The Use of Khislavichi Lect Morphological Tagging to Determine its Position in the East Slavic Group

Ilia Afanasev
National Research University Higher School of Economics
ilia.afanasev.1997@gmail.com

Abstract

The study of low-resourced East Slavic lects is becoming increasingly relevant as they face the prospect of extinction under the pressure of standard Russian while being treated by academia as an inferior part of this lect. The Khislavichi lect, spoken in a settlement on the border of Russia and Belarus, is a perfect example of such an attitude.

We take an alternative approach and study East Slavic lects (such as Khislavichi) as separate systems. The proposed method includes the development of a tagged corpus through morphological tagging with the models trained on the bigger lects. Morphological tagging results may be used to place these lects among the bigger ones, such as standard Belarusian or standard Russian.

The implemented morphological taggers of standard Russian and standard Belarusian demonstrate an accuracy higher than the accuracy of multilingual models by 3 to 15%. The study suggests possible ways to adapt these taggers to the Khislavichi dataset, such as tagset unification and transcription closer to the actual sound rather than the standard lect pronunciation. Automatic classification supports the hypothesis that Khislavichi is a border East Slavic lect that historically was Belarusian but got russified: the algorithm places it either slightly closer to Russian or to Belarusian.

1 Introduction

Automatic classification of lects that are both closely related and low-resourced has been the target of dialectology studies for the last two decades, because it provides insights on the linguistic variation, used both for developing language tools and language studies (Nerbonne et al., 1999) (Gooskens and Heeringa, 2004) (Snoek, 2013) (Campos et al., 2020b). However, the morphological tagging results were rarely considered to be the basis for automatic classification. The main goal of this research is to develop a morphological tagger for a low-resourced East Slavic lect of Khislavichi with the help of the lects that possess significantly more resources, standard Russian and standard Belarusian. After the tagger is built, the morphological variation becomes the main subject of study: what differences between standard Russian, standard Belarusian, and Khislavichi stop the tagger from the correct cross-prediction between these lects? The automatic classification of standard Russian, standard Belarusian, and Khislavichi lects with distance-tree matrix (Bapat, 2010) demonstrates how it is possible to specify the position of a low-resourced lect in the context of much bigger lects that are phylogenetically connected to it.

The neutral term *lect* is used instead of *dialect* and/or *language* because the latter often imply a hierarchy of subjugation and one lect being considered as an inferior part of the other. The distinction between a language and a dialect is significantly more connected to the sphere of sociolinguistics rather than pure linguistic variation (Otheguy and Stern, 2011). As language classification studies operate in terms of language distance and not language hierarchy, all three lects are studied as equal.

Avoiding the language/dialect distinction, particularly in the study of the Khislavichi lect, is a requisite, as its current research is heavily influenced by extralinguistic factors. By now, it is safe to assume that Khislavichi is an East Slavic lect, phylogenetically more closely connected to standard Belarusian while being heavily influenced by standard Russian as a part of the process of an intense russification (Bączkowski, 1958). This process continues now and affects lesser Russian lects (Daniel et al., 2019), East Slavic lects (Naza-Accent, 2023), and the lects of ethnic groups of diverse origins in Russia (Liberty, 2023). Even the very name Khislavichi is transcribed into English as *Khislavìči* (IPA: [hɪsɫavɪtɕɪ]), with the letters *ch* representing the Russian voiceless alveolo-palatal affricate *č̀* and not the native voiceless postalveolar

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 174–186
May 5, 2023 ©2023 Association for Computational Linguistics

affricate *č* which is more similar to Belarusian.

In this paper, which deals with the issue of linguistic variation and classification in relation to morphological tagging, Khislavichi lect is treated as a separate entity equally connected to standard Belarusian and standard Russian. Therefore, automatic language classification will provide data on its possible grouping with the other two, not its position within the language hierarchy or its inclusion into either standard Russian or standard Belarusian. However, while making conclusions, we should not ignore the historical context, mainly the intensive russification the Khislavichi lect has undergone.

Treating Khislavichi as a separate lect opens the road to its fully independent study, as it will no longer be considered a part of the Russian language (Zaharova and Orlova, 2004), contrary to its placement within the Russian lect in the earlier research. This study requires the development of natural language processing tools (NLP). The first tool in this pipeline is generally a morphological tagger. Morphological tagging is a process (and a product of this process) that includes assigning universal part-of-speech tags and morphological features to the tokens (Toleu et al., 2022). Morphological tagging is employed to get basic information on the grammar structure of the lect under study. Afterwards, it is utilised in both further processing, such as lemmatisation or masked language modelling, and the research of a lect, for instance, in the creation of a lect grammar. Some studies suggest that the results of lect automatic processing may also be used for its classification (Campos et al., 2020b). Morphological tagging was not considered to be the best candidate in comparison with perplexity (Campos et al., 2020a). However, morphological tagging seems to be useful for the preliminary classification that presents general information on the relationship between a small lect and the larger ones that surround it or influence it.

The classification of lects is a process of grouping lects by some meaningful characteristics. Among such characteristics may be the historical differentiation (Gooskens and Heeringa, 2004) or typological similarities (Hammarström and O'Connor, 2013) (McGregor, 2013) (Wälchli and von Waldenfels, 2013). The classification may give some insights into the development of a language or signal of a currently occurring intense language change. In this paper, the suggestion is to classify the lects based on the differences that cause problems in the work of a morphological tagger. The optimal algorithm is a distance-tree matrix (Bapat, 2010). To build a tree from a triangular distance matrix collected from the models accuracy scores we use a statistical method, UPGMA(Sokal and Michener, 1958), implemented via *biopython* package (Cock et al., 2009). It is generally used in evolutionary biology, and probably may be successfully adapted for language study, as the whole idea of tree classification had been (Schleicher, 1863).

