Lexicon Induction for Spoken Rusyn – Challenges and Results

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

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

Abstract

This paper reports on challenges and results in developing NLP resources for spoken Rusyn. Being a Slavic minority language, Rusyn does not have any resources to make use of. We propose to build a morphosyntactic dictionary for Rusyn, combining existing resources from the etymologically close Slavic languages Russian, Ukrainian, Slovak, and Polish. We adapt these resources to Rusyn by using vowel-sensitive Levenshtein distance, hand-written language-specific transformation rules, and combinations of the two. Compared to an exact match baseline, we increase the coverage of the resulting morphological dictionary by up to 77.4% relative (42.9% absolute), which results in a tagging recall increased by 11.6% relative (9.1% absolute). Our research confirms and expands the results of previous studies showing the efficiency of using NLP resources from neighboring languages for low-resourced languages.

1 Introduction

This paper deals with the development of a morphological dictionary for spoken varieties of the Slavic minority language Rusyn by leveraging the similarities between Rusyn and neighboring etymologically related languages. It is structured as follows: First, we give a brief introduction on the characteristics of the Rusyn minority language and the data our investigation is based upon. Afterwards, we describe our approach to lexicon induction using resources from several related Slavic languages and the steps we took to improve the matches from the dictionaries. Finally, we discuss the results and give an outlook on future work.

2 Rusyn and the Corpus of Spoken Rusyn

Rusyn belongs to the Slavic language family and is spoken predominantly in the Carpathian region, most notably in Transcarpathian Ukraine, Eastern Slovakia, and South Eastern Poland, where it is called Lemko.[1] Some scholars claim Rusyn to be a dialect of Ukrainian (Skrypnyk, 2013), others see it as an independent Slavic language (Pugh, 2009; Plishkova, 2009). While there is no denying the fact that Ukrainian is the standard language closest to the Rusyn varieties, certain distinct features at all linguistic levels can be detected. This makes the Rusyn varieties take an intermediary position between the East and West Slavic languages (for more details see, e.g., Teutsch (2001)). Nowadays, the speakers of Rusyn find themselves in a dynamic sociolinguistic environment and experience significant pressure by their respective roofing state languages Ukrainian, Slovak, or Polish. Thus, new divergences within the old Rusyn dialect continuum due to contact with the majority language, i.e., so-called border effects, are to be expected (Rabus, 2015; Woolhiser, 2005). In order to trace these divergences, and create an empirically sound basis for investigating current Rusyn speech, the Corpus of Spoken Rusyn (www.russinisch.uni-freiburg.de/corpus, Rabus and Šymon (2015)) has been created. It consists of several hours of transcribed speech as well as recordings.[2] Although the transcription in the corpus is not phonetic, but rather orthographic, both diatopic and individual varia-

[1] According to official data, there are 110 750 Rusyns, according to an "informed estimate" no less than 1 762 500, the majority of them living in the Carpathian region (Magocsi, 2015, p. 1).

[2] The corpus engine is CWB (Christ, 1994), the GUI functionality has been continuously expanded for several Slavic corpus projects (Waldenfels and Woźniak, 2017; Waldenfels and Rabus, 2015; Rabus and Šymon, 2015).

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 27–32,
Valencia, Spain, 4 April 2017.

tion is reflected in the transcription. The reason for that is that exactly this variation is what we want to investigate using the corpus, i.e., more "Slovak" Rusyn varieties should be distinguished from more "Ukrainian" or "Polish" varieties. Besides, variation in transcription practices of different transcribers cannot be avoided.

At the moment, Rusyn does not have any existing NLP resources (annotated corpora or tools) to make use of. The aim of this paper is to investigate first steps towards (semi-)automatically annotating the transcribed speech data. It goes without saying that the different types of variation present in our data significantly complicate the task of developing NLP resources.

