Multi-source morphosyntactic tagging for Spoken Rusyn

Yves Scherrer
Department of Linguistics
University of Geneva
Switzerland
yves.scherrer@unige.ch

Achim Rabus
Department of Slavonic Studies
University of Freiburg
Germany
achim.rabus@
slavistik.uni-freiburg.de

Abstract

This paper deals with the development of morphosyntactic taggers for spoken varieties of the Slavic minority language Rusyn. As neither annotated corpora nor parallel corpora are electronically available for Rusyn, we propose to combine existing resources from the etymologically close Slavic languages Russian, Ukrainian, Slovak, and Polish and adapt them to Rusyn. Using MarMoT as tagging toolkit, we show that a tagger trained on a balanced set of the four source languages outperforms single language taggers by about 9%, and that additional automatically induced morphosyntactic lexicons lead to further improvements. The best observed accuracies for Rusyn are 82.4% for part-of-speech tagging and 75.5% for full morphological tagging.

1 Introduction

This paper addresses the development of morphosyntactic taggers for spoken varieties of the Slavic minority language Rusyn by leveraging the resources available for the neighboring, etymologically related languages. Due to the lack of annotated and parallel Rusyn data, we propose to create Rusyn taggers by combining training data from related resource-richer languages such as Ukrainian, Polish, Slovak and Russian.

We start by giving a brief introduction to the characteristics of Rusyn and present related work in the domain of low-resource language tagging. After describing the training and test data, we present a set of experiments on different multi-source tagging approaches. In particular, we investigate the impact of majority voting, Brown clustering, training corpus adaptation, and the addition of automatically induced morphosyntactic lexicons. Finally, we give an outlook on future work.

2 Status of Rusyn and corpus data

Rusyn is a Slavic linguistic variety spoken predominantly in Transcarpathian Ukraine, Eastern Slovakia, and Southeastern Poland, and is linguistically close to the Ukrainian language. Its sociolinguistic status is disputed insofar as some scholars see Rusyn as a dialect of Ukrainian, others claim it to be an independent – the fourth East Slavic – language. Despite its closeness to Ukrainian, Rusyn exhibits numerous distinct features on all linguistic levels, which make Rusyn look more "West Slavic" as compared to Ukrainian.[1]

Nowadays, most speakers of Rusyn are bilingual and have native-like command of, e.g., Polish or Slovak. This has an impact on their Rusyn speech and leads to new divergences within the old Rusyn dialect continuum, which can be investigated using the Corpus of Spoken Rusyn (www.russinisch.uni-freiburg.de/corpus) that is currently in the process of being built up. The corpus comprises several hours of transcribed Rusyn speech from the different countries where Rusyn is spoken. This means that both diatopic and individual speaker variation is reflected in the transcription, which is one reason for the fact that the corpus data is orthographically (and morphologically) heterogeneous. Another reason is that variation in transcription practices due to several individual transcribers could not completely be avoided.

The goal of the research presented here is to automatically provide morphosyntactic annotations

[1] For further details on the status and the features of Rusyn see, e.g., Magocsi (2004), Plishkova (2009), Pugh (2009), Skrypnyk (2013), Teutsch (2001).

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 84–92,
Valencia, Spain, April 3, 2017.

PL	Na początku było Słowo a Słowo było u Boga, i Bogiem było Słowo.
Cyrillicized	На почутку было Слово а Слово было у Бога, и Богем было Слово.
RU	В начале было Слово, и Слово было у Бога, и Слово было Богом.
SK	Na počiatku bolo Slovo a Slovo bolo u Boha a Boh bol to Slovo.
Cyrillicized	На почиатку боло Слово а Слово боло у Бога а Бог бол то Слово.
UK	На початку було Слово, а Слово в Бога було, і Бог було Слово.
RUE	На початку было Слово, а Слово было у Бога, і Бог было Слово.

Figure 1: John 1:1 in the Slavic languages used for the experiments.

for the Corpus of Spoken Rusyn. However, there are virtually no NLP resources (annotated corpora or tools) available for Rusyn at the moment. The different types of variation present in the data complicate the task of developing NLP tools even more. Crucially, there is no parallel corpus available for Rusyn, which means that the popular projection-based approaches cannot be applied (see below).

