Evidence for Availability Effects on Speaker Choice in the Russian Comparative Alternation

Thomas Hikaru Clark[1] **(thclark@mit.edu)**

Ethan Gotlieb Wilcox[2] **(wilcoxeg@g.harvard.edu)**

Edward Gibson[1] **(egibson@mit.edu)**

Roger Levy[1] **(rplevy@mit.edu)**

[1] Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139

[2]Department of Linguistics, Harvard University
Cambridge, MA 02138

Abstract

When a language offers multiple options for expressing the same meaning, what principles govern a speaker's choice? Two well-known principles proposed for explaining wide-ranging speaker preference are Uniform Information Density and Availability-Based Production. Here we test the predictions of these theories in a previously uninvestigated case of speaker choice. Russian has two ways of expressing the comparative: an EXPLICIT option (*Ona bystree chem ja*/She fast-COMP than me-NOM) and a GENITIVE option (*Ona bystree menya/She fast*-COMP *me*-GEN). We lay out several potential predictions of each theory for speaker choice in the Russian comparative construction, including effects of post-comparative word predictability, phrase length, syntactic complexity, and semantic association between the comparative adjective and subsequent noun. In a corpus study, we find that the explicit construction is used preferentially when the post-comparative noun phrase is longer, has a relative clause, and is less semantically associated with the comparative adjective. A follow-up production experiment using visual scene stimuli to elicit comparative sentences replicates the corpus finding that Russian native speakers prefer the explicit form when post-comparative phrases are longer. These findings offer no clear support for the predictions of Uniform Information Density, but are broadly supportive of Availability-Based Production, with the explicit option serving as an unreduced form that eases speakers' planning of complex or low-availability utterances. Code for this study is available at `https://github.mit.edu/thclark/russian_uid`

Keywords: Uniform Information Density, Availability-Based Production, Corpus Study

Introduction

Unlike English, Russian has two ways of expressing a comparison of the form "A is ADJ-er than B". The first way, which we call the EXPLICIT construction, involves the insertion of an explicit comparative word чем ("than"), followed by the nominative case of the post-comparative noun phrase (i.e. 'B' in the earlier example). The second way, which we call the GENITIVE construction, uses no explicit comparator but requires a genitive post-comparative NP (see Figure 1). For comparatives of this form, these two constructions are interchangeable, making this alternation an interesting testbed for theories of speaker choice between two syntactic alternatives for a given meaning. In this work, we perform a corpus analysis of this alternation to test hypotheses made by two accounts of human language processing: a) Uniform Information Density (Jaeger, 2006; R. Levy & Jaeger, 2007), and b) availability-based production (Ferreira & Dell, 2000a). We find that the latter theory more fully explains the observed data for this alternation. We then conduct a production study with native Russian speakers to lend further support to the corpus findings.

Predictions for Russian Comparatives

Uniform Information Density

Uniform Information Density (UID) hypothesizes that speaker choices are influenced by a communicative pressure to distribute information uniformly across an utterance, avoiding spikes in information that exceed a theoretical channel capacity (R. Levy & Jaeger, 2007; Jaeger, 2010; Aylett & Turk, 2004; Genzel & Charniak, 2002; Meister et al., 2021). For example, English speakers are more likely to omit "that" in a relative clause when the upcoming content is more predictable in context (R. Levy & Jaeger, 2007). UID has also been tested in other languages and with other linguistic alternations, such as syntactic choices in Hindi (Jain, Singh, Ranjan, Rajkumar, & Agarwal, 2018) and subject omission in Russian (Kravtchenko, 2014).

Information density can be operationalized in terms of the information-theoretic quantity of surprisal, or Shannon information content (Shannon & Weaver, 1949; R. Levy & Jaeger, 2007). Surprisal is defined as the negative log probability of a word in context: $s(w) = -\log p(w \mid \mathbf{w}_{<t})$. According to UID, when a speaker has two alternatives for expressing a given meaning, they should prefer the option that leads to a more uniform distribution of information across the utterance. In the case of the Russian comparative construction, we posit that UID would predict a preference for the explicit construction when the post-comparative material is high-surprisal. The explicit construction offloads some information onto the explicit comparator чем, which conveys the fact that the upcoming phrase completes a comparative construction, while the following words convey their lexical meanings. This separation of information leads to a more uniform information density across the utterance and prevents an

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Она	быстрее,	чем	известная	китайская	спортсменка
Ona	bystr-ee,	chem	izvestn-aya	kitajsk-aya	sportsmenk-a
She	fast-COMP,	than	famous-FEM.NOM.SG	Chinese-FEM.NOM.SG	athlete-FEM.NOM.SG

Она	быстрее	известной	китайской	спортсменки
Ona	bystr-ee,	izvestn-oj	kitajsk-oj	sportsmenk-i
She	fast-COMP	famous-FEM.GEN.SG	Chinese-FEM.GEN.SG	athlete-FEM.GEN.SG

‘She is faster than the famous Chinese athlete.’