3 Lexicon Induction

We propose to build a morphosyntactic dictionary for Rusyn, using existing resources from etymologically related languages. The idea is that if we know that a Rusyn word X corresponds to the Ukrainian word Y, and that Y is linked to the morphosyntactic descriptions M_1, M_2, M_n, we can create an entry in the Rusyn dictionary consisting of X and M_1, M_2, M_n. The proposed approach is inspired by earlier work by Mann and Yarowsky (2001), who aim to detect cognate word pairs in order to induce a translation lexicon. They evaluate different measures of phonetic or graphemic distance on this task. While they show that distance measures adapted to the language pair by machine learning work best, we are not able to use them as we do not have the required bilingual training corpus at our disposal. Scherrer and Sagot (2014) use such distance measures as a first step of a pipeline for transferring morphosyntactic annotations from a resourced language (RL) towards an etymologically related non-resourced language (NRL).

Due to the high amount of variation and the heterogeneity of the Rusyn data (our NRL), we resolved to use resources from several neighboring RLs, namely from the East Slavic languages Ukrainian and Russian as well as from the West Slavic languages Polish and Slovak.[3] This makes sense, because the old Rusyn dialect continuum features both West Slavic and East Slavic linguistic traits, with more West Slavic features in the westernmost dialects and more East Slavic ones in the easternmost dialects. Moreover, the respective umbrella languages – Ukrainian, Slovak, and Polish – exert considerable influence on the Rusyn vernacular. In fact, the overwhelming majority of Rusyn speakers are bilingual.

[3] As a matter of fact, Russian is no neighboring language to Rusyn, but since for historical reasons there are numerous Russian borrowings in Rusyn and since NLP resources for Russian are developed quite well, we also include Russian.

Language	Source	Entries
Polish	MULTEXT-East	1.9M
Russian	MULTEXT-East	244k
Russian	TnT (RNC)	373k
Ukrainian	MULTEXT-East	300k
Ukrainian	UGtag	4.6M
Slovak	MULTEXT-East	1.9M

Table 1: Sizes of the morphosyntactic dictionaries used for induction.

3.1 Data

Our RL data consist of morphosyntactic dictionaries (i.e., files associating word tokens with their lemmas and tags) from Ukrainian, Slovak, Polish, Russian. All of them were taken from the MULTEXT-East repository (Erjavec et al., 2010a; Erjavec et al., 2010b; Erjavec, 2012). As Rusyn is written in Cyrillic script, we converted the Slovak and Polish dictionaries into Cyrillic script first. During the conversion process, we made the tokens more similar to Rusyn by applying certain linguistic transformations (e.g., denasalization in the Polish case) and thus excluded some output tokens that could not possibly match any Rusyn tokens for obvious linguistic reasons.

As mentioned above, the standard language closest to the Rusyn varieties is Ukrainian. Several Ukrainian NLP resources exist, e.g., the Ukrainian National Corpus.[4] However, these resources cannot easily be used to train taggers or parsers. UGtag (Kotsyba et al., 2011) is a tagger specifically developed for Ukrainian; it is essentially a morphological dictionary with a simple disambiguation component. Its underlying dictionary is rather large and can be easily converted to text format, making it a good addition to the small MULTEXT-East Ukrainian dictionary. For Russian, we complemented the small MULTEXT-East dictionary with the TnT lexicon file based on data from the Russian National Corpus (Sharoff et al., 2008). We also harmonized the MSD tags (morphosyntactic descriptions) across all languages and data

[4] www.mova.info

sources. Table 1 sums up the used resources.

Our NRL data consist of 10 361 unique tokens extracted from the Corpus of Spoken Rusyn (which currently contains a total of 75 000 running words). In addition, we were able to obtain a small sample of morphosyntactically annotated Rusyn, amounting to 1 047 tokens; the induction methods are evaluated on this sample.

3.2 Exact Matches

As a baseline, we checked how many Rusyn word forms could be retrieved by exact match in the four RL lexicons. Despite Rusyn being closely related to the dictionary languages, the results are rather poor: merely 55.47% of all Rusyn tokens were found in at least one RL lexicon (see Table 2, first column).

We further show the relative contributions of the four RLs in Table 2. Ukrainian is by far the most successful language, both with respect to the overall matched words (i.e., words matched with Ukrainian and possibly other RLs) and to uniquely matched words (i.e., words matched with Ukrainian but not with any other RL). This is due to several factors: e.g., Ukrainian is the RL with the smallest linguistic distance to the Rusyn varieties, the Ukrainian dictionary is considerably larger than the other dictionaries, and the relative majority of tokens in the corpus belongs to "Ukrainian" varieties of Rusyn.

