The Parallel Corpus of Russian and Ruska Romani Languages

Kirill Koncha[1,2*], **Abina Kukanova**[3], **Tatiana Kazakova**[3], **Gloria Rozovskaya**[3]
[1]University of Groningen, [2]Ghent University, [3]HSE University

Abstract

The paper presents a parallel corpus for the Ruska Romani dialect and Russian language. Ruska Romani is the dialect of Romani language attributed to Ruska Roma, the largest subgroup of Romani people in Russia. The corpus contains translations of Russian literature into Ruska Romani dialect. The corpus creation involved manual alignment of a small part of translations with original works, fine-tuning a language model on the aligned pairs, and using the fine-tuned model to align the remaining data. Ruska Romani sentences were annotated using a morphological analyzer, with rules crafted for proper nouns and borrowings. The corpus is available in JSON and Russian National Corpus XML formats. It includes 88,742 Russian tokens and 84,635 Ruska Romani tokens, 74,291 of which were grammatically annotated. The corpus could be used for linguistic research, including comparative and diachronic studies, bilingual dictionary creation, stylometry research, and NLP/MT tool development for Ruska Romani.

1 Introduction

Ruska Romani is the dialect of Romani language attributed to Ruska Roma, the largest subgroup of Romani people in Russia. Ruska Roma makes up at least 50% of all Romani people in Russia and the number of speakers of the dialect can be estimated at 70-90 thousand (Kozhanov, 2018).

In general, the Ruska Romani dialect is characterized by a significant influence on the Russian language at all levels It also contains a significant number of lexical grammatical borrowings from German, and Polish (Kozhanov, 2018). The Cyrillic script for the Ruska Romani dialect was developed by Dudarova and Pankov (1928).

This paper presents the parallel corpus for Ruska Romani and Russian languages. To create the corpus, we found Russian literature translated into Ruska Romani in the 1920-30s. We choose Russian literature of that period as a large number of Russian texts were translated then into minority languages as a part of government language policy (Gurbanova and Rangsikul, 2018). As a result, many of Ruska Romani written texts were created. Even today these texts make up the majority of literature written in Ruska Romani. Moreover, many of them are available in the machine-readable format. The translated texts include both fiction and nonfiction domains. We manually aligned a subsample of sentences from the translations with sentences from the original works and fine-tuned LaBSE model (Feng et al., 2022) on the aligned pairs and used it within *lingtrain-aligner* library[1] for Python to align the rest of the data. Finally, Ruska Romani sentences were annotated using *uniparser-soviet-romani*[2] library and our manually crafted rules.

The parallel corpus of Russian and Ruska Romani is available as the part Russian National Corpus (RNC)[3]. The data both in JSON format and XML format of the RNC together with code are also publicly available via our repository[4]. The corpus includes 88,742 Russian tokens and 84,635 Ruska Romani tokens, 74,291 of which are grammatically annotated.

Our parallel corpus could be used for:

- Comparative studies of Russian and Ruska Romani;
- Diachronic studies of vocabulary and grammar of Ruska Romani;
- Creation of bilingual dictionaries and study materials for Ruska Romani;

[*] Work is partially done while at HSE University.

[1] https://github.com/averkij/lingtrain-aligner

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

[3] https://ruscorpora.ru

[4] https://github.com/kirillkoncha/ruska_romani

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 1–5
August 16, 2024 ©2024 Association for Computational Linguistics

- Stylometry studies, investigating the influence of a translator on authorship attribution;
- Creation of NLP and MT tools for Ruska Romani.

Moreover, it contributes to the representation Ruska Romani dialect and Romani culture overall.

2 Resources for Ruska Romani

The Ruska Romani dialect is one of the most described Romani dialects. It was described in grammars by Shapoval (2007); Ventcel' (1964) and has several dictionaries created by Sergievskij and Barannikov (1938); Demeter-Charskaya (2007); Vasilevskij (2013).