The second section is dedicated to the previous research on the topics of morphological tagging, Khislavichi lect studies, and classification methods, including the automatic ones. The third section details the methods of morphological tagging (bi-LSTM neural network) and automatic classification (a distance matrix-based tree) that are going to be implemented in the experiments. The fourth section contains information about the datasets used for training, evaluation, and tests of taggers. The fifth section presents experiments and their analysis, performed on Russian, Belarusian, and Khislavichi material. The conclusion wraps up the research with the final analysis of the morphological tagger prediction efficiency and provides an outline for future research of the Khislavichi lect and the classification methods.

2 Related Work

The variation within the low-resourced closely-related territorially close lects, often joined under the term dialect, has been intensely studied for the last two decades (Nerbonne et al., 1999) (Arhangel'skij, 2021). Different methods have been used to study the lect variation. The most frequent, though heavily criticised (Prokić and Moran, 2013), is the edit distance group of methods (Kosmajac and Keselj, 2020), mainly represented by using Levenshtein distance on a certain list of words (Nerbonne and Heeringa, 1997) (Nerbonne et al., 1999) (Gooskens and Heeringa, 2004), generally the Swadesh list items (Nerbonne and Heeringa, 1997). Recent years, however, witnessed some changes in this situation. The Swadesh list items are no longer the ultimate solution, some other, topic-restricted, wordlists are used (Saxena and Borin, 2013) (Snoek, 2013). The methods changed as well: phonetical approach (Saxena et al., 2022), the perplexity of large language models on the task of masked language modelling (Campos et al., 2020b), sequence alignment approach (List, 2011), linguacultural approach (Lewandowska–Tomaszczyk, 2021), information theory approach (Wettig et al., 2013), and interdisciplinary approach (Carling et al., 2013). Morphological taggers generally were not used to measure language variation, but most were claimed to benefit from it (Magistry et al., 2019). This article inquires about the possible reverse situation when language variation is measured by the results of morphological tagging.

Automatic morphological tagging is an NLP task that has existed for a long time (Spyns, 1996) (Aduriz et al., 1996) (Branco and Silva, 2003) (Berdičevskis et al., 2016) (Sierra Martínez et al., 2018) (Ljubešić and Dobrovoljc, 2019). There are different approaches to it, especially when low-resourced lects are considered. Currently, the dominating approaches are the rule-based, adjusted for the needs of a particular language (Gambäck, 2012), and the more universal one based on recurrent neural networks (Straka et al., 2016). The current shift into the direction of language-independent morphological tagging (Toleu et al., 2022) leads to the development of taggers that can deal with close lects (Obeid et al., 2022), which is an essential problem, for instance, for Arabic (Inoue et al., 2022) (Fashwan and Alansary, 2022). Low-resourced morphological tagging is gaining increasing recognition (ImaniGooghari et al., 2022) (Wiemerslage et al., 2022). Now a lot of attention is paid to the selection of data to train, evaluate, and test a tagger on (Muradoglu and Hulden, 2022). The old models (Qi et al., 2018) (Qi et al., 2020) are adjusted (Scherrer, 2021) to meet the new requirements of efficient training on low-resourced closely-related lects.

Low-resourced closely-related East Slavic lects are currently undergoing extinction (Daniel et al., 2019), with Khislavichi being no exception (Ryko and Spiricheva, 2022). The Khislavichi lect is a lect of the Khislavichi settlement, which is located on the border between Russia and Belarus. It used to be a part of Belarusian territories until the beginning of the XX century, but became a part of Russia in 1924 (Ryko and Spiricheva, 2020). Its study began at the beginning of the XX century when it was characterised as a Northern Belarusian dialect (Karskij, 1903) (Durnovo et al., 1915) (though it is important to state that the Belarusian language itself was then considered a dialect of Russian by the “colonial scientists” from the Russian Empire). Since the 1960s Khislavichi was considered to be a Russian dialect with Belarusian elements becoming less and less prominent (Zaharova and Orlova, 2004). The current consensus is that the Khislavichi lect is a borderline lect between Russian and Belarusian, sharing key features with both these languages (Ryko and Spiricheva, 2022). Yet the Russian features are mostly borrowed or inflicted upon this lect, and Belarusian features form its historical core (Ryko and Spiricheva, 2022). The question of its classification remains uncertain. For a review of the historical and contemporary state of the Khislavichi lect and its overall linguistic description, one should refer to Ryko and Spiricheva (2022).

There are different ways to produce a classification of lects (Holman et al., 2008). The idea of splitting lects into non-hierarchical groups by performing hierarchical clustering became prevalent during the last decade (Buch et al., 2013). The clustering is made automatically, as recent years have witnessed an increasing use of quantitative methods for classification (Pastorelli, 2017) (Mironova, 2018). The algorithms that provide visualisations for the clusters were developed as well (Korkiakangas and Lassila, 2018). Some clustering solutions are distance-based (Rama and Kolachina, 2013).

3 Method

The research is split into three parts. The models for Russian and Belarusian are prepared via training and evaluation on the respective language datasets to get a general picture of their performance. After that, a cross-evaluation is performed to see the ability of both models to generalise based on the knowledge they have previously acquired. Next, the predictions on the Khislavchi lect material are made and evaluated manually. The final stage includes the automatic classification of the lects based on the results of the tagging evaluation.