Considering the lack of annotated Rusyn data and the etymological situation of Rusyn, our approach consists in training taggers for several related languages – namely, the East Slavic languages Ukrainian and Russian and the West Slavic languages Polish and Slovak – and combining and adapting them to Rusyn. This multi-source setting makes sense, because the Rusyn dialect continuum features both West Slavic and East Slavic linguistic traits to a different extent, depending on both the dialect region and the impact of the respective umbrella language. In order to get an idea of the similarities and differences of the Slavic languages involved, compare the different versions of John 1:1 in Figure 1.

3 Related work

The task of creating taggers for languages lacking manually annotated training corpora has inspired a lot of recent research. The most popular line of work, initiated by Yarowsky and Ngai (2001), draws on parallel corpora. They annotate the source side of a parallel corpus with an existing tagger, and then project the tags along the word alignment links onto the target side of the parallel corpus. A new tagger is then trained on the target side, with some smoothing to reduce the noise caused by alignment errors. Follow-up work has focused on the inclusion of several source languages (Fossum and Abney, 2005), more accurate projection algorithms (Das and Petrov, 2011; Duong et al., 2013), the integration of external lexicon sources (Li et al., 2012; Täckström et al., 2013), the extension from part-of-speech tagging to full morphological tagging (Buys and Botha, 2016), and the investigation of truly low-resource settings by resorting to Bible translations (Agić et al., 2015). A related approach (Aepli et al., 2014) uses majority voting to disambiguate tags proposed by several source languages. However, these projection approaches are not adapted to our setting as no parallel corpora – not even the Bible[2] – are electronically available for Rusyn.

Another approach consists in training a model for one language and applying it to another, closely related language. In this process, the model is trained not to focus on the exact shape of the words, but on more generic, language-independent cues, such as part-of-speech tags for parsing (Zeman and Resnik, 2008), or word clusters for part-of-speech tagging (Kozhevnikov and Titov, 2014). A related idea consists in translating the words of the model to the target language, either using a hand-written morphological analyzer and a list of cognate word pairs (Feldman et al., 2006), or using bilingual dictionaries extracted from parallel corpora (Zeman and Resnik, 2008) or induced from monolingual corpora (Scherrer, 2014).

Our work mostly follows the second approach: we train taggers on four resource-rich Slavic languages and adapt them to Rusyn using a variety of techniques.

4 Training data

While morphosyntactically annotated corpora exist for all four source languages, e.g. in the form

[2]We did not find any Rusyn material in the sources given by Christodouloupoulos and Steedman (2015) and Mayer and Cysouw (2014). The sentence cited in Figure 1 has been taken from the printed edition Krajnjak and Kudzej (transl.) (2009).

ID	Origin		Training data			Test data		
			Sentences	Tokens	Tags	Sentences	Tokens	Tags
PL	UD 1.4 Polish	train / dev	6 800	69 499	920	700	6 887	448
RU1	UD 1.4 Russian	train / dev	4 029	79 772	704	502	10 044	410
RU2	UD 1.4 SynTagRus	train / dev	48 171	850 689	580	6 250	109 694	501
SK	UD 1.4 Slovak	train / dev	8 483	80 575	657	1 060	12 440	426
UK	UD 1.4 Ukrainian	train / dev+test	200	1 281	1 040	55	395	92
	Additional data		3 962	70 299				
RUE1	Manually annotated gold standard					104	1 050	96
RUE2	Corpus of Spoken Rusyn					5 922	75 201	—

Table 1: Sizes of the training and test corpora used in our experiments.

of national corpora,[3] they use disparate tagsets and are often difficult to obtain in full-text format. The MULTEXT-East project (Erjavec et al., 2010; Erjavec, 2012)[4] provides annotated versions of the novel *1984* for several Eastern European languages, but Ukrainian and Russian versions are not available.