However, there are almost no electronic resources for Ruska Romani dialect. The only exceptions are Romani Corpus[5] and digitized version of Russian — Ruska Romani dictionary from Shapoval (2007)[6]. The Romani Corpus contains Ruska Romani texts published in the USSR in the 1920s and 1930s (both original works and translations from Russian). The corpus consists of 720K tokens. The morphological annotations of the tokens in the corpus were not disambiguated.

Despite a large number of translations from Russian to Ruska Romani, there are no parallel corpora for these two languages.

3 Corpus Generation

3.1 Data

To create the corpus, we used Russian texts and their translations to Ruska Romani created in the 1920s and 1930s[7]. Both original texts and their translations are written in Cyrillic script. All the text sources and their metadata are presented in Table 1. The texts we used for creation of the corpus partially overlap with texts in The Romani Corpus. However, The Romani Corpus does not contain aligned sentence equivalents in Russian.

3.2 Sentence Alignment

Methods. Sentence alignment is the task of matching up equivalent sentences within the same texts in different languages. Hunalign (Varga et al., 2007) is one of the most popular tools for sentence alignment. It uses statistical models and heuristics to identify corresponding sentences based on similarity measures, such as word order and context. Another solution for this task is Vecalign (Thompson and Koehn, 2019), which employs vector space models to align sentences in a parallel corpus.

Lingtrain-aligner. We used the *lingtrain-aligner* library for the alignment task, an approach that combines both sentence embedding similarities and heuristics. Firstly, *lingtrain-aligner* selects sentence pairs with the closest vector similarity obtained from a multilingual model from an unaligned text. Then, it computes the *chain score*, a metric that estimates how well sentence indexes align with each other based on the number of breaks or discontinuities in ordered sentence pairs. The metric will be equal to *0* if all pairs are selected randomly and equal to *1* if a single line without breaks is obtained. Finally, *lingtrain-aligner* automatically resolves conflicts (cases of breaks or discontinuities) by splitting or combining sentences from one sequence.

Model for Ruska Romani. However, a language model that is trained in both languages is needed to use the *lingtrain-aligner* library. As there was no model for Ruska Romani, we trained the LaBSE model (Feng et al., 2022)[8] on Russian and Ruska Romani sentence pairs using *chain score* as an evaluation metric. The chain score was aggregated over multiple batches of ordered sentence pairs (i.e., sentences were given as they appear in original texts) several times during each epoch.

Training Set. To train the model, we used sentence pairs from randomly selected titles: *Dubrovskij*, *Malen'kie rasskazy*, *Tri medvedya*, *Posle bala*. We matched each original sentence with a translated sentence by their indexes. Then, the linguist annotator manually checked and corrected matches using Ruska Romani grammar and dictionary from Shapoval (2007). If two sentences in one language corresponded to one sentence in another language, the annotator merged these sentences into one line. The cases, when a sentence in one language did not correspond to any in another language were allowed (but were not used during training). The following texts were aligned for model training: *Dubrovskij*, *Malen'kie rasskazy*, *Tri medvedya*, *Posle bala*. In total, these texts contain 24,700 Russian tokens and 25,170 Ruska Romani tokens.

[5] http://web-corpora.net/RomaniCorpus

[6] https://romadict.linghub.ru

[7] We express our gratitude to Kirill Kozhanov for sharing the machine-readable texts from Romani Corpus.