There are different models, including the multilingual pre-trained ones, that are used for the morphological tagging of heterogeneous lects (Straka and Straková, 2017) (Kondratyuk, 2019) (Kanerva et al., 2021). However, for this experiment, the model for training should be as unaware of the potential structure of Russian, Belarusian, or Khislavichi as possible. At the same time, after training on one lect, it should still not know the others while possessing the ability to generalise its previous findings for their tagging. One more requirement is that the model should preserve some level of consistency while being trained on the corpora that are not significantly low-resourced but yet do not achieve the old national corpus standard of 1 million words. The models stated previously are either universal or underprepared for the low-resource scenario. One possible pick that satisfies all the requirements is the Stanza tagger, developed at Stanford for Universal Dependencies tagging (Qi et al., 2018) (Qi et al., 2020). It was recently modified to use bidirectional character-level LSTM by default, and specifically adjusted to the aims of part-of-speech tagging, the starting point for low-resource NLP (Scherrer, 2021). This fork is used in this paper.

After the model is chosen, the training phase begins. It consists of two subsequent runs of the code, yielding two trained models, one for Russian, and one for Belarusian respectively. These models should satisfy the requirements for overall accuracy, overcoming the basic threshold of 50%, and, optimistically, getting close to the threshold of 85 – 90% overall accuracy. To exclude overfitting, the additional evaluation of the model on the previously chosen part of the dataset is performed. The models are also compared to the previous results of morphological tagging for the Russian and Belarusian languages to check whether their ability to tag is not significantly lower. In the latter case, the shift to the other language model is probably going to be necessary.

When the conditions are met, the models are cross-evaluated. The model that was trained on the Russian material is evaluated on the Belarusian material, and vice versa. This helps to evaluate the ability of both models to generalise before the final run is performed. There can be no expectations at this stage, as both the Russian and the Belarusian models are going to be trained on the monolingual corpora. However, as both lects are East Slavic, the models are probably going to demonstrate at least a 20 to 30 per cent level of accuracy. At this stage, a manual analysis by the researcher must be performed to highlight some common mistakes that can be made by the model that switches from Russian to Belarusian tagging and the other way around as well.

The final run of both models is going to be performed on the Khislavichi material. As it is with cross-evaluation runs, there is no particular threshold that the models are expected to overcome. And for Khislavichi tagging there is an additional obstacle of the gold dataset absence. So, the evaluation is performed only by overall accuracy, and both models are getting the easiest possible treatment: they may not guess all the tags, however, if the tags they assign are correct, they get a point. Their performance, thus, may seem to get significantly boosted, though, in fact, it is going to remain at the same level as in the cross-evaluation stage. The main expectation here is that predictions of both models demonstrate a close accuracy score, as Khislavichi lect is generally supposed to be located just in the middle of the spectre between standard Russian and standard Belarusian. If the expectation is not met, the reasons should be provided. This leads to the analysis of the errors the models make on the Khislavichi material, as well as to possible explanations of the most common mistakes. And if one of the models performs abnormally well (getting close to the monolingual evaluation level of accuracy, or dealing particularly well with some classes of units), or abnormally bad (getting close to the bilingual evaluation level of accuracy, or coping particularly badly with some classes of units) the rationale is also going to be given. If the results of the two models contrast in some meaningful manner, clarification is expected. The analysis should provide recommendations for developing future datasets and taggers for the Khislavichi dataset.

After all the accuracy scores are acquired, the overall analysis for the whole picture of Russian-Belarusian continuum morphological tagging should be provided.

The research finishes with an attempt to automatically classify the three lects of standard Russian, standard Belarusian, and Khislavichi. For this, the standard method of distance-tree matrix (Bapat, 2010) is applied. This method has been previously successfully used in phylogenetic studies in biology (Fitch and Margoliash, 1967) (Gilbert and Parker, 2022) (de Vienne et al., 2011). The borrowing of the automatic classification method from biology is justified, as the evolution of language and the evolution of life share a lot of similarities (Pasquini et al., 2023), which have been highlighted recently (Ladoukakis et al., 2022), and the concept of linguistic phylogeny tree is borrowed from biology (Schleicher, 1863). This may lead to a new potential way of lect classification, a possibility to clusterise lects with the same computational methods as species (Wattel, 1996) (Bapat, 2010). When the classification is performed, the resulting trees are drawn by a Python script that employs these methods. The resulting trees demonstrate the clusterisation of the standard Russian, standard Belarusian, and Khislavichi lects predicted through morphological tagging.

4 Data

Three datasets are employed for the experiments. The first one is the corpus of the Khislavichi lect (Ryko and Spiricheva, 2020). The second one is the Belarusian-HSE corpus, the Belarusian Universal Dependencies one (Shishkina and Lyashevskaya, 2021). The third one is the Taiga corpus, one of the Russian Universal Dependencies datasets (Lyashevskaya et al., 2017) (Shavrina and Shapovalova, 2017).

As the Khislavichi lect and its relationship to standard Russian and standard Belarusian form the centre of the research, the entirety of the currently available data should be investigated. These data in the corpus have been collected and digitised by A. Ryko and M. Spiricheva (Ryko and Spiricheva, 2020). These are transcribed recordings of the interviews with the native speakers of this lect, all born between the late 1920s and the late 1960s.

The Khislavichi data is heterogeneous. Not all texts are presented as transcriptions, most of them are edited into a cross between a transcription and a standard Russian text: only the differentiating lexemes are given in parentheses. For instance, in как (як) 'how', как is a standard Russian form, and як is a Khislavichi form. For some texts, however, transcription is available as well. Additional complications arise from the fact that the lect was under the process of intense russification during the Soviet period, which manifests in the speakers born in the late 1920s and the late 1960s speaking in different manners. The common features persist (such as using č, more similar to Belarusian, and not č̀, more similar to Russian, in words like Хиславичах 'Khislavichi'), however, some radical changes in lexis start manifesting (for example, using лук instead of цыбуля for onion). The texts are interviews, with interviewers speaking in standard Russian.

