause. To the best of our knowledge, no corpus-based study of this linguistic phenomenon has been conducted. In the study, we will carry out a set of analyses to test three accounts of language processing that attempt to explain the phenomenon while controlling for other potential effects of lexical idiosyncrasy and Grammaticalization. The relative strength and contribution of each predictor is examined in a multifactorial analysis with mixed-effects modeling.

Predictions for C-omission in Russian

Here we briefly summarize the three accounts that have been proposed in previous research: the Availability account, the Ambiguity Avoidance account and the Uniform Information Density account. Variables predicted to have an effect on C-omission by each account are presented in respective subsections.

Availability account

From the perspective of language production, Ferreira and Dell (2000) introduced 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)". The principle makes a straightforward prediction for CO: if the first word of the complement clause (CC) is selected quickly, then the complementizer should be omitted to accommodate immediate mention of that quickly selected word (Ferreira and Dell, 2000).

The variable most relevant to the account is CO-REFERENTIALITY: if the CC subject is in the pronominal form and/or is co-referential with the matrix clause (MC) subject, then it indicates that its corresponding referent has already been mentioned in the previous conversation and can be more easily accessed by the speaker. Therefore, an overt complementizer is more likely to be omitted.

An effect of FREQUENCY of matrix verb is also to be expected if we consider the "spill-over" effect mentioned in

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Jescheniak and Levelt (1994) and Baayen et al. (2006): increased processing load associated with the production of less frequent word forms may spill over from the matrix verb to the CC onset. In that case, the Availability account predicts a higher probability of the absence of complementizer following more frequent matrix verbs (Jaeger, 2010). At the same time, the effect of FREQUENCY is also predicted by the Grammaticalization theory, according to which highly frequent matrix verbs usually occur without an overt complementizer.

Ambiguity Avoidance account

Frazier (1985) proposed the impermissible ambiguity constraint, according to which constructions that would lead to ambiguities or misanalyses tend to be prohibited. Specifically for CO, a strong tendency was found in Elsness (1984) that the complementizer is more likely to be absent in English cases where the subject of the CC is a pronoun.

In our study, by encoding AMBIGUITY at the CC onset, we expect to see something similar in Russian: speakers should have a tendency to omit an overt complementizer if the first word of the CC is of a certain lexical category that is unlikely to create ambiguity. Moreover, an effect of SUBCATEGORIZATION PREFERENCE of the matrix verb is also relevant: if the matrix verb can be used in ways except co-occurring with a CC, then a speaker should potentially be less likely to produce the sentence in its reduced form to avoid temporary ambiguity.

Uniform Information Density (UID) account

UID predicts that the probability of omission of an optional linguistic element is inversely correlated with the information that it carries in context (Kravtchenko 2014). Since its formulation in Jaeger (2010), the effect of UID has received much attention and has been studied in various languages (see Horch and Reich, 2016 for article omission in German; Jain et al., 2018 for word order alternation in Hindi; Temperley, 2019 for supporting evidence in music, which is a universal language).

Mentioning čto at the CC onset distributes the same amount of information over one more word, thereby lowering information density (Jaeger, 2010: 27). In our study, the amount of information at the CC onset is encoded with two distinct variables: SEMANTIC CONTENTFULNESS and SURPRISAL: a speaker should be more likely to omit the complementizer if the CC onset is non-contentful and unsurprising.

Methods

As in Gries (2003), the perspective taken in the present study is "rigorously corpus-based in order to find out what is actually done by native speakers". The analyses are carried out using the Corpus of Spoken Russian, which is part of the Russian National Corpus (RNC; available at https://ruscorpora.ru/).

Defining the variable context

In order to find the most frequent verbs that can co-occur with a CC, we first searched for instances containing a string of a comma, a space and a complementizer čto, from an offline disambiguated version of the RNC. The corpus contains texts in modern Russian and is of about one million words in size, where fiction, academic and journalistic texts, transcripts of oral speech and blogs are represented in roughly equal proportion.