[8] https://huggingface.co/sentence-transformers/LaBSE

Russian Title	English Title	Domain	Original Author	Year	№. Tokens Russian	Romani Title	Transl. Author	Transl. Year	№. Tokens Romani
Dubrovskij	Dubrovsky	Fiction	A. S. Pushkin	1833	19,249	Dubrovsko	A. Svetlovo	1936	19,957
Posle bala	After the Ball	Fiction	L. N. Tolstoy	1903	3,103	Koli progyya balo	N. Pankovo	1936	2,764
Tri medvedya	Three Bears	Fiction		1875	493	Trin rychya		1937	497
Spat' hochetsya	Sleepy	Fiction	A. P. Checkhov	1888	1,584	Te soves kamelpe	A. Svetlovo	1934	1,552
Van'ka	Vanka	Fiction		1886	1,157	Van'ka		1934	1,241
Malen'kie rasskazy	A Small Stories	Fiction	A. S. Neverov	1922	1,855	Rakiribena vash tykne chyavorenge	G. Lebedevo	1930	1,952
V brigade proryv	There's a Breakthrough in the Brigade	Fiction	M. A. Sholokhov	1930	5,126	Dre brigada proriskiribe	O. Pankovo	1934	5,299
Esli vrag ne sdayotsya, – ego unichtozhayut	If The Enemy Does Not Surrender, He is to Be Destroyed	Publicism	A. M. Gor'kij	1930	815	Koli vrago na zdelape les has'kirna	M. Bezlyudskij	1930	583
Strasti-mordasti	Fat-Faced Passion	Fiction		1913	4,069	Strasti-mordasti	A. Svetlovo	1934	4,299
Druzhki	Buddies	Fiction		1898	3,819	Druzhke		1934	3,200
Zlodei	Villains	Fiction		1901	6,390	Zlodei		1934	6,038
Mal'va	Malva	Fiction		1897	12,146	Mal'va		1934	12,979
Rozhdenie cheloveka	The Birth of a Man	Fiction		1898	2,758	Manusheskiro biyanype		1935	2,358
Na plotah	On Rafts	Fiction		1895	3,409	Pro ploty		1936	3,379
Tovarishchi	Comrades	Fiction		1895	3,845	Tovarishshi		1937	3,519
Makar Chudra	Makar Chudra	Fiction		1892	6,062	Makar Chudra		1932	3,900
Emel'yan Pilyaj	Emelyan Pilyay	Fiction		1893	3,550	Emel'yano Pilyaj		1932	2,829
K rabochim i krest'yanam	To Workers and Peasants	Publicism		1930	1,159	Ko butyar'ya	M. Bezlyudskij	1930	1,102
Son Makara	Makar's Dream	Fiction	V. G. Korolenko	1885	7,613	Makaroskiro soibe	A. Svetlovo	1935	7,187
Total					88,742				84,635

Table 1: Texts Sources

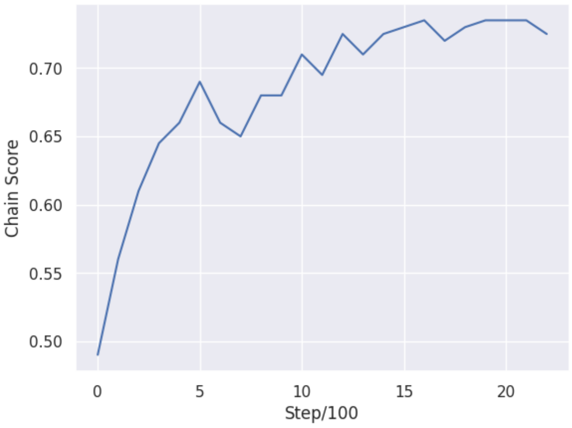

Figure 1: *Chain score* values during model training.

Training Model. The model was trained on *7* epochs or *2100* steps (each epoch had 300 steps) with batch size *6*. The evaluation was performed every *100* steps. The best *chain score* equal to *0.74* was achieved at *200* step of *5* epoch (*1600* step). The observed evaluation metrics during training are presented in Figure 1.

We used the best-trained model with *lingtrain-aligner* to automatically align the rest of the texts and resolve conflicts.

Errors Correction. Additionally, cosine similarities of each sentence pair were computed. Sentence pairs with cosine similarity below 0.5 were checked by the linguist annotator if necessary manually corrected the same way the training set was aligned. Overall, 881 sentence pairs out of 8,127 (11%) were assessed.

