Quantifying scope and disrtribution of Slavic Short Adjectives

Szymon T.J. Kossowski, Nikita L. Beklemišev

3 Mar 2025

1 Executive Summary

This project aims to investigate the use of short forms of adjectives and participles in a typologically representative sample of Slavic languages and to revisit the syntactic and semantic cues that influence the choice of the adjective or participle form in a data-driven way. As part of the project, we are going to quantify and visualize these typological differences in the spirit of gradient approaches to syntax. [11] This can be helpful for further comparison and speculating about the typological prerequisites that resulted in different roles of the short and long forms of Slavic adjectives and participles (this could be for example word order or the role of metonimy [5] or anything else).

2 Background

Proto-Slavic adjectives and participles had two declensional paradigms: long (for definite declension) and short (for indefinite declension). [Wiktionary]

Such distinction, or at least a remnant of it, can to this day be found in diverse modern Slavic languages such as Russian, Serbo-Croatian+, Czech, Polish, Slovene, and Ukrainian. With the loss of definiteness as a grammatical category in most of Slavic languages, the distinction of adjective and participle forms assumed new functions. Russian and Serbo-Croatian preserved this distinction to a fuller extent [Rusgram], [4], Serbo-Croatian have even preserved the distinct paradigms for long and short forms, respectively, with separate declensions for all genders [Wiktionary], [1], see Table 1 for an example. Russian and Czech maintain the distinction in the nominative for all genders in both numbers. In languages such as Ukrainian, Polish, and Slovene, this distinction was preserved partially [Slovene, Polish, Ukrainian], limited to singular masculine nominative forms, whereas in Bulgarian, Macedonian, Belarusian, and Slovak this distinction disappeared completely. [Belarusian, Bulgarian, Macedonian, Slovak]

However, the exact extent to which the distinction is present in the speech leaves room for specification. As such, languages vary in the number of lexemes that can have these forms. Another consideration is the flexibility of their use:

in Polish they can mostly be found in a few fixed formulae, for example *bądź zdrów/litościw* 'be healthy/merciful', *jestem pewien/ciekaw* 'I am sure/curious' [p.c.], while in Czech we expect it to be syntactically predictable, and in Russian the distinction also carries semantic/stylistic information [Rusgram].

Sentence	**Gloss**	**Translation**
Marko je star	m. is old-SHORT	'Marko is old.'
Marko je sta:ri:	m. is old-LONG	'Marko is the old one.' (*Marko is old)

Table 1: Example of short and long adjectives in Serbo-Croatian. Both can be in fact found in predicate position, although the long form will be interpreted as substantivised. [1]

Therefore the following research questions correspond to each stage.

1. Quantifying productivity

   How would a map depicting the productivity of the short forms in Slavic languages look like?

2. Measuring surprisal of the form and quantifying the degree to which different factors affect it

   What and to what extent determines the form of an adjective: the syntactic role (predicate vs. modifier), the position in the phrase, definiteness of the noun it attributes to, collocation, or the exact adjective itself, inherently?

3 Scope

The novelty of the project is in comparing diverse Slavic languages with the same tradition-neutral methods. Therefore the study has to encompass as many languages as possible, as long as the adjective form distinction is preserved at least in some cases of relic forms, and there is a corpus of sufficient size and detailed enough annotation. This left us with a nice list of languages of diverse philogeny: Czech and Polish, Russian and Ukrainian, Serbo-Croatian+ and Slovene.

One of the challenges is that in some languages the distinction is only overt for the masculine gender. This makes the comparison more difficult and requires us to make an arbitrary choice—to exclude it from the data of that language.

4 Methods

4.1 Data acquisition

Since we plan to predict the adjective form based on syntactic information, for the 2nd question we need a well annotated corpus. Using `UD` [6] corpora of

the following languages can insure the comparability and consistent syntactic annotation. The drawback is the limited size of some of the corpora.

The 1st question, e.g. the productivity and presence of short forms, might require to describe the language more fully, e.g. include more types of adjectives. Because of that the sizes of UD corpora may not be sufficient. This lead us to considering the national corpora of the languages, such as `National Corpus of Polish` [12], `Russian National Corpus` [10], `General Regionally Annotated Corpus of Ukrainian` [9] etc. We haven't yet looked at the availability of the data for use outside of web interfaces.

Since the distinction might be present in the speech less, maybe using spoken corpora would enhance the study. [7] We are also very fond of using parallax corpora, since pretty much all frequency effects are very much affected by text type and size. [13] Unfortunately, we think that they are too short to capture short adjectives.