Table 2 also shows some ambiguity measures. On average, a Rusyn token is found in 1.66 resourced languages and associated with 3.28 tags. Trivially, a Rusyn word is matched with exactly one RL word, as both forms need to be identical for exact match.

We evaluated the correctness of the induced lexicon on the annotated Rusyn sample. More than 84% of the 1 047 words were covered, and the correct tag was among the induced ones for more than 78% of words. (We do not attempt to disambiguate the tags here, which is why we only report recall.) We also report noise, which is defined as the amount of covered but wrongly tagged words (i.e., coverage - recall). With a noise of only 6%, we can characterize exact match as a high-precision, low-recall method.

The poor coverage often results from orthographic mismatches by merely one or a few different letters between the Rusyn token and its RL counterpart. In order to improve the coverage, we propose different types of transformations, as described in the following sub-sections.

3.3 Daitch-Mokotoff Soundex Algorithm

Soundex is a family of phonetic algorithms for indexing words and, in particular, names by their pronunciation and regardless of their spelling (Hall and Dowling, 1980). The principle behind a Soundex algorithm is to group different graphemes into a small set of sound classes, where all vowels except the first of a word are discarded. The Daitch-Mokotoff Soundex is a variant of the original (English) Soundex that is adapted to Eastern European names (Mokotoff, 1997).

Matching soundex-transformed RL words with soundex-transformed NRL words allowed us to obtain a coverage of 97.16% (i.e., almost all NRL words were matched), but in fact, each matched NRL word was associated with as many as 630 RL words on average. Thus, this algorithm proved to be too radical as it identified a multitude of unrelated tokens. In particular, vowel removal neutralized nearly all inflectional suffixes. While Soundex algorithms have proved useful for matching names with different spellings, they are clearly not adapted to our task. Therefore, we had to resort to less radical transformation methods.

3.4 Hand-Written Transformation Rules

The Slavic RLs in question differ with respect to regular sound changes and morphological correspondences that are reflected in orthography. For instance, Rusyn dialects reflect Common Slavic *ě as i, while Russian yields e. Moreover, Rusyn verbs in the infinitive end in -ти, while Russian has -ть. About 40 such transformation rules were formulated for each language and implemented in *foma* (Hulden, 2009).

During the lexicon induction process, each RL word was transformed with the appropriate rules to resemble Rusyn. All rule applications were optional, yielding a multitude of candidates for each RL word. Whenever one of the candidates corresponded to an existing Rusyn word, this was counted as a match. As shown in Table 2, applying these transformation rules yielded a considerable increase of matched words (compared with exact match) to more than 76%. Ambiguity levels rise slightly, and the contributions of the different languages rise uniformly. The better coverage is confirmed on the test set, and tagging recall also in-

	Exact	Soundex	Rules	Leven.	R+L	L+R
Words matched with any RL	55.47%	97.16%	76.38%	98.09%	98.38%	98.09%
Words matched with PL	13.92%	87.24%	19.17%	25.80%	24.66%	22.89%
Words matched with RU	20.03%	92.57%	30.30%	37.26%	38.03%	34.41%
Words matched with SK	19.43%	93.45%	28.17%	39.62%	37.68%	35.63%
Words matched with UK	38.84%	96.06%	49.49%	70.09%	64.89%	63.69%
Words matched with PL only	3.91%	0.10%	5.16%	5.76%	6.34%	6.81%
Words matched with RU only	3.94%	0.12%	7.44%	6.15%	8.79%	8.82%
Words matched with SK only	4.14%	0.31%	6.69%	8.64%	9.33%	10.49%
Words matched with UK only	21.69%	1.27%	26.25%	33.46%	33.23%	36.12%
Average RL language ambiguity	1.66	3.80	1.66	1.76	1.68	1.60
Average RL word ambiguity	1.00	630.74	1.29	2.17	1.81	1.51
Average tag ambiguity	3.28	271.62	3.66	5.08	4.34	3.93
Coverage on test set	84.2%	—	90.4%	99.0%	99.6%	99.0%
Tagging recall on test set	78.2%	—	81.9%	87.3%	87.1%	86.4%
Noise on test set	6.0%	—	8.5%	11.7%	12.5%	12.6%

Table 2: Results of the different lexicon induction methods. Percentages show how many distinct Rusyn words were matched with any of the four RLs, with at least one of the RLs, and with exactly one RL. The last rows show the coverage, tagging recall and noise on the annotated Rusyn sample.

creases by more than 3%,[5] while the noise level increases by 2.5%.

