Universal Dependencies for the Selice Variety of Romani

Lucie Zemanová
Charles University, Faculty of Arts
Prague, Czechia
lucka.zemanova007@gmail.com

Relevant UniDive working groups: WG1, WG2

1 Introduction

The aim of the project described in this abstract is to provide a treebank for the Selice variety of the Romani language within the framework of Universal Dependencies (UD)[1] (de Marneffe et al., 2021), hopefully accompanied by spoken data recordings.

Romani is an Indo-Aryan (macro)language spoken over a vast area mainly in Europe. Despite the sizable population (estimates vary but it seems safe to say that the total number of speakers exceeds three million), it is a minority language in all countries in which it is spoken. Romani has countless dialects, which are not always mutually intelligible. There are 12 dialect groups, our focus is on South Central dialect, spoken in Slovakia, Hungary, Austria and Slovenia (Elšík and Beníšek, 2020); specifically, we are interested in the variety spoken in the village of Selice; see also Figure 1.

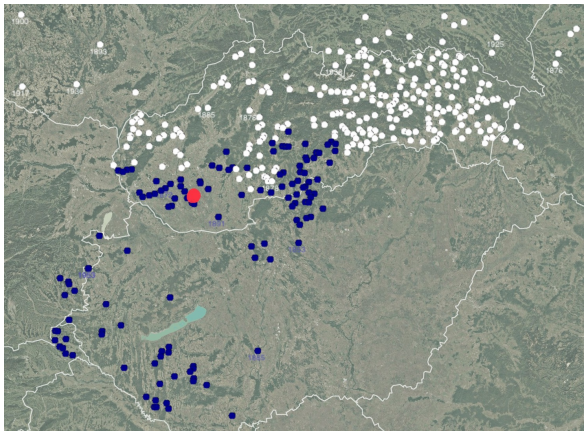

Figure 1: Central Romani dialects on the map: White – North Central, Blue – South Central, Red – Selice.[2]

Romani as a whole is a low-resource language (which may be due to the fact that Romani is in general mostly an oral language).

[1] https://universaldependencies.org/
[2] http://atlasofcentralromani.ff.cuni.cz/#/feature=1150

2 Selice Romani

There are many varieties of Romani in Slovakia. Larger share of them belongs to the North Central group (also called "Slovak Romani"), while the South Central dialects are found closer to the Hungarian border (and in Hungary). The variety we focus on is considered part of the latter group.

One of the interesting features of Selice Romani is the contact with other languages, resulting in borrowed vocabulary and grammar. Along with, for example, Persian, Greek and Slavic languages, it was mainly influenced by Hungarian (Elšík, 2007, 2009). Releasing the fieldwork data annotated as a UD treebank will provide an ideal resource for showing grammatical similarities and dissimilarities with other languages that already have data annotated in the UD style: Hungarian and Slovak (as contact languages), Hindi and other Indo-Aryan languages (as genealogical relatives), Persian, Greek and other Balkan languages (as older contact languages during the Romani migration to Central Europe). For illustration of Hungarian borrowings see Figure 2.

3 Methodology

3.1 Data collection

We are in contact with Viktor Elšík, who did fieldwork in Selice (mostly by letting the consultant tell stories from her life) and even more work afterwards, and kindly offered to share his data with us. More than 1000 sentences will be available for us for annotation and publication. His guidance is indispensable in our understanding the language.

We also want to collect more data ourselves in a planned fieldwork in Selice. To have a richer collection, we will use different techniques such as asking consultants to describe a picture or narrate a video and maybe even collect some folk stories or legends.

3.2 Annotation

We have created a finite-state morphological analyzer for Selice Romani, inspired by existing

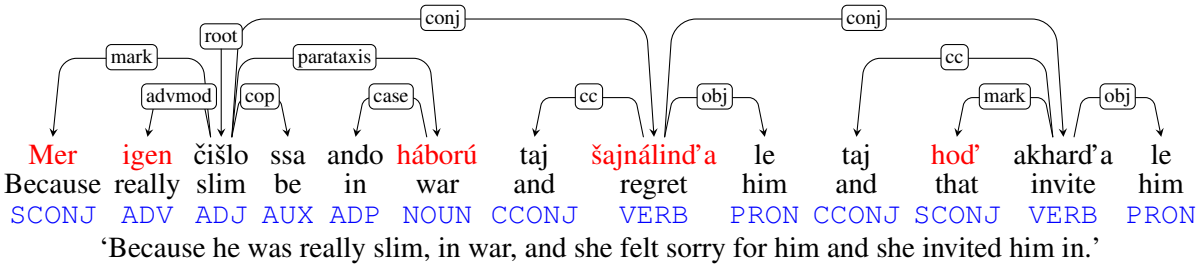

Figure 2: Selice Romani tree.
(Red – Hungarian borrowings.)

attempts at similar tools for Finnish Romani[3] (Granqvist, 2023) and Soviet Romani.[4] We initially started the syntactic annotation in TrEd[5] which we connected with the analyzer, so it could create UPOS tags and translation for the CoNLL-U file automatically. But then we decided to switch to annotation in Arborator-Grew (Guillaume, 2021), which allowed better cooperation. We are looking into binding Arborator-Grew with the morphological analyzer as well.

Once the first batch of sentences is annotated, the UDPipe 2 parser (Straka, 2018) will be trained and used to preprocess data for the subsequent batches, in order to speed up and facilitate the manual checking.

We have not annotated many sentences yet due to the fact that most of the collected spoken data is not digitized yet. We will continue once we have more transcribed data in digital form or new data that we will gather ourselves.

3.3 Spoken data

As we already mentioned, we would like to add the recordings to the annotation. We started the annotation of the spoken data during our STSM in September and created a set of tags for this purpose. We hope to be able to make at least some of them available; nevertheless, the selection will be constrained by the need to protect the personal data of the speakers.

4 Conclusion

We propose a work-in-progress. Its main outcome will be the first UD treebank in the Romani language; a rule-based morphological analyzer will be a secondary outcome. The work will help deepen the diversity within the UD language collection, as well as increase accessibility and visibility of the existing fieldwork data. The ultimate goal is to provide enough data for training a parsing model, which could help linguists and hopefully also contribute to documentation and digital support of the language.

[3] https://www.kielipankki.fi/cgi-bin/romtwol.py

[4] https://github.com/burushona/uniparser-soviet-romani

[5] https://ufal.mff.cuni.cz/tred/

Acknowledgements

We would like to thank Viktor Elšík, Dan Zeman and Nikolett Mus for their help and guidance.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Viktor Elšík. 2007. Grammatical borrowing in Hungarian Rumungro. In *Grammatical Borrowing in Cross-Linguistic Perspective*, pages 261–282. De Gruyter Mouton.

Viktor Elšík. 2009. Loanwords in Selice Romani, an Indo-Aryan language of Slovakia. In *Loanwords in the World's Languages: A Comparative Handbook*, pages 260–303. De Gruyter Mouton.

Viktor Elšík and Michael Beníšek. 2020. Romani dialectology. In *The Palgrave Handbook of Romani Language and Linguistics*, pages 389–425. Palgrave Macmillan, Cham, Switzerland.

Kimmo Granqvist. 2023. Kieliteknologiaa Linuxilla: Suomen romanikielen automaattinen muoto-opin analysaattori ja korjausohjelma.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th*

Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 168–175, Online. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

All internet resources were up to date on 6th April 2026.