Comparative adjective Explicit comparator Post-comparative NP Head noun

Figure 1: Example Russian explicit and genitive comparative sentences with labeled sentence parts.

information spike on the post-comparative word. Therefore, we should expect to see the explicit construction more often when the upcoming content is already information-dense.

Availability-Based Production

According to availability-based production, speaker production choice is influenced by the cognitive availability of words and structures (Bock, 1987; Ferreira & Dell, 2000b). Availability effects can be hard to disentangle from the surprisal-based predictions of UID, but Zhan and Levy (2018) argued that the two make opposing predictions for Mandarin classifier choice, and that speaker preference in that case favors availability-based production over UID. In the Russian comparative alternation, availability-based production should predict that when an upcoming word or structure has low availability, speakers are more likely to use the explicit construction, which inserts an extra high-frequency word and “buys time”. An upcoming word or phrase may have low availability for several reasons. The word or phrase may be low-frequency in the language, long, or syntactically complex. To disentangle the predictions of the two theories, we consider two syntactic features that may play a role in availability but which are not directly related to the surprisal of the post-comparative word: the total length of the post-comparative noun phrase and the presence of relative clauses. Availability-based production should predict that syntactically more complex post-comparative NPs are less available, and would therefore be more likely to take the explicit construction. The UID account, meanwhile, should not be sensitive to these factors when controlling for the surprisal of the post-comparative word.

Separately, availability-based production may predict an effect of the degree of association between the comparative adjective and the post-comparative noun. For example, given the context “He is braver than...", the word “giraffe" may be less available than “lion" purely because lions are conceptually associated with bravery. We operationalize and investigate this potential effect in our corpus study.

Corpus Study

We used a subset of the Russian-language Taiga dataset (6*B* tokens), which includes text from a range of genres (Shavrina & Shapovalova, 2017; Shavrina, 2018). The Taiga dataset comes preprocessed and tagged with parts of speech, grammatical features such as case, and dependency relations in the CONLL format (Buchholz & Marsi, 2006).

We read the CONLL-format data and used a simple rule-based method to identify instances of explicit and genitive comparative constructions. For the explicit construction, we searched for sub-sequences consisting of a comparative adjective, the comparator word чем , and an adjective/noun/pronoun in the nominative case. For the genitive construction, we searched for sub-sequences consisting of a comparative adjective followed by an adjective/noun/pronoun in the genitive case. We additionally use dependency parse information to exclude clausal comparatives of the form “A is ADJ-er than B VERB", such as “He is taller than she thought", since clausal comparatives are not compatible with the genitive construction. Our approach for extracting comparatives does not capture every possible comparative sentence in the corpus, yet still results in approximately 100*K* comparative sentences, of which we randomly sample 50% for our corpus analysis. In order to test the effect of our information-theoretic and syntactic factors on comparative construction choice, we fit a mixed-effects logistic regression model with several theoretically-motivated predictors, which we describe in the following sections, before turning to the results of our model.

Syntactic Complexity of Post-Comparative Noun Phrase

We look at two measures of syntactic complexity: NP length and presence of relative clauses. To do this, we built a dependency graph of each sentence using the provided dependency annotations and the Networkx graph library (Hagberg, Schult, & Swart, 2008), and identify the head noun of each post-comparative NP. The head noun was defined as a) the word immediately after the comparative if it was a noun or pronoun and its head was the comparative adjective or b) the head of the immediately post-comparative word if that word was an adjective. Sentences where no head noun could be extracted using the above rules were excluded. Post-comparative noun phrase length was calculated by taking the number of ancestors of the post-comparative head noun in the