3.5 Vowel-Sensitive Levenshtein Distance

As an alternative to hand-written rules, we also tested a vowel-sensitive variant of Levenshtein distance (Levenshtein, 1966), following Mann and Yarowsky (2001). In this variant, edit operations on vowels are assigned a weight of 0.5, whereas edit operations on consonants use the standard weight of 1. Using this variant was motivated by the fact that Rusyn vowels differ systematically and significantly from the vowels present in neighboring Slavic languages and also within different Rusyn varieties. We also normalize distances by the length of the longer word.

Initial experiments have shown that most NRL words lie within a small distance of an RL word, and that matches with high distance values are most often wrong. Because of that, we decided to discard all matches with distance values higher than 0.25. This considerably decreased word and tag ambiguity while losing merely 1.95% of matched tokens. Even with this threshold, the number of matched words as well as the tagging recall – but also the noise – is higher than with the rules.[6] Future research will show whether the optimal threshold can be found automatically, e.g., by using a small annotated development corpus.

Despite the good coverage, we were concerned by the higher ambiguity values, which is why we decided to combine Levenshtein distance with the transformation rules.

3.6 Rules and Levenshtein

In this first combined approach, we complement the rules with Levenshtein results in order to increase coverage: Whenever the rules do not succeed in creating a match for a Rusyn word, we back off to the corresponding Levenshtein results. This combination outperforms both individual methods in terms of matches (98.38%, as compared to 76.38% and 98.09%). As expected, the resulting ambiguity levels lie between those of the rules and those of the Levenshtein method. The coverage on the test set also increases, but this is not followed by better tag recall.[7]

[5] This increase is statistically significant with $p < 0.05$; $\chi^2(1; N = 1047) = 4.32$.

[6] The tagging recall difference is statistically significant: $\chi^2(1; N = 1047) = 11.89; p < 0.001$.

[7] The tagging recall difference is not statistically significant: $\chi^2(1; N = 1047) = 0.02; p = 0.90$.

3.7 Levenshtein and Rules

In the second combined approach, we start with the Levenshtein results and filter them using the rules in order to further reduce ambiguity. The underlying idea is that in case of ambiguity, some of the Levenshtein-induced results will be correct and some will not. The correct ones will relate to the Rusyn words by known correspondences such as those implemented in the rules, while the incorrect ones will not. Hence, we took all Rusyn words matched (using Levenshtein) with more than one distinct RL word and transformed these RL words using the rules. We then checked whether the rules were able to "move" the RL words closer to Rusyn, i.e., whether the minimum Levenshtein distance of any transformed word was lower than the original Levenshtein distance. We only kept those RL words for which this check succeeded.

For example, the Levenshtein method matched the Rusyn word *береме* 'we take' with Polish *беремы*, Russian *берем*, *беремя*, Slovak *берieме*, *берме*, and Ukrainian *берем*, *беремо*, all of which obtained a Levenshtein distance of $0.08\overline{3}$ (one vowel substitution, insertion, or deletion in a word of length 6). The rule base contains rules which transform the Ukrainian ending -мо, the Russian ending -м, and the Polish ending -мы to Rusyn -ме. Hence, the Russian, Ukrainian and Polish forms are transformed to *береме*, reducing the distance to the Rusyn word to 0 (exact match). Therefore, we only keep *беремы* and *берем* as well as *беремо* and discard the other candidates. Since all three forms share the identical tag, the Rusyn word is morphologically disambiguated and only receives the correct reading as a verb in first person plural present tense.