A total of 1428 instances were extracted from the offline corpus. Secondly, 584 instances where the word čto is not used as a complementizer were manually filtered out. The top 10 most frequent matrix predicates that can take a CC are chosen as the object of our following more detailed annotations and analyses.

Table 1: The top 10 most frequent matrix predicates that can co-occur with a subordinate CC.

Verb	Count
skazat' (say)	85
znat' (know)	63
kazat'sja (seem)	43
govorit' (say)	41
dumat' (think)	37
sčitat' (think)	28
ponimat' (understand)	26
čuvstvovat' (feel)	26
ponjat' (understand)	24
uznat' (get to know)	19

For each of the 10 chosen verbs, we randomly selected 300 instances from the Corpus of Spoken Russian and manually processed them to exclude cases where the verb is not followed by a subordinate CC. A total of 731 instances from 150 conversations are included for the final annotation of the above-mentioned potentially relevant variables.

Clarifying the coding process

For each instance, we coded whether the complementizer čto is present or absent (the dependent variable) and a number of potential influencing factors (independent variables), each of which tests a hypothesis reported in the literature on CO. Apart from the above-mentioned variables related to accounts of language processing, two other variables are also considered to control for potential effects of lexical idiosyncrasy (ΔP) and Grammaticalization (TYPE of matrix subject). The coding process for each variable is briefly described below.

Type of matrix subject Type of matrix subject is coded as a binary variable with two levels, namely (1) first-or-second-person pronoun and (2) other.

Frequency of matrix verb Frequency of matrix verb is coded as a continuous variable by counting how many times each verb occurs in the offline version of the RNC.

Co-referentiality Co-referentiality is coded as a binary variable with two levels, namely (1) co-referential and (2) non-co-referential, depending on whether the referent mentioned (usually by the subject) in the MC is also mentioned at the CC onset.

Ambiguity at the CC onset Ambiguity at the CC onset is coded on the basis of the categorization of the first word of the CC by determining whether a particular word class at the onset can potentially create a garden path. Instances where the first word belongs to prepositions, pronouns and adverbs are coded as ambiguous, while instances where it belongs to nouns, verbs, adjectives, conjunctions, numbers and particles are coded as non-ambiguous.

Subcategorization preference of matrix verb Subcategorization bias of each verb is measured by estimating how often it co-occurs with a subordinate CC (with or without an overt complementizer). For instance, ponjat' in Russian co-occurs mainly with a simple NP object or is used as an intransitive verb. A subordinate CC follows the verb in only 56 of the 300 instances we checked. Therefore, for ponjat', the value for the variable is 0.186 (56 / 300).

Lexical idiosyncrasy It has long been suggested that lexical items often have preferences for certain constructions. Lexical idiosyncrasy of each matrix verb is quantified with ΔP, which is a directional association measure discussed in Ellis (2006) in the field of associative learning. The value of ΔP for each verb is calculated with equation (1), where the cue refers to the presence of a particular matrix verb and the outcome refers to the omission of complementizer.

$$\Delta P = p(outcome \mid cue) - p(outcome \mid no\ cue) \quad (1)$$

Surprisal As in Wulff et al. (2018) and Gires (2021), conditional surprisal of the CC onset was measured by considering "the last word in the MC prior to the clause juncture, regardless of whether or not the complementizer is present". The operationalization of conditional surprisal is based on equation (2).

$$S_c(x|y) = -log_2 p(x|y) \quad (2)$$

Additionally, the multifactorial analysis includes a random intercept for each conversation, which can be thought of as an individual adjustment to the personal preference of each pair of interlocutors.

Results and Discussion

As stated above, a sample of 731 instances of verbs with subordinate CCs as their direct objects was extracted from The Corpus of Spoken Russian. A complementizer čto is present in 474 and absent in 257 instances. Overall, Russian speakers are more "conservative" than English speakers in the sense that the rate of C-omission is much lower in Russian (35.2%) than in English (cf. 82.5% in Jaeger, 2010; 67.9% in Gries, 2021).

Table 2: The semantic class of each matrix verb and their value of ΔP. The classification is based on the definitions given in Noonan (1985).

