A Translation Mining Approach to Grammatical Aspect: Insights from Slavic and Baltic Languages

Keywords: grammatical aspect; Slavic&Baltic languages; translation mining; parallel corpus

Grammatical aspect encodes the internal temporal structure of events and has been extensively studied in Slavic languages. One of the most influential frameworks for analyzing Slavic aspect is Dickey's (2000, 2015) typology, which proposes an East-West divide. While this model has shaped research on aspectual systems, its empirical validation across a broader set of data and languages, particularly those outside the Slavic family, remains underexplored. This study applies the Translation Mining methodology that leverages a multilingual parallel corpus to detect cross-linguistic patterns in the distribution of grammatical aspect by analyzing how verbs are rendered across multiple translations of the same source text. Our study applies this approach to a newly compiled multilingual parallel corpus, consisting of the English original text of *Harry Potter and the Philosopher's Stone* and its translations into Polish, Czech, Slovak, Russian, Belarusian, Ukrainian, Serbian, Slovene, Macedonian, Bulgarian, Lithuanian, and Latvian. The corpus contains approximately 1,940 verb forms per language, manually annotated for tense and aspect.

We expect languages belonging to Dickey's Eastern and Western aspectual groups to differ in terms of aspect distribution. Polish and Serbian are expected to be more transitional with Polish leaning towards the Eastern group and Serbian towards the Western one.

We analyzed the data by employing a binomial regression model to examine the relationship between language and the likelihood in each of them to use each aspect. This method is appropriate for modeling binary dependent variables, where the response variable, Aspect, was coded as 1 for Perfective and 0 for Imperfective; therefore, the results presented below will show the estimate for a verb to be perfective across Baltic and Slavic languages. The predictor was Language. We fitted the model using a generalized linear model (GLM) with a binomial family and a logit link function, estimated via maximum likelihood. Model significance and effect sizes were assessed using Wald tests and odds ratios. The results of the statistics model are depicted in **Figure 1**. They confirm that Czech and Slovak represent the prototypical West aspectual system, showing that Slovak strongly decreases the chance of the verb being perfective (estimate= -0.533): verbs in Slovak are more likely to be imperfective; the result is statistically significant (p= 1.33e-13). Czech also strongly decreases the likelihood of a verb being perfective (estimate= -0.498); (p= 3.42e-12). The second group of languages also showing a significant decrease in the likelihood of verbs being perfective includes Serbian (estimate= -0.396; p= 2.58e-08), Macedonian (estimate= -0.394; p= 3.09e-08), Croatian (estimate= -0.331; p= 3.56e-06), Slovenian (estimate= -0.313; p= 1.70e-05), Bulgarian (estimate= -0.278; p= 9.64e-05), Polish (estimate= -0.277; p= 1.83e-04) and_Latvian (estimate= -0.269; p= 0.000253). Lithuanian (estimate= -0.146) and Ukrainian (estimate= -0.134) also decrease the likelihood of being perfective, but the effect is weaker: in both languages, the effect is not significant (the p-values are 0.0527 in Lithuanian and_0.0714 in Ukrainian). Finally, the model also showed that the two prototypical East aspectual systems, Russian and Belarusian, are the only languages where the likelihood of perfective aspect is increased. Russian only slightly increases the likelihood of perfective aspect (estimate= 0.012), but the effect is not significant (p= 0.8779). Conversely, Belarusian significantly increases the likelihood of perfective aspect (estimate= 0.324; p = 1.54e-09). Our findings confirm the existence of an East-West aspectual divide, with Russian, Belarusian, and Ukrainian forming the core of the Eastern aspectual system and Czech and Slovak exemplifying the Western type. However, our results also reveal transitional systems that were not fully accounted for in Dickey. In particular, Bulgarian and Macedonian, classified as the Eastern group, exhibit aspectual behaviors aligning them closer to the Western group. Moreover, our analysis highlights that Baltic languages, not included in Dickey's model, display aspectual tendencies that align them with different sections of the aspectual spectrum. Lithuanian appears closer to the Eastern group, whereas Latvian patterns more closely with the Western-oriented transitional systems.

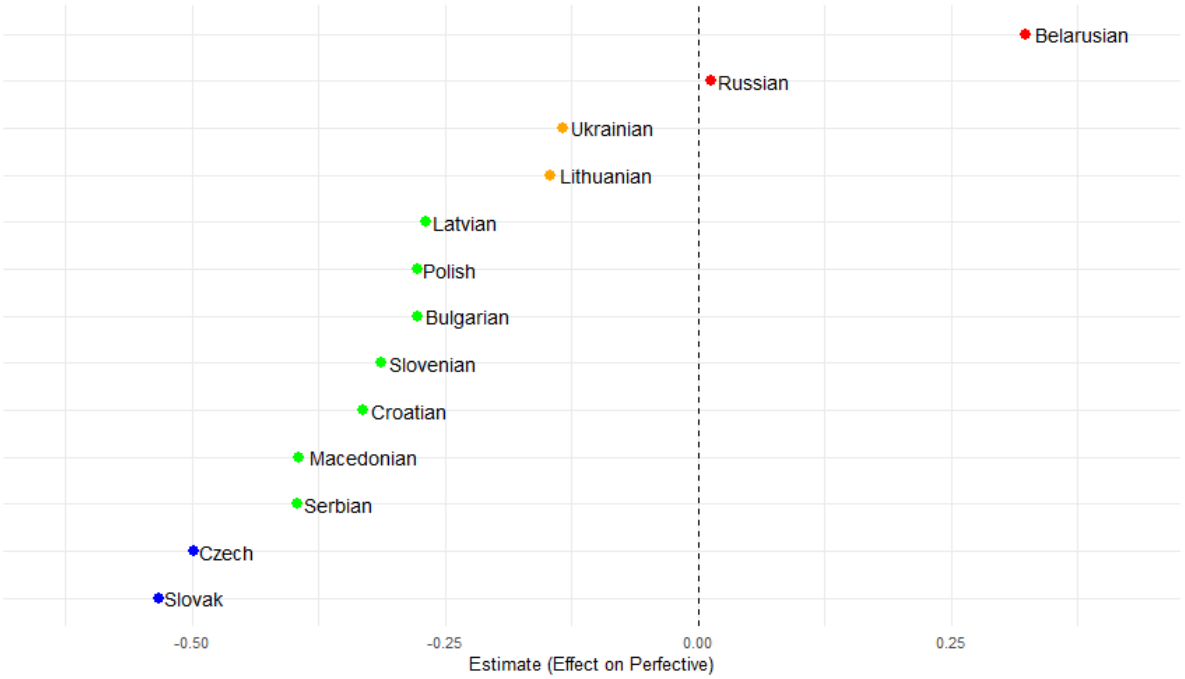

Figure 1. Effect of language on aspect use (estimate of the likelihood for a verb to be perfective in Baltic and Slavic languages)

For space reasons, we do not present corpus data to support our findings in the abstract but we will provide a lot of interesting cross-linguistic data during the talk.

Selected references

Metadata listing all the texts in the corpus: https://osf.io/5vwkg

Dickey, S. M. 2000. *Parameters of Slavic Aspect: A Cognitive Approach*. Stanford: Center for the Study of Language and Information.

Dickey, S. M. 2015. “Parameters of Slavic Aspect Reconsidered: The East-West Aspect Division from a Diachronic Perspective.” In: Shrager, M., E. Andrews, G. Fowler and S. Franks (eds.). *Studies in Accentology and Slavic Linguistics in Honor of Ronald F. Feldstein*. Bloomington: Slavica Publishers, 29–45.