Fortunately, since version 1.4, the Universal Dependencies project[5] contains treebanks for the four relevant languages with unified part-of-speech tags and morphosyntactic descriptions (Nivre et al., 2016; Zeman, 2015). Two corpora are available for Russian, but the Ukrainian corpus is still rather small (see Table 1). Additionally, we were able to obtain more Ukrainian data developed by the non-governmental Institute of Ukrainian[6] and planned to be included in one of the upcoming Universal Dependencies releases; we converted these additional data from the MultextEast-style tags to universal tags and morphological features.

As Rusyn is written in Cyrillic script, we converted the Slovak and Polish corpora into Cyrillic script. During this process, we applied certain transformation rules in order to "rusynify" our training data (e.g., transform Polish *ć* to Cyrillic *мь* or Polish *ą* to Cyrillic *у*, which is in line with well-known historical phonological processes).

Initial experiments have shown that additional morphological dictionaries, such as those made available for the four languages within the MULTEXT-East project, do not have a positive impact on Rusyn tagging. We therefore do not include these additional resources (except for the derived lexicons discussed in Section 5.5).

We evaluate our methods on a small hand-annotated sample of Rusyn containing 104 sentences and 1 050 tokens and 96 distinct tags (henceforth RUE1). At the time of conducting the experiments, the Corpus of Spoken Rusyn (RUE2), which we aim to annotate with the presented methods, contains 5 922 sentences with 75 201 tokens. We also report OOV rates on the latter and use it as additional unlabeled data for some of the adaptation processes described below.

5 Experiments

5.1 The MarMoT tagger

We use the MarMoT tagger for all of our experiments. MarMoT (Mueller et al., 2013) is a state-of-the-art toolkit for morphological tagging based on Conditional Random Fields (CRFs). It has been shown to work well on full morphological tagging with hundreds of tags (as opposed to part-of-speech tagging, which typically only uses a few dozen tags), thanks to pruning and coarse-to-fine decoding. Unless stated otherwise, we use the default parameters for morphological tagging.

We evaluate the different models on the development sets of the five source corpora as well as on RUE1. A token is considered correctly tagged if its part-of-speech tag is correct and if all morphological features present in the gold annotation are found with the same value.[7]

[3] Ukrainian National Corpus: `www.mova.info`; Russian National Corpus: `www.ruscorpora.ru`; Polish National Corpus: `www.nkjp.pl/`; Slovak National Corpus: `http://korpus.juls.savba.sk/index_en.html`.

[4] `http://hdl.handle.net/11356/1043`

[5] `www.universaldependencies.org`

[6] `https://mova.institute`

[7] The gold annotation of RUE1 does not distinguish proper from common nouns, auxiliary from main verbs, and coordinating from subordinating conjunctions; these mismatches were not penalized.

	Accuracy (%)						OOV rate (%)						
	PL	SK	UK	RU1	RU2	RUE1	PL	SK	UK	RU1	RU2	RUE1	RUE2
PL	**85.87**	49.08	39.2	40.47	43.15	49.5 ±1.0	**20.02**	60.61	59.0	65.56	60.87	50.5	46.00
SK	46.77	**79.87**	37.2	41.94	45.54	43.3 ±0.4	58.05	**33.87**	57.7	63.72	58.56	53.1	43.73
UK	38.25	35.71	**79.8**	41.24	44.81	**63.4** ±0.4	63.13	67.98	**15.4**	69.11	66.07	**37.1**	**39.67**
RU1	39.19	42.60	36.5	**85.73**	79.39	46.0 ±0.6	64.93	65.14	62.0	24.53	27.51	54.1	46.58
RU2	40.79	46.33	40.8	80.68	**93.79**	50.9 ±0.0	59.36	60.35	55.7	**19.73**	**7.98**	49.1	42.72

Table 2: Tagging accuracies and OOV rates for single-language taggers. Rows represent models, columns represent test sets.

5.2 Single-language taggers

We start by training five distinct taggers on the five training corpora and apply these taggers to the five source-language test corpora as well as to the Rusyn corpora. The results are shown in Table 2.