Verb	Semantic class	ΔP
skazat' (say)	utterance	0.024
znat' (know)	knowledge	0.219
kazat'sja (seem)	propositional attitude	0.308
govorit' (say)	utterance	0.139
dumat' (think)	propositional attitude	0.120
sčitat' (think)	propositional attitude	-0.238
ponimat' (understand)	knowledge	-0.296
čuvstvovat' (feel)	immediate perception	-0.148
ponjat' (understand)	knowledge	-0.280
uznat' (get to know)	knowledge	-0.297

The influence of verb semantics

Thompson and Mulac (1991) suggested that, in English, matrix verbs of the same semantic class may exhibit similar preference in terms of C-omission. In order to find out if verbs of certain semantic classes have a consistent preference for the presence of absence of complementizer in Russian, we semantically classified the 10 matrix verbs into four classes, as shown in Table 2.

Of the three classes that contain more than one verb, no consistent preference is observed, which is in agreement with the result in Cacoullos and Walker (2009), where verb semantics was not found to be a significant predictor in either direction. For instance, idiosyncratic preference of the three verbs of thinking varies greatly: kazat'sja has a relatively high preference for complementizer omission; scitat' has a strong tendency of retaining the complementizer; and dumat' shows no specific preference in either direction. Therefore, we propose that in Russian, the preference for C-omission may be purely lexically specific, although a future study that includes more matrix verbs is necessary to make a more compelling argument. In the following multifactorial analysis, ΔP is treated as a control variable.

Multifactorial analysis

A mixed-effects regression analysis is implemented with the `lme4` package in R. The four continuous variables (frequency, subcategorization preference, lexical idiosyncrasy and surprisal) are z-standardized before incorporated into the model. No issue of multicolinearity is detected in the initial model as the VIF values for all the predictors are below 6. Model selection is performed following the two-step strategy outlined by Zuur et al. (2009). For the selection criterion of figuring out the right fixed-effects structure, we used likelihood ratio tests (LRTs) with a significance threshold set to 0.05. $R^2_{marginal}$, $R^2_{conditional}$, and C-score of the final model are 0.329, 0.378 and 0.824, respectively. Classification accuracy of the final model is 0.763, which is significantly higher than the baseline ($p_{binom} < 0.001$). A summary of results is

provided in Table 3. Variables related to each of the four accounts (three processing accounts plus the Grammaticalization account) are discussed separately.

Grammaticalization account Our results provide strong support for the Grammaticalization account. As shown in Table 3, TYPE of matrix subject has the largest effect size in the model: the predicted probability of the presence of an overt complementizer is significantly lower if the matrix subject in an utterance is a first-or-second-person pronoun (β = 1.330, z = 4.982, $p < 0.001$). In analyzing C-omission in English, Thompson and Mulac (1991: 242) found that I and you disfavor the presence of complementizer more than other matrix subjects, which they explain by the higher frequency of I and you in discourse and their capacity to express epistemicity or subjectivity. The observed effect provides support for the Grammaticalization account in Thompson and Mulac (1991) with novel evidence from a different language (i.e. Russian). As a matter of fact, TYPE of matrix subject is the second strongest (after lexical idiosyncrasy[1]) predictor in terms of its contribution to the model's likelihood ($\chi^2(1) = 27.665$, $p < 0.001$). Notably, the improvement in model quality more than matches that of all the factors motivated by processing accounts ($\chi^2(6) = 20.882$, $p = 0.002$).

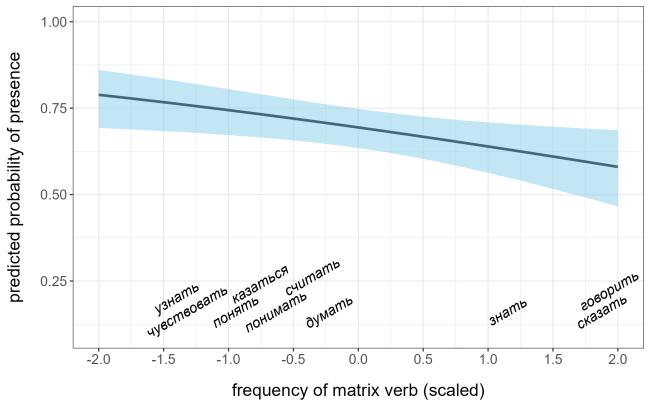

Figure 1: Effect plot for FREQUENCY of matrix verb. The variable is z-standardized. The shaded area denotes 95% CIs.