With these issues in mind, the Khislavichi dataset is additionally preprocessed. The latter issue is resolved by the exclusion of the interviewers' lines from the final dataset. The issue of an in-lect heterogeneity is treated as a matter of fact, no additional splits are performed. Where the transcriptions are available, they are taken. If a pair of standard Russian lexemes and a differentiating Khislavichi lexeme in parentheses is met in the texts made to resemble standard Russian, only the differentiating Khislavichi lexeme is taken. The resulting set of texts is transferred into the CoNLL-U format, which turns it into a corpus of nearly 100 000 tokens. This corpus is split into the training, evaluation, and test datasets (80 000, 10 000, and 10 000 tokens respectively). As the research does not imply training the model for the Khislavichi lect, only the test dataset is going to be used for the later evaluation of the models trained on the Belarusian and Russian material.

The Belarusian-HSE corpus (Shishkina and Lyashevskaya, 2021) is chosen to get the model able to perform morphological tagging for Belarusian. While there are some much larger corpora, for instance, Belarusian N-corpus (N-corpus, 2023), their tagging is not disambiguated and thus is impossible to be used for training. Additionally, these corpora are not in open access. The Belarusian-HSE corpus, in turn, is available in CoNLL-U format from the start and was designed with the tagger training in mind, as this is the requirement for the Universal Dependencies corpora. This corpus consists of different text genres, from the newspapers to the Telegram messages and community posts, so it may safely be called a balanced representation of the modern Belarusian language (Shishkina and Lyashevskaya, 2021). The size of the corpus is 305 000 tokens, which makes it sufficient for the modern taggers to be trained on, especially for the ones that are well-adjusted for the Universal Dependencies low-resourced datasets (Qi et al., 2018). The corpus is split into the training, evaluation, and test parts in 80/10/10% proportion, as is generally the case with the Universal Dependencies datasets. In contrast to the Khislavichi dataset, training and evaluation parts are going to be used in the training phase to get the model for tagging Belarusian texts. The test part is going to be used for the final evaluation of this model, as well as for testing the ability of the model trained for the tagging of Russian texts, which subsequently will aid the automatic classification of standard Russian, standard Belarusian, and Khislavichi lects.

The Russian corpus, in its turn, should meet one, but a very strict condition. It should match the Belarusian-HSE corpus in size, the precision of manual tagging, and adjustment for the taggers that are designed for the data in Universal Dependencies (CoNLL-U) format. Again, there are giant corpora, consisting of billions of words, such as the Russian National Corpus (Corpus, 2023), but they are not in open access, and, what is much more important, their tagging is not fully disambiguated. There is even a big Universal Dependencies corpus, 1.5 million words SynTagRus (Droganova et al., 2018). However, its use is going to give an advantage to the Russian model. It is going to train better on a significantly bigger corpus. So, a smaller corpus should be used. The Taiga corpus (Lyashevskaya et al., 2017) (Shavrina and Shapovalova, 2017), with an overall size of 197 000 tokens, is proposed as a suitable candidate. This corpus is prepared in a Universal Dependencies format and designed specifically for tagging tasks. It is smaller, though not greatly, than the Belarusian-HSE corpus. It is also balanced, and quite representative of modern Russian, containing blog texts, news texts, fiction (including poetry) texts, Wikipedia articles, as well as different texts from social media (Lyashevskaya et al., 2017). It is split into the training (80%), evaluation (10%), and test (10%) parts. As with Belarusian, the model is going to be trained with the use of the training and the evaluation parts of the dataset. After this, the test part of the dataset will be used for the evaluation of the trained model as well as the model trained on the Belarusian dataset, supplying the data for the automatic classification of the lects.

Both Belarusian-HSE and Taiga contain a significant amount of texts from social media, a genre that is as close to the main Khislavichi corpus genres, everyday talks and events retelling. This should eliminate genre elements from affecting accuracy scores of the models.

5 Experiments and Analysis

The starting point is testing the models in the languages they were trained on. Thus, the first experiment includes testing the model that was trained on the standard Russian Taiga corpus on the test subset of this corpus, and testing the model that was trained on the standard Belarusian corpus on the test subset of its corpus.

After that, cross-evaluation is performed: the model that was trained in standard Russian is tested on the test dataset from the Belarusian corpus, and vice versa. The evaluation highlights the main difficulties the models face while tagging a closely-related lect.

Model	PoS + Feats	PoS	UFeats
UD	63.0%	86.4%	69.2%
Stanza(m)	**92.99%**	**96.96%**	**83.81%**

Table 1: Comparison of morphological tagging for Belarusian-HSE by UDPipe (Straka and Straková, 2017) and modified Stanza (Scherrer, 2021). The best results, here and afterwards, are given in **bold**.

In the last experiment, both these models are tested on the restricted test dataset from the Khislavichi lect, with an in-depth analysis of the reasons why each of the models succeeds in tagging of particular language units and fails in others. Some preliminary predictions on how the classification may look like are made at this stage.

The final analysis includes the comparison and the discussion of the experiments results, putting each of them in the general context of the research. Two possible ways for the following automatic classification of lects, based on the morphological tagging evaluation results, are suggested and realised. The *pro et contra* for both of them is given.

5.1 Monolingual Experiments

The first model to train was a Belarusian one. It achieved an almost perfect part-of-speech tagging accuracy of 97% and morphological features tagging accuracy of close to 85%. The results of the model run on the test part of the dataset were compared to UDPipe, a multilingual morphological tagging model presented in Straka and Straková (2017). The comparison is performed by PoS + Feats (both morphological features tagging and part-of-speech match), PoS (part-of-speech match), and UFeats (morphological tagging exact match). The summary of this comparison is presented in Table 1.

A modified Stanza tagger provides a more effective tagging than UDPipe. However, UDPipe is a multilingual model, and this Stanza version was specifically trained for Belarusian, so it is incorrect to make a direct comparison. For this research, it is enough to state that the model is sufficiently trained, and does not overfit.