Unsurprisingly, each test set is best tagged with the tagger based on its own training set. Polish and Russian fared somewhat better than Slovak and Ukrainian. The differences between RU1 and RU2 give an indication of the loss resulting from annotation/conversion differences as well as domain differences within the same language. For Rusyn, the best accuracy is obtained using the Ukrainian tagger, which is in line with the claims on linguistic proximity made above, followed by RU2, which is due to its large size rather than to small etymological distance. Also note that for none of the models, Rusyn is the worst-performing test language, hinting at its role as a bridge language between East and West Slavic.

In order to quantify the reliability of the Rusyn tagging results given the somewhat small test corpus, we split it into two equally-sized parts and computed the accuracies on both parts. The deviation of the accuracy values of these parts from the mean accuracy is indicated after the ± sign in Table 2.

While no single-language tagger achieves satisfactory accuracy on Rusyn, the results suggest that a combination of the five taggers (or of their training data) could yield improved accuracy on Rusyn. There are essentially two ways of combining taggers: using the five source language taggers and choosing the majority vote, or using a single tagger trained on merged data from the five source corpora.

5.3 Majority-vote tagging

Aepli et al. (2014) develop a tagger for Macedonian by transferring morphosyntactic annotations from multiple source languages by word alignment, choosing one annotation by majority vote, and training a new tagger on the annotated corpus. We follow a similar method. We start by annotating the Rusyn data with the five source language taggers. A majority annotation is determined in two steps: first, the majority part-of-speech tag is determined, and second, the majority morphological features are determined on the basis of the taggers that have predicted the majority part-of-speech tag. We propose two ways of dealing with ties: we either randomly resolve ties (*Random*) or weight the tags on the basis of *a priori* knowledge about the etymological distances of the languages (*Weighted*).[8]

We report results on this direct annotation (see Table 3, rows MAJ-D), but also use the annotated RUE2 corpus to retrain a new tagger (see Table 3, rows MAJ-R). Only the weighted method yields similar tagging accuracies as the best single-language tagger. The impact of retraining is negative, probably due to the fact that the OOV rate on RUE1 hardly decreases. While we could have tuned the weights of the majority-vote models to further improve their accuracy, this option did not look worthwhile in the light of the better results obtained with the approaches discussed below.

5.4 Creating multi-source taggers

For the multi-source tagger, we concatenate the five training sets, using only the first 10% of RU2 in order to keep the distribution better balanced. As shown in Table 3 (row MS), this simple combination of training resources yields better accuracy than all majority-vote systems and outperforms the best single-language model (UK) by nearly 9%, although with a high variance between the two parts of the corpus. If only parts-of-speech are eval-

[8]The following weights are used: PL: 1.5, SK: 3, UK: 4, RU1: 1, RU2: 1.

	Accuracy (%)						OOV rate (%)						
	PL	SK	UK	RU1	RU2	RUE1	PL	SK	UK	RU1	RU2	RUE1	RUE2
Majority-vote – direct annotation (R=random, W=weighted):													
MAJ-D-R	55.35	59.13	46.3	70.31	75.34	54.9 ±0.7	18.08	28.17	11.9	13.37	7.33	24.9	23.83
MAJ-D-W	51.91	57.55	64.1	49.93	55.12	63.4 ±1.3							
Majority-vote – after retraining (R=random, W=weighted):													
MAJ-R-R	47.38	45.82	42.0	48.30	52.37	54.7 ±0.3	55.34	61.45	31.7	63.54	58.34	23.5	0.00
MAJ-R-W	44.62	43.36	57.2	41.29	46.07	63.0 ±1.2							
Multi-source tagger (B=with Brown clusters):													
MS	84.23	**79.61**	81.5	85.91	88.00	72.0 ±1.3	18.66	29.08	13.2	20.17	16.40	26.4	24.99
MS-B	84.07	79.32	**83.3**	86.44	**88.31**	72.3 ±2.0							
Taggers with additional lexicons (R=rules, L=Levenshtein):													
LEX-R	83.72	79.34	81.8	86.03	88.06	73.9 ±0.1	18.51	28.82	11.7	20.03	16.31	9.6	7.94
LEX-L	83.65	79.54	82.0	86.25	88.04	**75.5** ±0.0						1.1	1.01
Taggers trained on adapted corpora (R=rules, L=Levenshtein, B=with Brown clusters):													
COR-R	83.04	78.30	80.3	85.16	86.68	71.3 ±0.6	20.75	31.54	14.2	22.88	19.81	23.2	22.04
COR-L	80.83	77.59	79.2	84.01	85.71	70.6 ±0.8	26.32	34.22	19.2	26.15	22.69	12.8	12.83
COR-L-B	**84.27**	78.79	82.3	**86.53**	88.30	73.0 ±0.9							