Figure 1 displays the effect plot for FREQUENCY of matrix verb: the predicted probability of the presence of an overt complementizer decreases as the frequency of the matrix verb increases (β = -0.248, z = -2.431, p = 0.015), which is also in line with the Grammaticalization account. The highly significant positive correlation between frequency of the matrix verb and complementizer omission replicates earlier results (Elsness, 1984; Garnsey et al., 1997; Roland et al., 2005, Jaeger, 2010). According to Jaeger (2010: 41), the effect is compatible with the Availability account under the assumption that "production resources are limited so that high processing load can 'spill over' into the planning of upcoming material". To be more precise, less frequent matrix verbs are assumed to be less available and, correspondingly, take more effort to produce. An overt complementizer is thus more likely to be present to allow for additional time to deal with the high processing load.

Ambiguity Avoidance account Strategic ambiguity avoidance has been argued to have an effect on C-omission in English in two ways: lexically and syntactically (Temperley, 2003). Our analysis in Russian provides weak support for the latter one.

Subcategorization preference of matrix verb, which potentially reflects ambiguity avoidance at the lexical level, did not survive the model selection process[2]. More specifically, verbs which have a high subcategorization preference for co-occurring with a subordinate CC may either prefer or disprefer C-omission. For instance, kazat'sja and sčitat' have the highest values for subcategorization preference, but show distinct patterns in terms of C-omission. As a side note, we did find some correlation between verb semantics and subcategorization preference: verbs of propositional attitude strongly prefer co-occurring with a subordinate CC, while verbs of knowledge show the opposite tendency.

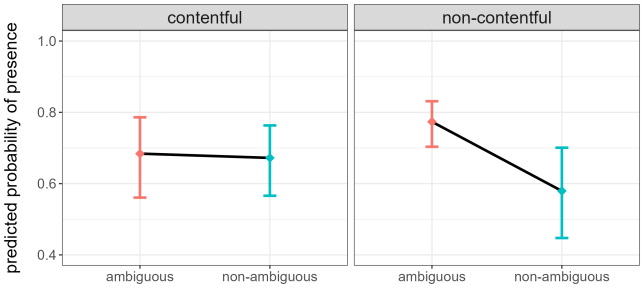

Figure 2: Effect plot for the interaction between semantic contentfulness and ambiguity at the CC onset. Error bars denote 95% CIs.

An interaction between semantic contentfulness and ambiguity at the CC onset is detected in our analysis. As shown in Figure 2, the factor encoding ambiguity at the CC onset only has an effect on C-omission when the onset is non-contentful: the predicted probability of the presence of an overt complementizer is significantly lower if the onset is non-ambiguous (β = -0.909, z = -3.039, p = 0.002). We interpret the result as indicating that Russian speakers use their language in a conservative way in that a complementizer tends to be omitted only when the onset is both non-ambiguous and semantically non-contentful.

UID account Apart from semantic contentfulness that is discussed above, UID also predicts an effect of surprisal. Figure 3 displays the effect plot for SURPRISAL: the predicted

[1]The fact that lexical idiosyncrasy emerges as the strongest predictor is unsurprising as we are simply modeling C-omission in Russian with the omission preference of each verb.

[2]A Spearman's correlation test indicates that the two variables [subcategorization preference] and [lexical idiosyncrasy] are not significantly correlated (r = 0.379, p = 0.281).

Table 3: Result summary: coefficient estimates, standard errors, z scores and p values for predictors in the final model.