The next model to train was a Russian one. It achieved the part-of-speech tagging accuracy of 94%, and 76% accuracy for morphological features. The results for the Russian model were compared to the results of the UDPipe model on the Taiga

Model	PoS + Feats	PoS	UFeats
UD	86.4%	75.8%	74.0%
Stanza(m)	**87.98%**	**93.63%**	**76.45%**

Table 2: Comparison of morphological tagging for Russian-Taiga by UDPipe (Straka and Straková, 2017) and modified Stanza (Scherrer, 2021).

Direction	PoS + Feats	PoS	UFeats	OOV
Ru > Bel	56.47%	67.46%	43.83%	73.84%
Bel > Ru	**58.9%**	**68.07%**	**51.63%**	60.8%

Table 3: Comparison of morphological tagging for Belarusian-HSE by the model trained on Taiga (Ru > Bel) and morphological tagging for Taiga by the model trained on Belarusian-HSE (Bel > Ru). The architecture of both models is the modified Stanza (Scherrer, 2021).

corpus. The comparison is given in Table 2.

The modified Stanza tagger again outperformed the UDPipe one (and proved its ability to efficiently operate under the lacking lect resources - both Taiga and Belarusian-HSE are not particularly big corpora), though this time not by a huge margin. This may be due to the fact that the training material for UDPipe run on Russian included more data than the training material for UDPipe run on Belarusian, or to the inner workings of the modified Stanza tagger, which was unable to train on the Taiga corpus. In any case, as this tagger beat the multilingual one, its results for Russian may also be called sufficient for further experiments.

5.2 Cross-evaluation between Russian and Belarusian

The next experiment was the run of the Belarusian-HSE-trained model on the test set of Taiga, and the run of the Taiga-trained model on the test set of Belarusian-HSE. This was conducted to evaluate the generalisation ability of the models and to detect whether some specific factors make cross-prediction between the lects easier or harder. The results are presented in Table 3.

The out-of-vocabulary (OOV) rate of the Belarusian model is smaller than the out-of-vocabulary rate of the Russian model, which is probably due to the Russian influence on Belarusian, and overall heterogeneuity of the Belarusian corpus. The size of the corpus hardly matters: the model, trained on downsampled Belarusian corpus, showed the same results in cross-evaluation experiments. In each possible category of comparison this model is slightly better, which is especially obvious in morphological features tagging. However, its accuracy falls more significantly (for instance, 34.09% against 31.41% in the exact morphological tagging category), which may indicate that it is overfitting for the Belarusian language.

Not all the errors that the Belarusian model makes support this theory. There are some strange ones, like tagging *)* as a punctuation mark and not a symbol when it is used as a smile. This is clearly an annotation schema difference. Sometimes not all glosses are used: for instance, *Tense=Pres* (the one that denotes present tense) is missing from the tagging of verb решается ‘solve-PRES.3SG.REFL’.

However, most errors are connected to the fact that the model was trained on the monolingual dataset. There are cases of words unknown in Belarusian but very frequent in Russian, such as отлично ‘excellently’ consistently tagged as nouns. Words that end with ть are often verbs in Belarusian, yet in Russian, there are words like пусть ‘let it be’, which are not verbs but particles, and this confuses the model. One more type of error that is connected with language interference: the model tags да ‘and’ as a preposition ‘to’, which it is in Belarusian.

The errors that the Russian model makes while tagging the Belarusian dataset are of the similar type. For instance, it incorrectly adds glosses like *NameType=Geo* ‘geographical proper name’ to the words like Еўрасаюза‘European Union.-GEN.SG’. A lot of mistakes are connected to the difference in the alphabets. Thus, Belarusian і and Russian и both mean ‘and’, which are pronounced in basically the same way, but due to the graphic differences і is not tagged as a coordinative conjunction and is tagged as a noun instead.

The errors of the Russian model put the errors of the Belarusian model in context. The bigger vocabulary of the Belarusian model leads to a higher level of language interference in this model, and thus its generalisation ability seemed to be worse than that of the Russian one. In fact, though, the generalisation ability of both the models is not great, both models fail in tagging a completely different lect in comparison to tagging the lect they were trained on. However, they retain a level of accuracy in which it is preferable to use them instead of the

Direction	Accuracy
Ru > Khi	**70.06%**
Bel > Khi	54.75

Table 4: Comparison of morphological tagging for the Khislavichi dataset by the models trained on Taiga (Ru > Khi) and Belarusian-HSE (Bel > Khi). The architecture of both models is the modified Stanza (Scherrer, 2021).

random assignment of parts of speech and morphological features (which would get 40 to 60% accuracy score). It is true for all the cases of the Belarusian model; the Russian model fails in the exact morphological features match task, yet it is often due to the tagset differences. Russian model tends to overtag, assigning more tags than there are in the original dataset. Sometimes it may be even treated as a correct assignment: *PronType=Dem*, demonstrative pronoun tag, for гэта 'that'. Thus, the Russian and the Belarusian models both may be used for preliminary tagging of closely-related East Slavic lects. The Khislavichi lect is a suitable candidate due to its strong connection to both standard Russian and standard Belarusian.

5.3 Evaluation on the Khislavichi Dataset

The last run of the models was conducted on the test part of the Khislavichi dataset, consisting of nearly 8000 tokens.

The issue with the Khislavichi dataset is that there are no gold data for it, and thus the evaluation had to be done manually, which may have led to some errors and inconsistencies, as any kind of human validation is going to. The only criterion of the evaluation was the total accuracy. As these models were previously proven to not perform efficiently on the lects they had not been trained on, the criterion is made to be very soft. It is enough for a model to not make an overt error to score. A model may not guess all the glosses, due to the annotation schema differences, but if all the glosses that model predicted are correct, it scores. The results of the experiments on Belarusian and Russian models are presented in Table 4.