Table 3: Tagging accuracies and OOV rates for the multi-source tagging experiments.

uated, the multi-source tagger achieves 79.2% of accuracy, compared to 69.7% for the best single-language model (UK).

Following e.g. Owoputi et al. (2013), we include word clusters as an additional feature for tagging. We obtain hierarchical word clusters (c=1 000) with the Brown clustering algorithm (Brown et al., 1992) on the concatenation of all source language and Rusyn texts (1.5M running tokens), and add the clusters as an additional feature to the tagger. This addition yields small improvements for some source languages and for Rusyn (see Table 3, row MS-B), although the latter impact is inconclusive due to the high variance between the two corpus parts. We observe that all word clusters spread over words from more than one language, suggesting that the clustering algorithm generalizes well over data from different languages. While larger amounts of unlabeled data will undoubtedly further increase source language tagging, it is less clear whether this will also have a positive impact on Rusyn tagging. In any case, larger Rusyn corpora will be hard to come by.

The idea behind tagger combination was that a lot of Rusyn words can be found in one of the source languages. This has been confirmed, as the OOV rates of the combined taggers (around 24% for Rusyn, see Table 3, rows MAJ-D and MS) are much lower than those of the single language taggers (between 37% and 54% for Rusyn, see Table 2). However, we assume that even more Rusyn words could be found in a source language if some transformations were applied. In the following two subsections, we investigate two different approaches.

5.5 Adding automatically induced lexicons

In Rabus and Scherrer (2017), we describe the automatic induction of morphosyntactic lexicons for Rusyn. In a nutshell, we match Rusyn words extracted from RUE1 and RUE2 with source language words extracted from the Polish, Slovak, Ukrainian and Russian MULTEXT-East lexicons as well as the morphological dictionary of UGtag[9] (Kotsyba et al., 2011), using vowel-sensitive Levenshtein distance, hand-written rules, and a combination of both. The Rusyn words are then associated with the morphosyntactic descriptions of the matched source-language words. The resulting lexicon contains 51 600 token-tag tuples when induced with Levenshtein distance, and 28 900 tuples when induced with rules.

Table 3 (rows LEX-R and LEX-L) reports tagging results, where one of the induced lexicons is added to the multi-source tagger. As expected, the OOV rates drop considerably.[10] Both the

[9] UGtag is a tagger specifically developed for Ukrainian, but essentially consists of a large morphological dictionary and a simple disambiguation component.

[10] OOV rates do not completely drop to 0 because the induction methods failed to find correspondences for a few Rusyn

rule-induced and the Levenshtein-induced lexicon improve accuracy, the latter by 3.5% to 75.5%, the best observed result. Moreover, these results are stable between the two parts of the RUE1 corpus, with only 0.2% difference for the rule-induced lexicon and less than 0.1% difference for the Levenshtein-induced lexicon. If evaluated on the parts-of-speech only, the accuracies increase from 79.2% to 81.3% for the rule-induced lexicon and to 82.4% for the Levenshtein-induced lexicon. Combinations of rule-induction and Levenshtein-induction do not lead to further tagging improvements with respect to the Levenshtein model.