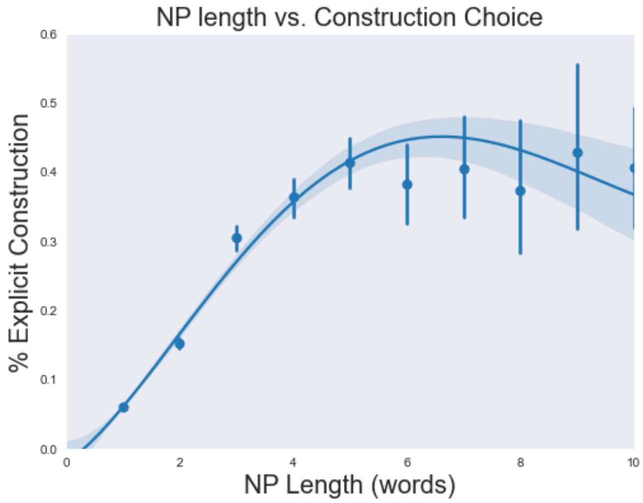


Figure 2: The relationship between length of post-comparative noun phrase and choice of construction with polynomial fit curve. As post-comparative NPs become longer, they become more likely to appear with the explicit construction. Error bars denote 95% CIs.

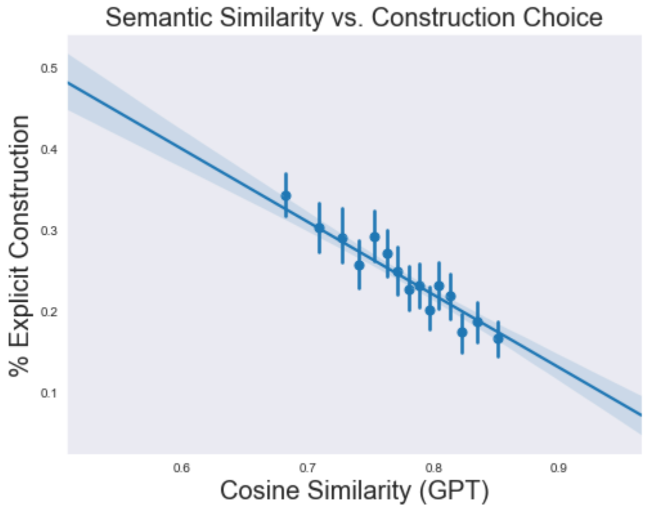


Figure 3: Adjective-noun cosine similarity (using Russian GPT model) vs. share of explicit construction. All instances of an (adjective, head noun) pair are collapsed into a single datapoint with the average explicit percentage, binning data into 15 equally sized bins. Error bars denote 95% CIs.

dependency graph and removing punctuation, the comparator word, and any subtrees connected via the dependency arc labels PARATAXIS or CONJ, which tended to create spuriously long NPs. After calculating NP length for each comparative sentence, we plotted the percentage of sentences using the explicit construction for each value of NP length, as shown in Figure 2. While larger values of NP length are more rarely attested and thus noisier, there is a clear positive correlation between NP length and the explicit construction.

A comparative sentence can also be made more complex by the attachment of a relative clause, as in the sentence “She is taller than the blonde girl who is standing by the window”. In Russian, relative clauses are introduced by a relative pronoun such as который . We identify relative clauses within a post-comparative NP using the ACL:RELCL arc label in a sentence’s dependency parse. We add a binary indicator for the presence of a relative clause but do not include this clause in the calculation for length of NP. Within our corpus, approximately 25% of sentences with relative clauses used the explicit construction, while only 11% of sentences without relative clauses used the explicit construction. We include presence of relative clause as a predictor in our mixed-effects model (see Table 1).

Calculating Surprisal with Counterfactual LMs

We computed surprisals for the post-comparative words in the corpus to test the UID hypothesis that the explicit construction would be preferred when the word following the comparative is high-surprisal. We estimated surprisal using neural network-based language models (LMs), which score words based on their probability of occuring given the preceding linguistic context. However, evaluating surprisal using LMs that are pretrained on naturalistic Russian text introduces a logical circularity - the predictions of an LM will depend on the distribution of the training data, which itself encodes the existing patterns for the comparative alternation in Russian. To avoid this circularity, we trained a language model from scratch on counterfactual data: a corpus that reflects a hypothetical version of Russian in which the explicit construction is the only way to express comparisons (see R. Levy & Jaeger, 2007, Zhan & Levy, 2018, and R. P. Levy, 2018 for discussion of this issue). This was created by automatically converting instances of the genitive construction in a training corpus to a semantically equivalent form that uses the explicit construction. For each sentence, we then calculated the surprisal of the post-comparative word in the counterfactual context. This allows a fair comparison of word surprisals across explicit and genitive examples. Two different LM architectures were used: a recurrent neural network based on (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018) and a Transformers model based on Fairseq (Ott et al., 2019). The RNN model backs off to the unigram surprisal (negative log word frequency) to handle out-of-vocabulary tokens (OOVs), while the Fairseq model uses byte-pair encoding to handle OOVs. We included post-comparative word surprisal as a predictor in our mixed-effects model (see Table 1).