This filtering approach resulted in an even further decrease of ambiguity while maintaining a high match rate: Average source word ambiguity dropped from 2.17 using the Levenshtein approach via 1.81 using rules and Levenshtein to 1.51 using Levenshtein and rules. This is close to the average source word ambiguity of 1.29 achieved when using exclusively the rules. However, a high amount of matched tokens could be maintained. While the combined Levenshtein and rules approach seems to be most successful in terms of matched words and ambiguity levels, the tagging recall actually suffers slightly.[8] This is to be expected, as reducing the ambiguity mainly increases the precision (sometimes at the expense of recall), which is not measured here.

4 Conclusion and Further Work

We have shown that a morphological dictionary for Rusyn can be created by leveraging existing resources of four etymologically closely related languages. Induction methods based on Levenshtein distance and hand-written philological rules significantly outperform exact match, both in terms of matched words and in terms of tagging recall. Also, the figures show that while there are significant differences in the individual contribution of each language, all languages contribute to the induction process.

Further work will be devoted to extending our work to lemmatization (which is available in the four RL dictionaries) and to making use of the newly created resources by statistical taggers (cf. Scherrer and Rabus (2017)).

Acknowledgments

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

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

References

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, pages 23–32.

Tomaž Erjavec, Ştefan Bruda, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Peter Holozan, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Igor Shevchenko, Kiril Simov, Lydia Sinapova, Han Steenwijk, Laszlo Tihanyi, Dan Tufiş, and Jean Véronis. 2010a. MULTEXT-east free lexicons 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Ivan Derzhanski, Dagmar Divjak, Anna Feldman, Mikhail Kopotev, Natalia Kotsyba, Cvetana Krstev, Aleksandar Petrovski, Behrang QasemiZadeh, Adam Radziszewski, Serge Sharoff, Paul Sokolovsky, Duško Vitas, and Katerina Zdravkova. 2010b. MULTEXT-east

[8] The difference in tagging recall compared to Levenshtein is again not statistically significant: $\chi^2(1; N = 1\,047) = 0.34; p = 0.56$.

non-commercial lexicons 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.

Patrick A. V. Hall and Geoff R. Dowling. 1980. Approximate string matching. *ACM Computing Surveys*, 12(4):381–402.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece, April. Association for Computational Linguistics.

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.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, February.

Paul R. Magocsi. 2015. *With Their Backs to the Mountains: A History of Carpathian Rus' and Carpatho-Rusyns*. Central European University Press, Budapest.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Pittsburgh, PA, USA.

Gary Mokotoff. 1997. Soundexing and genealogy. `http://www.avotaynu.com/soundex.htm`. Accessed: 2016-12-20.

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 Andrianna Šymon. 2015. Na novŷch putjach isslidovanja rusyns'kŷch dialektu: Korpus rozhovornoho rusyns'koho jazŷka. In Kvetoslava Koporová, editor, *Rusyn'skŷj literaturnŷj jazŷk na Slovakiji: Zbornyk referativ z IV. Midžinarodnoho kongresu rusyn'skoho jazŷka*, pages 40–54. Prjašiv.

Achim Rabus. 2015. Current developments in Carpatho-Rusyn speech – preliminary observations. In Patricia A. Krafcik and Valerij Ivanovyč Padjak, editors, *Juvilejnyj zbirnyk na čest' profesora Pavla-Roberta Magočija*, pages 489–496. Užhorod.

Yves Scherrer and Achim Rabus. 2017. Multi-source morphosyntactic tagging for spoken Rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, Valencia, Spain.

Yves Scherrer and Benoît Sagot. 2014. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Proceedings of LREC 2014*, pages 502–8, Reykjavik, Iceland.

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating Russian tagsets. In *Proceedings of LREC 2008*, Marrakech, Morocco.

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.

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.

Ruprecht von Waldenfels and Achim Rabus. 2015. Recycling the Metropolitan: building an electronic corpus on the basis of the edition of the Velikie Minei Čet'i. *Scripta & e-Scripta*, 14–15:27–38.

Ruprecht von Waldenfels and Michał Woźniak. 2017. SpoCo – a simple and adaptable web interface for dialect corpora. *Journal for Language Technology and Computational Linguistics*, 31(1).

Curt Woolhiser. 2005. Political borders and dialect divergence/convergence in Europe. In Peter Auer, Frans Hinskens, and Paul Kerswill, editors, *Dialect change*, pages 236–262. Cambridge Univ. Press, Cambridge.