5.6 Adapting the corpora to Rusyn

An alternative to adding Rusyn data in the form of lexicons is to modify the source language training corpora directly by making them look more Rusyn-like. The idea behind this method is to provide the tagger with additional Rusyn tokens in sentential context. We proceed as follows: for each source language word, we search for the most similar Rusyn word in the RUE1 and RUE2 corpora, again using Levenshtein distance or the hand-written rules. If the most similar Rusyn word is different from the source word, we replace the source word with the former.[11]

As the number of known Rusyn words is small in comparison with the number of source words, there is a risk of replacing a source word by a non-related Rusyn word because the related one simply is not known. In this case, we prevent the replacement whenever another source word is closer to the Rusyn candidate. For example, the word *презыдент* in the Polish corpus (converted from *prezydent* 'president') would be replaced by the most similar Rusyn word, which happens to be the word *презенті* but which is unrelated. This replacement is blocked because another Polish word, *презенты* (< *prezenty* 'gifts'), is even closer to *презенті*. When more than one Rusyn word exists with the same distance, no replacement takes place. This phenomenon mostly occurs with Levenshtein distance, where 3-5% of tokens are concerned, but more rarely with the rules, where 1-3% of tokens are concerned. In the end, between 8% and 12% of source tokens are replaced with Levenshtein, and between 1% and 5% of source tokens with the rules.

The results presented in Table 3 (rows COR-R and COR-L) show that these conversions slightly decrease tagging accuracy for the source languages (which is expected, as training corpora now look less like the source languages), but do not improve the accuracy for Rusyn either compared to the simple multi-source model. We also reran the word clustering tool on the Levenshtein-converted data, under the assumption that the increased frequency of the Rusyn words would improve the reliability of the induced clustering. This assumption was indeed borne out with an accuracy increase of 2.4% absolute (row COR-L-B). However, this result did not surpass the one obtained with induced lexicons.

6 Conclusion and future work

We have investigated several approaches to morphosyntactic tagging of spoken Rusyn without relying on annotated Rusyn training data nor on annotation projection from aligned parallel data. Instead, we argued that fair tagging accuracy could be achieved by training taggers on the etymologically related languages Ukrainian, Slovak, Polish and Russian. The experiments also showed that although Ukrainian is most closely related to Rusyn, all four related languages are useful for tagging. We have shown that a multi-source tagger trained on a balanced set of source language corpora performs rather well and even outperforms majority vote approaches. In contrast, Brown clustering has only been modestly useful in our setting, which may be due to the low amount of unlabeled data used.

We have presented two additional techniques to adapt the taggers to the specificities of Rusyn: adding automatically induced morphosyntactic lexicons, or adapting the training corpora. We oriented the first technique towards maximising recall (e.g., keeping all possible readings of a Rusyn word in the induced lexicons) and the second towards high precision (e.g., only replacing unambiguous words in the corpus). The first approach turned out to be more successful.

However, we believe that further improvements can be achieved. First, the RUE1 corpus – currently our only gold standard – is not completely representative of the material found in RUE2. In fact, the RUE1 test set may actually underesti-

words.

[11] For relative Levenshtein distance, we introduce a threshold at 0.25 – as already in the lexicon induction experiments – above which word matches are considered noise and are discarded.

mate the impact of the tagger adaptation methods, as it contains only Rusyn varieties spoken in Ukraine, with a low amount of orthographic variation, whereas RUE2 also contains Rusyn from Poland and Slovakia. As an illustration, compare the OOV rates of the UK tagger (Table 2), which is 2.5% higher in RUE2 than in RUE1. A cursory evaluation of the results confirms this hypothesis, but we cannot quantify it at the moment. Only the manual annotation of a balanced subset of the different RUE2 parts would provide us with a broader data basis for evaluation.

Second, it is crucial to keep in mind that both RUE1 and RUE2 – as opposed to the training corpora – are oral corpora with distinct features such as corrections, repetitions, incomplete sentences, unintelligible words or phrases, markers for pauses, etc. Any tagger trained on written data and applied to oral data will inevitably perform worse than when applied to written data (Nivre and Grönqvist, 2001; Westpfahl, 2014).

The final annotation of the Rusyn corpus is not only expected to consist of morphosyntactic descriptions, but also of lemmas. Therefore, we intend to train a separate lemmatization model on the tagged Rusyn corpora. The multi-source approach will be more problematic here, as we do not want the predicted lemmas to be a mix of the four source languages. The prediction of Rusyn lemmas is prevented by two factors: none of our Rusyn data are annotated with Rusyn lemmas, and the orthographic variation would also carry over to the lemmas, which we would like to avoid. Therefore, one goal could be to annotate the Rusyn tokens with Ukrainian lemmas such as those available in the UGtag lexicon.