Quantifying Semantic Association

Word embeddings are a dense vector representation of the distributional semantics of words (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The semantic association between two words can be modeled using the cosine similarity between their embeddings. We use cosine similarity to test the availability-based hypothesis that semantically associated adjective-noun comparisons would be more likely to use the genitive construction.

We generated word embeddings for comparative adjectives and head nouns in our dataset using two different models: GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018) [1] and Fasttext (Bojanowski, Grave, Joulin, & Mikolov, 2016). For GPT, we compute word embeddings for isolated words, not words in their sentence contexts; Fasttext word embeddings are non-contextual. For each sentence in the dataset, we computed the cosine similarity between the word embeddings of its comparative adjective and head noun. We included cosine similarity as a predictor in our mixed-effect model in the following section (see Table 1).

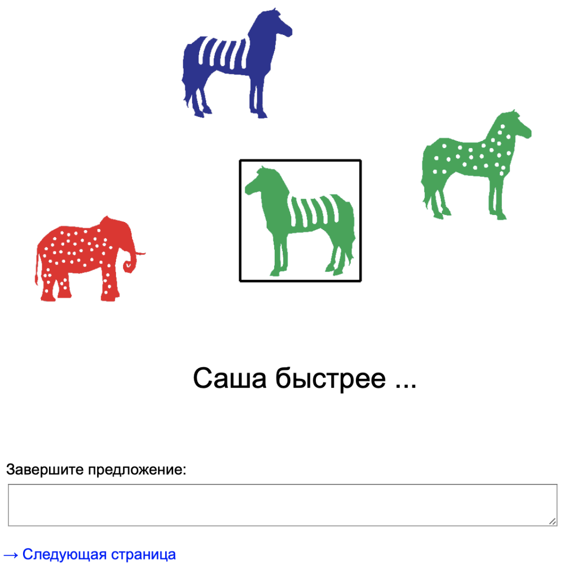
Mixed-Effects Model for Corpus Data

We fit a mixed-effects logistic regression model (using the `glmer` package in R) to predict whether a comparative sentence would be realized with the explicit construction. This model included random intercepts for the different post-comparative words and comparative adjectives, under the assumption that each word may have some idiosyncratic preference for one construction over the other. The model then fits fixed effects for the following predictors: length of post-comparative NP in words, LM-estimated surprisal of the post-comparative word, presence of a relative clause attached to the post-comparative NP, cosine similarity of comparative adjective and post-comparative head noun, and pronominality of post-comparative word. The fitted model parameters can be seen in Table 1. Figure 3 shows that there is a strong raw correlation between this measure of similarity and construction choice, with higher similarity favoring the genitive.

These parameters show that post-comparative NP length and presence of relative clause correlate significantly with the explicit construction. Meanwhile, post-comparative word pronominality and adjective-noun semantic association correlate significantly with the genitive construction. The relationship between word embedding cosine similarity and construction choice was directionally similar for both GPT and Fasttext word embeddings. Lastly, there was a significant correlation between the post-comparative word surprisal and the genitive construction (contrary to the UID predictions) regardless of whether the Transformer or RNN LM was used to estimate surprisal. Though not shown in the table, the results were qualitatively the same when the average surprisal of the entire NP was used instead of the immediately post-comparative word surprisal. The effect of surprisal had the opposite direction when not including random intercepts for the post-comparative word.

Behavioral Study

To complement our corpus study, we designed and conducted a production study with native Russian speakers to probe their choice in the comparative alternation in a production setting. This helps to rule out the possibility that the corpus study results are an artefact of the formal, published prose found in the text corpus. While this task still does not involve spoken production, participants generate language in a direct and spontaneous way that is closer to real-time production than published prose is.



Саша быстрее ...

Завершите предложение:

Figure 4: Example scene shown to participants, requiring species, color, and pattern to fully disambiguate the boxed animal. A sentence prompt in Russian (“Sasha is faster...") is shown below the scene; participants enter their completions in the text box.