	Estimate	SE	z value	p value
(Intercept)	0.422	0.237	1.782	0.075
Lexical idiosyncrasy (ΔP)	-0.927	0.109	-8.482	< 0.001
Type of matrix subject $_{\text{other}}$	1.330	0.267	4.982	< 0.001
Frequency of matrix verb	-0.248	0.102	-2.431	0.015
Semantic contentfulness $_{\text{non-contentful}}$	-0.398	0.334	-1.193	0.233
Ambiguity $_{\text{ambiguous}}$	0.055	0.306	0.179	0.858
Surprisal	0.285	0.128	2.226	0.026
Co-referentiality $_{\text{co-referential}}$	-0.685	0.338	-2.027	0.043
Semantic contentfulness $_{\text{non-contentful}}$: Ambiguity $_{\text{ambiguous}}$	0.854	0.421	2.028	0.043

probability of the presence of an overt complementizer increases as the degree of surprisal increases (β = 0.285, z = 2.226, p = 0.026). The effect has also been observed in Wulff et al. (2018) and Gries (2021) for C-omission in English and corroborates the UID account.

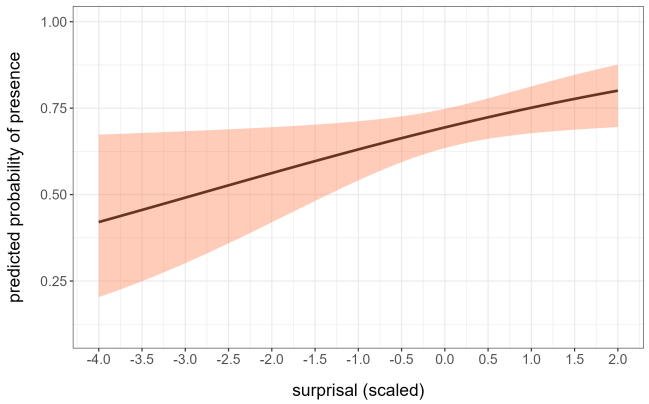

Figure 3: Effect plot for SURPRISAL. The variable is z-standardized. The shaded area denotes 95% CIs.

Availability account The factor that is usually attributed to the effect of lexical availability is CO-REFERENTIALITY (Jaeger, 2010). The model summary indicates that the predicted probability of the presence of an overt complementizer is significantly lower if the referent of the matrix subject is also mentioned at the CC onset (β = -0.685, z = -2.027, p = 0.043). The effect direction of the variable is in line with the prediction of the Availability account, although the effect does not quite achieve significance at the 0.01 level, thereby providing only weak support for the account in Russian.

However, as pointed out by one reviewer, the effect can also potentially be explained by the UID account: a co-referential CC subject is generally less informational than a non-co-referential one. Therefore, in order to disentangle the availability effect from the surprisal-based predictions of UID, we consider a syntactic feature which may play a role in availability but is not directly related to the surprisal at the CC onset: the complexity of the noun phrase at the CC onset (Clark et al., 2022). Availability-based production should predict that syntactically more complex NPs are less available, and would therefore be more likely to retain the complementizer. The UID account, meanwhile, should not be sensitive to the factor when controlling for the surprisal of first word.

To test whether Russian speakers are sensitive to the availability effect in C-omission, we extracted from our sample cases where the first word at the CC onset is a noun, adjective or number. NPs that are modified by adjectives or numbers are coded as having a complex structure (e.g. the main issue, two albums). We fitted a mixed-effects model with the subset of our sample and found that the main effect of surprisal remained significant, whereas that of complexity failed to reach significance (β = -0.133, z = -0.305, p = 0.760). The result indicates that compared with the Availability account, UID more fully explains the observed alternation.

Interestingly, in Clark et al. (2022), an opposite conclusion was drawn in their analysis of the comparative alternation in Russian (the Availability account better explains the phenomenon than UID). It could be argued that the availability effect is manifested in the ordering of constituents in the CC rather than in the optional realization of complementizer at the CC onset so that the more available constituent is produced earlier. Still, whether and how availability has impact on C-omission in Russian merits future investigation.