Finally, all source language corpora used in our experiments are annotated with syntactic dependencies. We assume that a Rusyn dependency parser could be created using similar methods as those presented here for morphosyntactic tagging.

Acknowledgments

We would like to thank Christine Grillborzer, Natalia Kotsyba, Bohdan Moskalevskyi, Andrianna Schimon, Peter Schwarz, and Ruprecht von Waldenfels. The usual disclaimers apply.

Sources of external funding for our research include the German Research Foundation (DFG).

References

Noëmi Aepli, Ruprecht von Waldenfels, and Tanja Samardžić. 2014. Part-of-speech tag disambiguation by cross-linguistic majority vote. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 76–84, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China, July. Association for Computational Linguistics.

Peter F. Brown, Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany, August. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised pos tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 634–639, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tomaž Erjavec, Ana-Maria Barbu, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Cvetana Krstev, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Behrang QasemiZadeh, Adam Radziszewski, Kiril Simov, Dan Tufiş, and Katerina Zdravkova. 2010. MULTEXT-east "1984" annotated corpus 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European

languages. *Language Resources and Evaluation*, 1(46):131–142.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 549–554, Genoa, Italy.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, page 862–873, Jeju Island, Korea.

Natalia Kotsyba, Andriy Mykulyak, and Ihor V. Shevchenko. 2011. UGTag: morphological analyzer and tagger for the Ukrainian language. In Stanisław Goźdź-Roszkowski, editor, *Explorations across Languages and Corpora*, pages 69–82, Frankfurt a. M.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585, Baltimore, Maryland, June. Association for Computational Linguistics.

František Krajnjak and Josif Kudzej (transl.). 2009. *Tetrajevanhelije*. Svitovŷj kongres Rusyniv, Prešov.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398, Jeju Island, Korea, July. Association for Computational Linguistics.

Paul R. Magocsi, editor. 2004. *Rusyn'skŷj jazŷk*. Najnowsze dzieje języków słowiańskich. Uniw. Opolski Inst. Filologii Polskiej, Opole.

Thomas Mayer and Michael Cysouw. 2014. Creating a Massively Parallel Bible Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Joakim Nivre and Leif Grönqvist. 2001. Tagging a corpus of spoken Swedish. *International Journal of Corpus Linguistics*, 6(1):47–78.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Anna Plishkova. 2009. *Language and national identity: Rusyns south of Carpathians*, volume 14 of *Classics of Carpatho-Rusyn scholarship*. Columbia University Press and East European Monographs, New York.

Stefan M. Pugh. 2009. *The Rusyn language: A grammar of the literary standard of Slovakia with reference to Lemko and Subcarpathian Rusyn*, volume 476 of *Languages of the World/Materials*. Lincom Europa, München.

Achim Rabus and Yves Scherrer. 2017. Lexicon induction for spoken Rusyn – challenges and results. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain.

Yves Scherrer. 2014. Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 30–38, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

H. A. Skrypnyk, editor. 2013. *Ukrajinci-Rusyny: Etnolinhvistyčni ta etnokul'turni procesy v istoryčnomu rozvytku*. Instytut mystectvoznavstva, fol'klorystyky ta etnolohiji im. M.T. Ryl's'koho, Kyjiv.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Alexander Teutsch. 2001. *Das Rusinische der Ostslowakei im Kontext seiner Nachbarsprachen*, volume 12 of *Heidelberger Publikationen zur Slavistik. A, Linguistische Reihe*. Lang, Frankfurt am Main, Berlin, Bern.

Swantje Westpfahl. 2014. STTS 2.0? Improving the tagset for the part-of-speech-tagging of German spo-

ken data. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 1–10, Dublin, Ireland.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL'01*, pages 200–207.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP'08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Daniel Zeman. 2015. Slavic languages in Universal Dependencies. In *Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning*, pages 151–163, Bratislava, Slovakia.