Our experimental paradigm is designed to elicit comparative sentences with post-comparative noun phrases of varying lengths.[2] We achieved this through a sentence completion task using visual scene stimuli. Participants were shown a series of scenes of animals with varying colors and patterns, in the context of a story about a child named Sasha visiting the zoo. Accompanying each visual scene was the beginning of a sentence in Russian, such as Саша быстрее ... (“Sasha is faster...”). Subjects were instructed to complete the sentence by referring to the animal in the scene framed by a black box.

An example screen from the study is shown in Figure 4. Experimental stimuli were procedurally generated to randomize the locations and characteristics of the animals.[3] Crucially, the number of adjectives required to disambiguate the indicated animal from others in the scene varied from 0 to 3 in a controlled manner across stimuli. Each scene had a main animal and between 0 and 3 (inclusive) distractor animals, each of which differed from the main animal along one axis (i.e. color, pattern, or size), and an additional number of unrelated animals to bring the total number of animals per scene to 4. Subjects were instructed that their sentence completions should be specific enough that someone reading the sentence completion could select the intended animal without

[1] We use pretrained model weights for a Russian GPT-3 model released by Sberbank `https://github.com/sberbank-ai/ru-gpts`

[2] The study was pre-registered at `https://aspredicted.org/WLF_BL2`

[3] Original animal images from Freepik:macrovector, using free license, with modifications by the first author.

	Transformer/GPT	RNN/GPT	Transformer/Fasttext	RNN/Fasttext
(Intercept)	0.38 (0.39)	0.70 (0.42)	$-1.69\ (0.11)^{***}$	$-1.43\ (0.17)^{***}$
Post-comp Surprisal	$-0.03\ (0.01)^{***}$	$-0.04\ (0.01)^{**}$	$-0.03\ (0.01)^{***}$	$-0.04\ (0.01)^{***}$
NP Length	$0.38\ (0.01)^{***}$	$0.39\ (0.01)^{***}$	$0.39\ (0.01)^{***}$	$0.39\ (0.01)^{***}$
Relative Clause	$0.73\ (0.14)^{***}$	$0.77\ (0.14)^{***}$	$0.77\ (0.14)^{***}$	$0.78\ (0.14)^{***}$
Post-comp Pronoun	$-0.92\ (0.31)^{**}$	$-1.00\ (0.32)^{**}$	$-0.91\ (0.31)^{**}$	$-1.06\ (0.32)^{***}$
Adj-N Association	$-3.07\ (0.50)^{***}$	$-3.37\ (0.51)^{***}$	$-2.11\ (0.32)^{***}$	$-2.14\ (0.32)^{***}$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Mixed-effects logistic regression models. Column headers indicate the type of LM (Transformer vs. RNN) and type of word embedding (GPT vs. Fasttext) used. Positive coefficients indicate a correlation with the explicit construction. Coefficients are in logit-space, surprisal is in nats, NP length is in words, and cosine similarity ranges between 0 and 1. The remaining predictors are binary variables. Positive coefficients indicate that larger values of the predictor favor the explicit construction.

Саша сильнее пятнистого зеленого волка
Саша сильнее зеленого волка в белую точку
Саша сильнее чем зеленый волк в крапинку

Table 2: Example responses for a scene in which a spotted, green wolf was boxed. Participants used both comparative constructions and a variety of lexical forms to indicate the intended animal.

seeing the black box. This design was successful at eliciting sentence completions of varying lengths from the participants; the presence of each additional distractor animal led to a longer average completion length for that stimulus class.

The study was written using the Ibex software. 102 Russian native speaker participants were recruited using Prolific.[4] After being shown instructions and 3 practice examples, participants completed 24 experimental items and 24 filler items in randomized order, with 6 stimuli from each distractor condition. Filler items consisted of the same task, but with sentence prompts not containing comparative adjectives (e.g. “Sasha is not as fast as ..."). Participants were paid $3.75 for their time, and took approximately 15-20 minutes on average to complete the study. The study resulted in a small corpus of 2376 comparative sentences after excluding 96 responses for being blank, containing multiple or run-on sentences, or not answering the prompt. Example responses can be seen in Table 2. Rather than annotating the syntactic structure of every response, we simply used the length in words (excluding the explicit comparator word if present) of the sentence completion as a proxy for post-comparative NP length.