Further analysis of the random effect

As rightly pointed out in Gries and Wulff (2021), the lack of attention to results of random effects is lamentable: "researchers use them to get more robust or generalizable results, but do nothing else with them, neither visualization nor further exploration nor correlating them with predictors not included in the model". In this section, a further exploration of the random adjustments for File is carried out.

The effect of random intercept adjustments for conversations is positively correlated with the presence of complementizer: a positive intercept adjustment increases the predicted probability of the presence of complementizer, whereas a negative slope adjustment decreases it.

The factor we chose to explore from the random effect is

the potential influence of formality, which has been reported to have an effect on C-omission in English (Biber, 1999). To explore the additional effect, we classified the conversations according to their formality based on a subjective judgement of the content of the conversation revealed by the corresponding file name. For instance, the file name "What kind of opposition Russia needs and whether it needs it at all from Program" clearly indicates a formal conversation, while "A conversation between two female students about movies and classmates" is typical of an informal conversation. Conversations are labeled uncertain if their file names do not imply the content.

The average intercept adjustment for conversations classified as formal is 0.115 (SD = 0.301), while the average adjustment for conversations classified as informal is -0.273 (SD = 0.354). A one-tailed t-test for independent samples indicated a statistically significant difference (t_{Welch} = 2.639, df = 20.121, $p_{1\text{-}tailed}$ = 0.008).

The result is consistent with the findings in Biber (1999): conversations, which typically have the characteristics of being produced on-line, having involved, interpersonal purposes and being casual and informal in tone, prefer the absence of complementizer.

Exploration of the middle ground

In Wulff et al. (2014: 272), it is mentioned that the variable presence of complementizer in English "appears to be a matter of gradient probabilistic preference rather than discrete grammaticality: omitting or producing that never renders an utterance ungrammatical, but depending on the specific context, omitting or producing that may render the utterance more or less idiomatic". In this section, we consider the possibility that in certain contexts in Russian, the presence and absence of complementizer are just about equally acceptable.

As pointed out in Gries and Deshors (2020), one potential shortcoming of performing a multifactorial analysis is that "the regression model might too eagerly label a certain choice as misclassified". For instance, the probability of the presence of complementizer in (2) is predicted by the final model to be 48.8%. If we take 0.5 as the threshold of classification, then it would be considered as a misclassified case. However, the handling of similar situations with the above-mentioned method is counter-intuitive as it ignores the middle ground, where either constructional choice is acceptable to Russian speakers.

(2) Mne kažetsja, čto huže čem etot god ne budet.
It seems to me that it won't be worse than this year.

Following Gries and Deshors (2020) and Gries (2021), we improve on our multifactorial analysis by considering the middle ground of C-omission in Russian. On the basis of the final model, we generated a 95% confidence interval of the predicted probability for each instance with the `bootMer` function. For each instance, if the confidence interval includes 0.5, then the categorical prediction is changed to "either". Confusion matrix of the three-way classification is shown in Table 4.

Table 4: Confusion matrix of the three-way classification.

	absent	present
prediction: absent	36	14
prediction: either	176	164
prediction: present	45	296

The new way of classification indicates that 340 (176 + 164) instances are labeled as "either". If we consider the "either" cases as correctly predicted, then the accuracy becomes much higher (91.9%). Therefore, the result, which is now more consistent with the intuition and linguistic knowledge of native speakers, indicates that a large degree of freedom exists in choosing to retain or omit the complementizer.

Conclusion

We believe that much insight can be gained from detailed investigations of syntactic alternations in Russian and other languages as "even similar constructions may be processed differently in different languages (Jaeger and Norcliffe, 2009: 13)". By expanding the empirical base with the approach pursued in our study, we can gradually uncover potentially universal patterns of usage and finally identify to what extent linguistic behavior can be explained by mechanisms of language processing.