3.3 Morphological Annotation

For morphological annotation, we used the *uniparser-soviet-romani* library. It is a morphological analyser for Ruska Romani created based on *uniparser-morph*[9], a parser developed primarily for under-resourced languages. The parser for Ruska Romani does not perform disambiguation of analyses. Therefore, all possible annotations are given for each token. Annotations include lemma and its translation, part of speech, case, person, gender, tense, and many other features. Frequently, *unipaser-soviet-romani* dictionaries do not contain entries for loanwords from Russian or proper nouns. In total, only 84% (71,287) of tokens were annotated.

In cases where the *uniparser-soviet-romani* library did not provide the annotation, we implemented a multi-step approach to analyse nouns and proper names. Firstly, we examined whether a word has a Ruska Romani suffix. Subsequently, we removed the Ruska Romani suffix from the word. Then, we checked the word presence in Russian and its grammatical properties using *PyMorphy2* (Korobov, 2015). The final annotation process took into account both the grammatical properties of the Russian word and the grammatical properties associated with Ruska Romani suffixes. These rules allowed us to increase the amount of annotated tokens by 4% or 3,004 tokens (Figure 2).

After annotation, we converted each *uniparser-soviet-romani* word tag into RNC format. Explanations of each tag are given in the project repository (see the link above).

We did not annotate original sentences as RNC uses its tools to annotate texts in Russian (see Lyashevskaya et al. 2023; Savchuk et al. 2024).

[9] https://github.com/timarkh/uniparser-morph

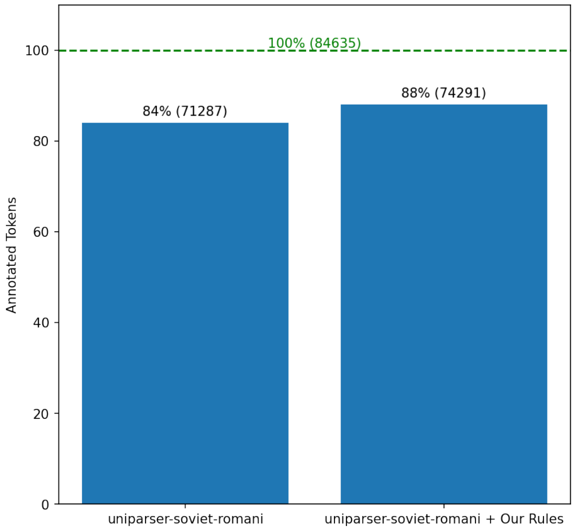

Figure 2: Annotated tokens without and with our annotation rules.

4 Data Format

Our parallel corpus is available in two formats: JSON and RNC XML. See the project repository for a more detailed description.

4.1 JSON Format

JSON annotations consist of the following fields:

- **sentence_rus**: sentence in Russian;
- **sentence_roma**: sentence in Ruska Romani;
- **sentence_id**: id of a sentence;
- **words_roma**: annotations of each Ruska Romani token in a sentence.

The first two fields are strings, the third field is numeral, and the field *words_roma* is a nested list. Each item in **words_roma** is a list of dictionaries with all possible annotations of a corresponding word in a sentence. For example, the first list in the field will contain all possible annotations of the first word in a sentence.

The annotation dictionary has the following fields:

- **wf**: word form;
- **lemma**: normalised form of a word;
- **gramm**: grammatical features of a word, such as part of speech, gender, case, etc;
- **wfGlossed**: word form divided into morphological elements by hyphens;
- **trans_en**: English translation of a word;
- **trans**: Russian translation of a word.

4.2 Russian National Corpus XML-format

We automatically converted JSON data into the XML format of RNC. The XML body consists of the following containers:

- **<para>**: container for a sentence pair, includes attributes **id** and **id_str**;
- **<se>**: container for a sentence within a sentence pair, includes attribute **lang** that could either be **rus** for Russian or **rom** for Ruska Romani;
- **<w>**: container for a word level annotation, applied only to sentences in Ruska Romani;
- **<ana>**: container inside **<w>** that stores grammatical features of a word.