These results seemingly differ from the ones that were acquired previously. Here, the Russian model demonstrates a much higher level of accuracy than the Belarusian. Its score is closer to its part-of-speech score in Belarusian, while the Belarusian model score is closer to its joined part-of-speech and morphological feature tagging score in Russian. What does this mean?

Input / Target lect	Russian	Belarusian	Khislavichi
Russian	0.88		
Belarusian	0.59	0.93	
Khislavichi	0.55	0.7	0

Table 5: Russian-centred distance matrix

Input / Target lect	Belarusian	Russian	Khislavichi
Belarusian	0.93		
Russian	0.57	0.88	
Khislavichi	0.7	0.55	0

Table 6: Belarusian-centred distance matrix

The models do not perform in an unusual way for them, even statistically. The ability of the Russian model to generalise is enough to get 70% of part-of-speech tags correctly in at least some East Slavic lects. The Belarusian model may meet a higher concentration of its *faux amis* in the dataset, as it has a bigger vocabulary. The Russian model, however, also meets a lot of language interference. Thus, both models are confused with the aforementioned word да. In the Khislavichi dataset, it mostly takes an interjection role and the meaning 'yes'. The Russian model treats it as a coordinative conjunction, and the Belarusian one – as a preposition. Both are mistaken.

Both models are sometimes fined due to the features of the Khislavichi dataset. Thus, //, which is meant to be a punctuation mark of a big pause in the speech, is consistently tagged by both models as a symbol. The same may be said about all the fragmentary tokens, denoted with the = sign at the end of the word. They should be tagged as X, non-word, yet the datasets the models were trained on do not possess examples of such cases, so the models fail in tagging these particular units.

Nearly 80% of the Khislavichi dataset is presented not as a transcription, but almost as a translation into the Russian language, so it is significantly easier for the Russian model to tag. When it meets raw transcribed tokens, such as захочыть 'want-FUT.3.SG' (which it tags as an infinitive), it often does not score. In contrast with the Russian model, the Belarusian one, while facing these transcribed tokens of Khislavichi origin, successfully tags them: the transcription often makes

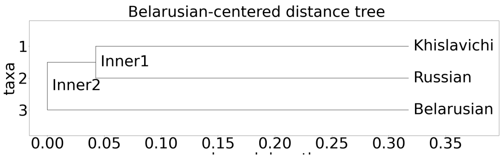

Figure 1: Belarusian-centered distance tree, built with UPGMA (Sokal and Michener, 1958)

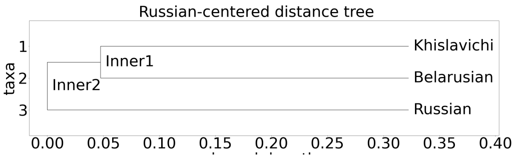

Figure 2: Russian-centered distance tree, built with UPGMA (Sokal and Michener, 1958)

Khislavichi tokens look more similar to Belarusian.

The Belarusian model meets the same kind of issues with the words that contain <w> (the labio-velar approximant transcription, designated as <ў> in Belarusian. It does not correctly tag these words due to the alphabet differences. The same may be said for the items like иГ 'them', which is written as іх and их in Belarusian and Russian respectively.

There are few mistakes connected to the systemic differences between Russian or Belarusian and Khislavichi. They are mostly sparse and hidden by writing system-based errata. There are, however, some common issues. Belarusian-trained model often predicts nouns like боль 'pain' as masculine, which they are in Belarusian. Russian-trained model assigns *Aspect=Perf* to words like ляжыт in the contexts where they are *Aspect=Cont*. This preference probably comes from the perfective contexts being more frequent in Russian.

Contrarily to the first impression, the attempt at tagging Khislavichi did not create an anomaly. There are two key reasons for the Russian-trained model performance boost, and the Belarusian-trained model performance remaining at the same level: the Khislavichi dataset transcription is forced to a form resembling standard Russian, and the sounds that Khislavichi and Belarusian share are transcribed differently in the dataset. It produces a lot of noise that interferes in the actual results.

5.4 Automatic Classification

The experiments results are grouped into the two possible distance matrices, presented in Table 5 and Table 6. The first matrix is centred around the standard Russian model, and the second one - around the standard Belarusian one. The per cent values of accuracy are replaced by a floating-point value. Khislavichi do not have gold data, so the accuracy score of the model, trained on it, is 0.

Using these two matrices, two distance trees are built with UPGMA (Sokal and Michener, 1958). They are presented in figures 1 (Belarusian-centered) and 2 (Russian-centered). The trees are very similar. When the focus is on the failures of a model trained on one of the standard Russian and standard Belarusian lects, it shifts the other one closer to the Khislavichi lect. For the classification to be more precise, grapholinguistic issues should be resolved and an additional model should be trained on the Khislavichi material.

6 Conclusion

The models for morphological tagging of Russian and Belarusian, based on the architecture provided in Scherrer (2021), beat the previous results set by the multilingual models by a significant margin for both Belarusian and Russian. Both models demonstrated the ability to perform a moderately successful tagging of the closely-related lects.

The cross-evaluation results and the results of evaluation on the Khislavichi dataset show that the Russian and Belarusian models may be used for the preliminary tagging of closely related low-resourced East Slavic lects. The classification by the results of morphological tagging retains the same uncertainty level of the Khislavichi lect position among the other East Slavic lects as the classifications reviewed in Ryko and Spiricheva (2022). In the current dataset orthography state, the reasonable conclusion is the Khislavichi dataset being classified as borderline between Russian and Belarusian datasets, not the Khislavichi lect - as borderline between Russian and Belarusian (the same notion for Italian presents Davis (2017)).

We are going to modify and implement the presented automatic classification method for the bigger number of lects, and use the transformed and tagged Khislavichi corpus for the further Khislavichi lect processing. The corpus probably will later become a Universal Dependencies part.