On the one hand, our study successfully replicated some results in previous research. We refuted the absolute optionality of C-omission in Russian and showed that, as in English, there are clear conditions underlying a Russian speaker's decision to retain or omit the complementizer in ordinary conversation. Specifically, we found that morphosyntactic (type of matrix subject, co-referentiality, lexical category at the CC onset), information-theoretic (frequency, surprisal) and pragmatic (formality of conversation) factors have a significant effect on C-omission in Russian.

On the other hand, the results go beyond the confirmation of previous research in several cases. First, compared with previous findings in English, our study provides relatively weak support for accounts of language processing as none of the effects of variables motivated by the accounts reached significance at the 0.01 level. Second, in our post-hoc analysis attempting to disentangle the availability effect from the surprisal-based predictions of UID, the predictor related to availability was not found to be significant. We therefore argue that UID more fully explains C-omission in Russian than the Availability account and that the availability effect could potentially be observed in the ordering or constituents in the CC. Third, by exploring the middle ground of the alternation, we propose that for Russian speakers, a large degree of freedom exists in choosing to retain or omit the complementizer.

Acknowledgments

The research was sponsored by the Academic Mentorship Project (Project No. 2023DSYL010). We are grateful to the four anonymous reviewers for their constructive comments that greatly improved the quality of the paper.

References

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of mono-syllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313.

Biber, D. (1999). A register perspective on grammar and discourse: variability in the form and use of English complement clauses. *Discourse studies*, *1*(2), 131–150.

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

Boye, K., & Poulsen, M. (2011). Complementizer deletion in spoken Danish. In *Indlægved societas linguistica europaea - 44th annual meeting.*

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

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics*, *27*(1), 1–24.

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

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

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation.* Cambridge: Cambridge University Press.

Frazier, L. (1985). Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, 129–189.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of memory and language*, *37*(1), 58–93.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. A&C Black.

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

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

Gries, S. T., & Deshors, S. C. (2020). There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, *44*(1), 69–96.

Hansen, B. (2010). Constructional Aspects of the Rise of epistemic sentence adverbs in Russian. *Wiener Slawistischer Almanach*, *74*, 75–86.

Hansen, B., Letuchiy, A., & Błaszczyk, I. (2001). Complementizers in Slavonic (Russian, Polish, and Bulgarian). In K. Boye & P. Kehayov (Eds.), *Complementizer semantics in European languages.* Berlin: Mouton de Gruyter.

Horch, E., & Reich, I. (2016). On "article omission" in German and the "uniform information density hypothesis". In *Proceedings of the 13th Conference on Natural Language Processing.*

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

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Linguistics Compass*, *3*(4), 866–887.

Jain, A., Singh, V., Ranjan, S., Rajkumar, R., & Agarwal, S. (2018). Uniform Information Density effects on syntactic choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing.*

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of experimental psychology: learning, Memory, and cognition*, *20*(4), 824–843.

Kravtchenko, E. (2014). Predictability and syntactic production: Evidence from subject omission in Russian. In *Proceedings of the annual meeting of the Cognitive Science Society.*

Liang, Y., Amsili, P., & Burnett, H. (2021). New ways of analyzing complementizer drop in Montréal French: Exploration of cognitive factors. *Language Variation and Change*, *33*(3), 359–385.

Morgunova, E. (2021). Complementizer-trace effects in Russian. In *Proceedings of ConSOLE XXIX.*

Poletto, C. (1995). Complementizer deletion and verb movement in Italian. *Working Papers in Linguistics*, *5*(2), 1–15.

Roland, D., Elmanand, J. L., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition*, *98*(3), 245–272.

Temperley, D. (2019). Uniform information density in music. *Music Theory Online*, *25*(2).

Thompson, S. A., & Mulac, A. (1991). The discourse conditions for the use of the complementizer that in conversational English. *Journal of pragmatics*, *15*(3), 237–251.

Wulff, S., Gries, S. T., & Lester, N. (2018). Optional that in complementation by German and Spanish learners. In *What is applied cognitive linguistics?* De Gruyter Mouton.

Wulff, S., Lester, N., & Martinez-Garcia, M. T. (2014). hat-variation in German and Spanish L2 English. *Language and Cognition*, *6*(2), 271–299.

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