The **<ana>** container has following attributes:

- **lex**: lemma of a word;
- **wordf**: word form;
- **gr**: grammatical features of a word;
- **transl**: Russian translation of a word.

  One word container **<w>** could include several annotation containers **<ana>**.

5 Conclusion

We presented the parallel corpus for Russian and Ruska Romani languages. For sentence alignment, we used a model trained on manually aligned sentences. We also manually checked alignment in sentence pairs, where the model predicted low similarity for sentences. Ruska Romani sentences in the corpus were annotated using *uniparser-soviet-romani* library and our own manually crafted rules. The data is available in JSON and RNC XML formats.

Our work could be used in different areas: from linguistic research and language teaching to the creation of NLP tools and resources for Ruska Romani. It also contributes to the representation of Ruska Romani dialect and Romani culture as the corpus is available in RNC, one of the largest platforms with resources for the Russian language and minority languages of Russia.

Limitations

The present work has several limitations. The first limitation is the absence of disambiguation in morphological annotation. Secondly, in the case of Russian sentence annotation, we rely on RNC annotation tools which are not publicly available. Finally, the corpus includes only translations of Russian literature and does not include any spoken language. Moreover, the texts were translated a long time ago and might not fully reflect the current state of the Ruska Romani dialect.

Acknowledgments

We express gratitude to Kirill Kozhanov for sharing texts in Ruska Romani; to Timofey Arhangel'skij for creating *uniparser-soviet-romani*; to Dmitry Sitchinava for help with data conversion to RNC format; to Olga Lyahevskaya, Svetlana Toldova, Denis Pisarenko, and Maxim Bazhukov for valuable feedback; to Fedor Samodov for help with manual sentence alignment.

References

O.S. Demeter-Charskaya. 2007. *Cygansko-russkij i russko-cyganskij slovar' (dialekt russkih cygan)*. Moskva.

N. A. Dudarova and N. A. Pankov. 1928. *Nevo drom. Bukvare vash bare manushenge*. Moscow.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Nubar Gurbanova and Rungthum Rangsikul. 2018. Language in politics features of the soviet language policy in 1920s-1930s. In *Proceedings of the International Conference on Language Phenomena in Multimodal Communication (KLUA 2018)*, pages 418–422. Atlantis Press.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Kirill Kozhanov. 2018. Cyganskij yazyk i ego dialekty. In N.G. Demeter and A.V. Chernyh, editors, *Cygane*, Narody i kul'tury, pages 156–159. Nauka, Moscow, Russia.

Olga Lyashevskaya, Ivan Afanasev, Sergey Rebrikov, Yulia Shishkina, Elvira Suleymanova, Ivan Trofimov, and Natalia Vlasova. 2023. Disambiguation in context in the russian national corpus: 20 years later. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conf. "Dialogue"*, volume 22, pages 307–318.

S. O. Savchuk, T. A. Arhangel'skij, A. A. Bonch-Osmolovskaya, O. V. Donina, YU. N. Kuznecova, O. N. Lyashevskaya, B. V. Orekhov, and M. V. Podryadchikova. 2024. Nacional'nyj korpus russkogo yazyka 2.0: novye vozmozhnosti i perspektivy razvitiya. *Voprosy yazykoznaniya*, 2:7–34.

M.V Sergievskij and A.P. Barannikov. 1938. *Cygansko-russkij slovar': Okolo 10000 slov s prilozheniem grammatiki cyganskogo yazyka*. Moskva.

V.V. Shapoval. 2007. *Samouchitel' cyganskogo yazyka*. Moskva.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Dávid Varga, Péter Halácsy, András Kornai, Nagy Viktor, László Nagy, Lajos Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. In *Amsterdam Studies in the Theory and History of Linguistic Science. Series 4, Current Issues in Linguistic Theory*, pages 247–258.

N.A. Vasilevskij. 2013. *Romany chib: Cygansko-russkij slovar'*. Kaliningrad.

T.V. Ventcel'. 1964. *Cyganskij yazyk (severnorusskij dialekt)*. Moskva.