7 Acknowledgements

We thank the anonymous reviewers for their comments on the proposed method and datasets, which greatly aided the final paper. The remaining errata are, of course, ours.

References

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, and Ruben Urizar. 1996. Euslem: A lemmatiser/tagger for basque. In *Proceedings of the 7th EURALEX International Congress*, pages 27–35, Göteborg, Sweden. Novum Grafiska AB.

Timofei A. Arhangel'skij. 2021. Primenenie dialektometricheskogo metoda k klassifikacii udmurtskih dialektov. *Uralo-altajskie issledovaniya*, 2(41):7–20.

Włodzimierz Bączkowski. 1958. *Russian colonialism: the Tsarist and Soviet empires*. Frederick A. Praeger, New York.

R. B. Bapat. 2010. Distance matrix of a tree. In *Graphs and Matrices*, pages 95–109, London. Springer London.

Aleksandrs Berdičevskis, Hanna Eckhoff, and Tatiana Gavrilova. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of middle russian. In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog»*, pages 99–111, Moscow, Russia. RSSU.

Antònio Branco and Jõao Silva. 2003. Portuguese specific issues in the rapid development of state of the art taggers. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, pages 7–9, Paris. European Language Resources Association.

Armin Buch, David Erschler, Gerhard Jäger, and Andrei Lupas. 2013. Towards automated language classification: A clustering approach. In *Approaches to Measuring Linguistic Differences*, pages 303–328, Berlin, Boston. De Gruyter Mouton.

José Campos, Pablo Gamallo, Iñaki Alegria, and Marco Neves. 2020a. A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, 28:1–31.

José Ramom Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. 2020b. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering*, 26(4):433–454.

Gerd Carling, Love Eriksen, Arthur Holmer, and Joost van de Weijer. 2013. Contrasting linguistics and archaeology in the matrix model: Gis and cluster analysis of the arawakan languages. In *Approaches to Measuring Linguistic Differences*, pages 29–56, Berlin, Boston. De Gruyter Mouton.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

The Russian National Corpus. 2023. The russian national corpus.

Michael Daniel, Ruprecht von Waldenfels, Aleksandra Ter-Avanesova, Polina Kazakova, Ilya Schurov, Ekaterina Gerasimenko, Daria Ignatenko, Ekaterina Makhlina, Maria Tsfasman, Samira Verhees, and et al. 2019. Dialect loss in the russian north: Modeling change across variables. *Language Variation and Change*, 31(3):353–376.

Joseph Davis. 2017. The semantic difference between Italian vi and ci. *Lingua*, 200:107–121.

Damien M. de Vienne, Gabriela Aguileta, and Sébastien Ollier. 2011. Euclidean Nature of Phylogenetic Distance Matrices. *Systematic Biology*, 60(6):826–832.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, 155, pages 52–65. Linköping University Electronic Press.

Nikolaj N. Durnovo, Nikolaj N. Sokolov, and Dmitrij N. Ushakov. 1915. *Opyt dialektologicheskoj karty russkogo jazyka v Evrope s prilozheniem Ocherka russkoj dialektologii*. Sinodal'naja tipografija, Moscow.

Amany Fashwan and Sameh Alansary. 2022. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for Egyptian Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 142–160, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Walter M. Fitch and Emanuel Margoliash. 1967. Construction of phylogenetic trees. *Science*, 155(3760):279–284.

Björn Gambäck. 2012. Tagging and verifying an amharic news corpus. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, pages 79–84, Paris. European Language Resources Association.

Gregory S. Gilbert and Ingrid M. Parker. 2022. Phylogenetic distance metrics for studies of focal species in communities: Quantiles and cumulative curves. *Diversity*, 14(7).

Charlotte Gooskens and Wilbert Heeringa. 2004. Perceptive evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change*, 16:189 – 207.

Harald Hammarström and Loretta O'Connor. 2013. Dependency-sensitive typological distance. In *Approaches to Measuring Linguistic Differences*, pages 329–352, Berlin, Boston. De Gruyter Mouton.

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1577–1589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Yefim F. Karskij. 1903. *Belorusy. Vol. I. Vvedenie v izuchenie jazyka i narodnoj slovesnosti*. Warsaw Academic County Publishing, Warsaw.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Timo Korkiakangas and Matti Lassila. 2018. Visualizing linguistic variation in a network of Latin documents and scribes. *Journal of Data Mining & Digital Humanities*, Numéro spécial sur le traitement assisté par ordinateur de l'intertextualité dans les langues anciennes.

Dijana Kosmajac and Vlado Keselj. 2020. Language distance using common n-grams approach. In *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE.

Manolis Ladoukakis, Dimitris Michelioudakis, and Elena Anagnostopoulou. 2022. Toward an evolutionary framework for language variation and change. *BioEssays*, 44:210–216.

Barbara Lewandowska–Tomaszczyk. 2021. Comparing languages and cultures: Parametrization of analytic criteria. *Russian Journal of Linguistics*, 25(2):343–368.

Radio Free Europe/Radio Liberty. 2023. In Russia, there are 40% fewer Komi-Permyaks, 35% fewer Mordovians (Erzya and Moksha), 30% fewer Udmurts, and more than 20% fewer Chuvash and Mari. *Radio Free Europe/Radio Liberty*.

Johann-Mattis List. 2011. Multiple sequence alignment in historical linguistics: A sound class based approach. In *Proceedings of ConSOLE XIX*, page 241–260.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Olga Lyashevskaya, Victor Bocharov, Alexey Sorokin, Tatiana Shavrina, Dmitry Granovsky, and Svetlana Alexeeva. 2017. Text collections for evaluation of Russian morphological taggers. *Journal of Linguistics/Jazykovedný casopis*, 68:258–267.

Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset. 2019. Exploiting languages proximity for part-of-speech tagging of three French regional languages. *Language Resources and Evaluation*, 53:865–888.

William B. McGregor. 2013. Comparing linguistic systems of categorisation. In *Approaches to Measuring Linguistic Differences*, pages 387–428, Berlin, Boston. De Gruyter Mouton.

Dina M. Mironova. 2018. *Avtomatizirovannaja klassifikacija drevnih rukopisej : na materiale 525 spiskov slavjanskogo Evangelija ot Matfeja XI - XVI vv. : avtoreferat dis. ... kandidata filologicheskih nauk : 10.02.21*. SPbU, Saint-Petersbourg.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Belarusian N-corpus. 2023. Belarusian N-corpus.

NazaAccent. 2023. Rosstat: From 2010 to 2021, There are One Million Fewer Ukrainians in Russia. *NazaAccent*.

J. Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonolby.*, pages 11 – 18. Association for Computational Linguistics (ACL).

John Nerbonne, Wilbert Heeringa, and Peter Kleiwig. 1999. Edit distance and dialect proximity. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd Ed.*, pages i–xviii. CSLI, Stanford, CA.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

Ricardo Otheguy and Nancy Stern. 2011. On so-called Spanglish. *International Journal of Bilingualism - INT J BILING*, 15:85–100.

Michele Pasquini, Maurizio Serva, and Davide Vergni. 2023. Gradual modifications and abrupt replacements: Two stochastic lexical ingredients of language evolution. *Computational Linguistics*, pages 1–23.

David Pastorelli. 2017. A Classification of Manuscripts Based on A New Quantitative Method. The Old Latin Witnesses of John's Gospel as Text Case. *Journal of Data Mining & Digital Humanities*, Numéro spécial sur le traitement assisté par ordinateur de l'intertextualité dans les langues anciennes.

Jelena Prokić and Steven Moran. 2013. Black box approaches to genealogical classification and their shortcomings. In *Approaches to Measuring Linguistic Differences*, pages 429–446, Berlin, Boston. De Gruyter Mouton.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Taraka Rama and Sudheer Kolachina. 2013. Distance-based phylogenetic inference algorithms in the subgrouping of dravidian languages. In *Approaches to Measuring Linguistic Differences*, pages 141–174, Berlin, Boston. De Gruyter Mouton.

Anastasiia Ryko and Margarita V. Spiricheva. 2020. *Corpus of the Russian dialect spoken in Khislavichi district*. Linguistic Convergence Laboratory, HSE University, Moscow.

Anastasiia Ryko and Margarita V. Spiricheva. 2022. The degree of preservation of dialectal features in different generations (khislavichi district of the smolensk region). *RSUH/RGGU Bulletin. "Literary Theory. Linguistics. Cultural Studies" Series*, 5:121–141.

Anju Saxena and Lars Borin. 2013. Carving tibeto-kanauri by its joints: Using basic vocabulary lists for genetic grouping of languages. In *Approaches to Measuring Linguistic Differences*, pages 175–198, Berlin, Boston. De Gruyter Mouton.

Anju Saxena, Anna Sjöberg, Padam Sagar, and Lars Borin. 2022. 4 linguistic variation: a challenge for describing the phonology of kanashi. In *Synchronic and Diachronic Aspects of Kanashi*, pages 131–144, Berlin, Boston. De Gruyter Mouton.

Yves Scherrer. 2021. Adaptation of morphosyntactic taggers. In *Similar Languages, Varieties, and Dialects: A Computational Perspective*, Studies in Natural Language Processing, page 138–166. Cambridge University Press.

August Schleicher. 1863. *Die Darwinsche Theorie und die Sprachwissenschaft – offenes Sendschreiben an Herrn Dr. Ernst Haeckel*. H. Boehlau, Weimar.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In *Proceedings of the International Conference "CORPORA 2017", Saint-Petersbourg, Russia*, pages 78–84. Linköping University Electronic Press.

Yana Shishkina and Olga Lyashevskaya. 2021. Sculpting enhanced dependencies for Belarusian. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, page 137–147, Berlin, Heidelberg. Springer-Verlag.

Luz Sierra Martínez, Carlos Cobos, Juan Corrales, Tulio Curieux, Enrique Herrera-Viedma, and Diego Peluffo-Ordóñez. 2018. Building a Nasa Yuwe language corpus and tagging with a metaheuristic approach. *Computacion y Sistemas*, 22:881–894.

Conor Snoek. 2013. Using semantically restricted wordlists to investigate relationships among Athapaskan languages. In *Approaches to Measuring Linguistic Differences*, pages 231–248, Berlin, Boston. De Gruyter Mouton.

R. R. Sokal and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Peter Spyns. 1996. A tagger/lemmatiser for Dutch medical language. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, page 1147–1150, USA. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Alymzhan Toleu, Gulmira Tolegen, and Rustam Mussabayev. 2022. Language-independent approach for morphological disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5288–5297, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Evert Wattel. 1996. Clustering stemmatological trees. In M. van Mulken P. van Reenen, editor, *Studies in Stemmatology*, page 123 – 134. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Hannes Wettig, Javad Nouri, Kirill Reshetnikov, and Roman Yangarber. 2013. Information-theoretic modeling of etymological sound change. In *Approaches to Measuring Linguistic Differences*, pages 507–532, Berlin, Boston. De Gruyter Mouton.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Bernhard Wälchli and Ruprecht von Waldenfels. 2013. Measuring morphosemantic language distance in parallel texts. In *Approaches to Measuring Linguistic Differences*, pages 475–506, Berlin, Boston. De Gruyter Mouton.

Kapitolina F. Zaharova and Varvara G. Orlova. 2004. *Dialektnoe chlenenie russkogo yazyka*. URSS, Moscow.