This means that we have to take some measures to ensure comparability of corpora data (and we haven't decided which yet: sample subcorpora of similar size? only look at top N most frequent lexemes?). For Serbo-Croatian and Slovene CLASSLA-Web Corpora [8] look like a possible solution.

4.2 Processing

Here are some measures we consider implementing for the processing of the data.

Lemmatization is the obvious first step.
To count the overall productivity of short adjectives:

- Baayen's formula $P = \frac{n_1}{N}$ [3, 2], where P stands for productivity rank, n_1 is the number of hapax legomena in the morphological category and N is the total number of tokens in the category.
- entropy of encountering a short adjective averaged across the vocabulary: Let $p(short)$ be the proportion of short adjectives for a given adjective lemma, then summing across adjectives $H = -\sum_{adjective} p(short) log_2 p(short)$.

To count dependence on predictors:

- linear regression on UD link (e.g. amod, root), position in a phrase (e.g. last, first), the role of the NP (nsubj / other—to approximate definiteness), and the adjective itself (must be a very important factor). We could also try to exclude this factor by using a mixed-effects model, but this could be difficult because of non-normailty and so on, and also if the the impact of the word group is present, it would be a significan contribution.
- information gain

Importantly, we assume that there is a semantic contribution, which would account for some uncertainty in prediction, for example, in Russian the choice between *Ja zdorov* 'I'm doing well' / *Ja zdorovy* 'I'm healthy' will depend on

the intended relevance of the attribute to "now". To calculate how unpredictable the form is, we can use conditional entropy.

4.3 Visualisation

The relevant data, as well as the findings, will be plotted using matplotlib and seaborn libraries for Python. GIS and mapping could be performed through the `geopandas` library.

5 Planning and roles

1. 17-23.03: reviewing corpora in terms of relevant data (Nikita), loading relevant data into the project (Szymon), extraction of instances (Szymon)
2. 24-30.03: data processing and organisation (Szymon), statistical hypothesis testing (Nikita)
3. 31.03-6.04: data visualisation (Nikita), GIS (Szymon), interpretation (Nikita), conclusions (Szymon)
4. 7.04-13.04: preparation for publication (Szymon), writing readme (Nikita)

NOTE: The schedule and role assignment are preliminary and may change in the course of the project.

6 Bibliography

Here is the link to our Drive where you can find some of the literature referenced.

References

[1] Nadira Aljović. Syntactic positions of attributive adjectives. In Patricia Cabredo Hofherr and Ora Matushansky, editors, *Adjectives: Formal analyses in syntax and semantics*, pages 29–52. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010.

[2] R. Harald Baayen. *41. Corpus linguistics in morphology: Morphological productivity*, pages 899–919. De Gruyter Mouton, Berlin, New York, 2009.

[3] R. Harald Baayen and Rochelle Lieber. Productivity and english derivation: a corpus-based study. *Linguistics*, 29:801–843, 1991.

[4] Leonard H. Babby. The syntactic differences between long and short forms of russian adjectives. In Patricia Cabredo Hofherr and Ora Matushansky, editors, *Adjectives: Formal analyses in syntax and semantics*, pages 53–84. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010.

[5] Mario Brdar, Rita Brdar-Szabó, Tanja Gradečak-Erdeljić, and Gabrijela Buljan. Predicative adjectives in some germanic and slavic languages: On the role of metonymy in extending grammatical constructions. *Suvremena lingvistika*, 51–52:35–57, 2001.

[6] Universal Dependencies Consortium. Universal dependencies. `https://universaldependencies.org/`, 2025.

[7] Nina Dobrushina and Elena Sokur. Spoken corpora of slavic languages. *Russian Linguistics*, 46:77–93, 2022.

[8] Centre for Language Resources and University of Ljubljana Technologies. Classla-web corpora. `https://www.classla.uni-lj.si/resources/corpora/`. Accessed: 2025-03-10.

[9] Ukrainian Language Corpus Research Group. General regionally annotated corpus of ukrainian (grac). `https://uacorpus.org/k-centre/corpora.html`. Accessed: 2025-03-10.

[10] Russian Academy of Sciences Institute for the Russian Language. Russian national corpus. `https://ruscorpora.ru/new/`. Accessed: 2025-03-10.

[11] Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883, 2023.

[12] Institute of Computer Science at the Polish Academy of Sciences. National corpus of polish. `https://nkjp.pl/index.php?lang=1&page=0`, 2012. Accessed: 2025-03-10.

[13] Stefan Schnell and Nils Norman Schiborr. Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics*, 8:171–191, 2022.