The top of Figure 5 shows a count histogram of NP length for each comparative form. For the genitive construction, the mass is concentrated towards shorter NPs and drops off quickly as NP length increases. For the explicit construction, the mass is distributed across a wider range of NP lengths, with a higher mean. While the genitive form is more common overall, the relative share of the explicit construction increases with NP length, as visualized in Figure 5. This aligns with the pattern observed in the corpus study.

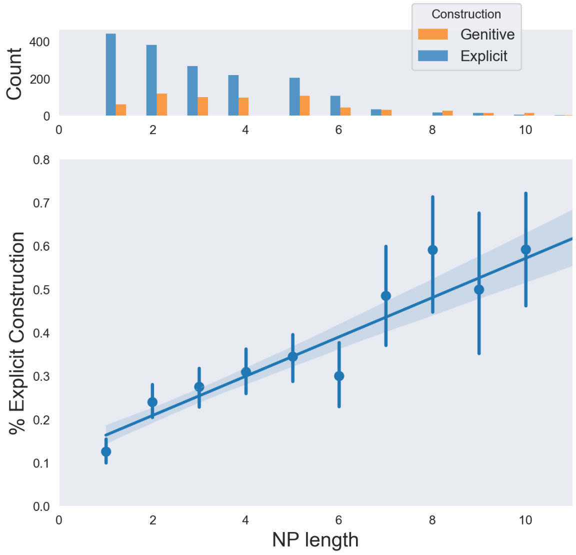


Figure 5: Length of post-comparative NP vs. share of explicit construction. Error bars denote 95% CIs.

Mixed Effects Model for Human Data

We analyzed the results in a similar way to the corpus study, creating a mixed-effects logistic regression model to predict construction from post-comparative NP length. The mixed-effects model used per-adjective and per-subject random slopes and intercepts to account for potential idiosyncratic preferences among subjects and adjectives for one form or the other. To avoid overfitting due to the small and constrained nature of the dataset, we regressed with only the NP length variable rather than the full suite of predictors used in the corpus study. The model's results (Table 3) are consistent with the raw proportions of Figure 5: longer NPs favor the explicit construction. This result holds despite substantial variation among participants in overall preference for the genitive versus the explicit construction.

[4] The study recruited 100 participants; 2 additional participants were timed out by Prolific due to technical issues and therefore did not count towards the participant limit, but their data were still saved.

	Model
(Intercept)	$-2.77\ (0.49)^{***}$
NP length	$0.19\ (0.07)^{**}$
Num. obs.	2376
Num. groups: participant	102
Num. groups: adjective	6

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Mixed effects model predicting choice of construction from length of NP, with random slopes and intercepts for participants and comparative adjectives. Positive coefficients indicate that larger values of a predictor favor the explicit construction.

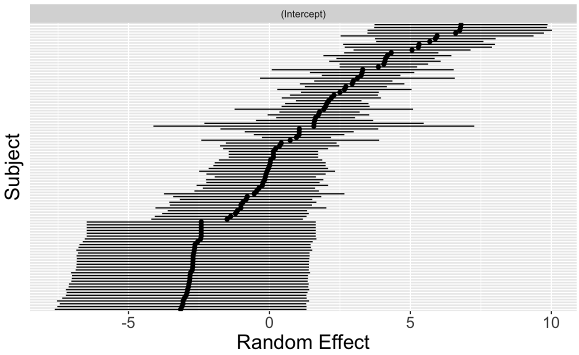


Figure 6: By-participant intercepts.

Discussion

The corpus study data are consistent with an availability-based account of speaker choice between the two comparative construction alternatives in Russian. Longer or syntactically more complex post-comparative noun phrases were correlated with a higher usage of the explicit comparative than shorter or more syntactically simple NPs. Additionally, higher cosine similarity of GPT-derived word embeddings between the comparative adjective and the head noun of the post-comparative NP was correlated with a preference for the genitive construction. In this case, cosine similarity of word embeddings is a proxy for semantic association, and suggests that adjective-noun combinations that are less associated or related (and therefore potentially less available) can trigger usage of the explicit form, resulting in an inserted word.

The correlation between post-comparative NP length and the explicit comparative construction held in our web-based native speaker production study as well, further supporting the predictions of an availability-based account. Interestingly, our data show that there is wide variability among native Russian speakers in their preferences for these two constructions - some speakers always use the genitive construction, some always use the explicit construction, and most fall somewhere in the middle. It is worth noting that the tendency to prefer the explicit construction in less available utterances is far from absolute - the genitive is still attested even with long and complex sentence completions.

This study intentionally focused on NP length as the sole predictor of construction choice because this could be controlled by varying the visual stimuli. Given the restricted domain of the task (dealing only with physical comparisons towards animals using a small set of adjectives and animals), measuring semantic association using word embeddings would likely have been subject to noise. Future work could use a modified experimental setup to keep NP length relatively constant and vary semantic association across a wide range of adjective-noun pairs.

The UID account does not explain the corpus study data, as the effect of post-comparative word surprisal (when including random intercepts) was in the opposite direction to what was predicted. UID also does not explain the significant impact of NP length. However, a variant of the UID account can potentially still be salvaged. While the explicit and genitive constructions are interchangeable for comparatives of the form “A is ADJ-er than B", a key difference between the two constructions is that the explicit construction can introduce not only a noun phrase, but an entire clause (e.g. “He is taller than my brother *thought*.”). The genitive construction can only introduce a noun phrase. As a result, when hearing or reading the explicit comparator чем followed by a nominative noun after a comparative adjective, a Russian speaker may actually experience increased uncertainty about the overall structure of the sentence. It may be the case that speakers follow UID by using the genitive construction when the post-comparative word could also start a clause with high probability. This may explain the higher tendency for pronouns to appear with the genitive construction, since pronouns are often the subjects of clauses and using the explicit construction could create temporary uncertainty about the syntactic structure of the utterance. Additionally, the relationship between semantic association and surprisal, including when and to what degree these quantities are correlated, merits further investigation.

Conclusion

This work used a corpus study and a native speaker elicitation study to provide support for an availability-based production account of speaker choice in the Russian comparative alternation. We take advantage of a large-scale, automatically annotated dataset of Russian to conduct a corpus study, and also show the feasibility of using visual scene stimuli to elicit comparative sentences of varying length. Our data and analysis indicate that speakers choose a construction with an extra word preferentially more when the post-comparative noun phrase is long, syntactically complex or less semantically associated with the comparative adjective, all factors that could influence availability-based production. When accounting for these other predictors, the post-comparative word surprisal does not have the effect predicted by UID, posing a challenge for this account. At the same time, this comparative alternation is likely a multi-faceted phenomenon, and other factors may influence it, such as the idiosyncratic variation between native speakers in their preference of construction as shown in the elicitation study.

Acknowledgments

This work has benefited from presentation at the 2022 Human Sentence Processing Conference. We gratefully acknowledge support from NSF grant BCS-2121074.

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*. doi: 10.1177/00238309040470010201

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, *26*. doi: 10.1016/0749-596X(87)90120-3

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Buchholz, S., & Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing.. doi: 10.3115/1596276.1596305

Ferreira, V. S., & Dell, G. S. (2000a). Effect of ambiguity and lexical availability on syntactic and lexical production. , *40*(4).

Ferreira, V. S., & Dell, G. S. (2000b). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*. doi: 10.1006/cogp.1999.0730

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 199–206).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In (Vol. 1). doi: 10.18653/v1/n18-1108

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (p. 11 - 15). Pasadena, CA USA.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University Stanford, CA.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*. doi: 10.1016/j.cogpsych.2010.02.002

Jain, A., Singh, V., Ranjan, S., Rajkumar, R., & Agarwal, S. (2018). Uniform information density effects on syntactic choice in hindi. *LC&NLP 2018*.

Kravtchenko, E. (2014). Predictability and syntactic production: Evidence from subject omission in russian. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction.. doi: 10.7551/mitpress/7503.003.0111

Levy, R. P. (2018). Communicative efficiency, uniform information density, and the rational speech act theory. In *Proceedings of the 40th annual meeting of the cognitive science society* (p. 684–689).

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021, November). Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 963–980). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.emnlp-main.74` doi: 10.18653/v1/2021.emnlp-main.74

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality..

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., … Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of naacl-hlt 2019: Demonstrations*.

Radford, A., Narasimhan, T., Salimans, T., & Sutskever, I. (2018). [gpt-1] improving language understanding by generative pre-training..

Shannon, C. E., & Weaver, W. (1949). The theory of mathematical communication. *International Business*.

Shavrina, T. (2018). Differential approach to web-corpus construction. In (Vol. 2018-May).

Shavrina, T., & Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: "taiga" syntax tree corpus and parser..

Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in mandarin chinese. In (Vol. 1). doi: 10.18653/v1/n18-1181