HAL
archives-ouvertes.fr

Dependency Parsing of Code-Switching Data with Cross-Lingual Feature Representations

Niko Partanen, Kyungtae Lim, Michael Rießler, Thierry Poibeau

▶ **To cite this version:**

Niko Partanen, Kyungtae Lim, Michael Rießler, Thierry Poibeau. Dependency Parsing of Code-Switching Data with Cross-Lingual Feature Representations. International Workshop on Computational Linguistics for Uralic Languages, Jan 2018, Helsinki, Finland. ACL, pp.1 - 17, 2018, <aclweb.org/anthology/W18-0200>. <hal-01722243>

HAL Id: hal-01722243
https://hal.archives-ouvertes.fr/hal-01722243

Submitted on 3 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dependency Parsing of Code-Switching Data with Cross-Lingual Feature Representations

Niko Partanen
LATTICE
CNRS & ENS / PSL & Université Sorbonne nouvelle / USPC
nikotapiopartanen@gmail.com

KyungTae Lim
LATTICE
CNRS & ENS / PSL & Université Sorbonne nouvelle / USPC
kyungtae.lim@ens.fr

Michael Rießler
Faculty for Linguistics and Literary Sciences
University of Bielefeld
michael.riessler@uni-bielefeld.de

Thierry Poibeau
LATTICE
CNRS & ENS / PSL & Université Sorbonne nouvelle / USPC
thierry.poibeau@ens.fr

Abstract

This paper describes the test of a dependency parsing method which is based on bidirectional LSTM feature representations and multilingual word embedding, and evaluates the results on mono- and multilingual data. The results are similar in all cases, with a slightly better results achieved using multilingual data. The languages under investigation are Komi-Zyrian and Russian. Examination of the results by relation type shows that some language specific constructions are correctly recognized even when they appear in naturally occurring code-switching data.

Tiivistelmä

Tutkimus arvioi dependenssianalyysin menetelmää, joka perustuu kaksisuuntaiseen LSTM-piirrerepresentaatioon ja monikieliseen 'word embedding' -malliin, sekä arvioi tuloksia yksi- ja monikielisissä aineistoissa. Tulokset ovat samantapaisia, mutta hieman korkeampia moni- kuin yksikielisissä aineistoissa. Tutkitut kielet ovat komisyrjääni ja venäjä. Tulosten yksityiskohtaisempi analyysi riippuvuuksien mukaan osoittaa, että tietyt kielikohtaiset suhteet on tunnistettu oikein jopa niiden esiintyessä luonnollisissa koodinvaihtoa sisältävissä lauseissa.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details:
http://creativecommons.org/licenses/by/4.0/

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 1–17,
Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

1 Introduction

Spontaneous speech data of small, endangered languages most commonly contain code-switching, ad-hoc borrowings and other kinds of language contact phenomena originating from the non-target contact language(s). Consequently, spoken corpora originating from such data contain numerous utterances in which linguistic elements from at least two languages co-occur. The most usual occurrences are combinations of target-language utterances including lexical and morphological elements from the contacting majority language. Corpus data of this type represents a particular challenge for morphological analysis and especially for dependency parsing. Although the basic morphological properties can usually be analyzed on the basis of individual languages and parsers can be targeted towards those, the syntactic dependencies are inevitably interspersed individual tokens from different languages, and thereby cannot be easily approached with tools that are able to target only monolingual data.

The present paper looks at an approach that has been introduced as The Multilingual BIST-Parser by Lim and Poibeau (2017). The tool was developed in order to perform dependency parsing on considerably low-resource languages, and the work was originally carried out within the CONLL-U Shared Task for 2017. Lim and Poibeau (2017) have shown that multilingual word embeddings can be used to train a model that combines data from multiple languages, and these seem to be particularly useful in low-resource scenarios where one of the languages has only a small amount of available training data.

The target language in the present paper is Komi-Zyrian (henceforth Komi), which belongs to the Permic branch of the Uralic language family. The language is spoken predominantly in the Komi Republic of the Russian Federation by approximately 160,000 speakers. Computational linguistic research on Komi is so far only in a development stage. However, an FST morphological analyzer and a (rudimentary) syntactic parser based on Constraint Grammar are available at *Giellatekno/Divvun – Saami Language Technology* at UiT The Arctic University of Norway[1] and work on a complete Constraint Grammar description to be implemented into a rule-based syntactic parser is currently carried out in collaboration by Giellatekno, the Izhva Komi Documentation Project Gerstenberger et al. (2016, 2017) and FU-Lab[2], which has also created a written Komi National corpus (with over 30M words), free electronic dictionaries and a Hunspell checker (including morpheme lists).

Our own initial dependency parsing tests were conducted by testing various different language pairs with Komi as parts of multilingual word-embedding models, in order to find out which combinations can reach the best performance. In our earlier tests, the best results were achieved when the majority of the training data were from a genealogically related language, in this case Finnish. This went against our hypothesis that the genealogically unrelated contemporary contact languages would have been particularly useful from a NLP perspective due to prolonged language contact and resulting convergence in Komi grammar and lexicon. Although it is possible to build truly multilingual models, such as a parser that combines Finnish, Russian and Komi word embeddings and training corpora in order to operate on Komi, we found that a bilingual Finnish-Komi model performed best in our tests for monolingual Komi data. However, especially if the results were analyzed more in detail beyond the LAS

[1] `http://giellatekno.uit.no`; for the technical documentation of the research on Komi, see `http://giellatekno.uit.no/doc/lang/kom/`; Jack Rueter (Helsinki) has been the main developer

[2] "The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages"; `http://fu-lab.ru`

and UAS scores (for explanation of these evaluation metrics, see Kübler et al., 2009, 79), the different language pairs will likely show different benefits and drawbacks in distinct areas of analysis, and testing the parsing method on data that naturally contains materials from both languages used in training is used here as one method to tease apart language specific changes in parser's behavior.

The next part of the paper describes this problem in further detail with examples from spoken language corpora.

2 Problem Description

The first example 1 is taken from the spoken Iźva (dialectal) Komi corpus Blokland et al. (2009–2017) (henceforth called IKDP) and represents naturally occurring spoken language mixed with Russian elements (Russian marked in boldface).

(1) ***До*** ***школьн-ого*** ***возраста*** *ветл-і* ***родитель****-яс-кед* *тундра-ын.*
until school-GEN age-GEN go-1SG.PST parents-PL-COMIT tundra-INES

'Until the school age I went to the tundra together with my parents.'

The example starts with a Russian prepositional phrase meaning 'until the school age', but it is followed by a direct shift to Komi. The word for 'parents' is also Russian, but it is inflected according to Komi morphological rules and in the same manner as native Komi words would be inflected. Such morphologically integrated nouns are often described as Russian loanwords in Komi, but as will be argued in Section 6 below, this approach may not be very applicable in the context of Uralic languages spoken in Russia. We are therefore referring to it as a "mixed" form. In order to compare the sentences, two bilingual Komi-Russian native speakers[3] have translated the example into both languages. It must be noted that because both Komi and Russian have rather flexible word orders, this aspect is not taken into account in the present analysis, although there is clear variation in both languages with respect to the semantic nuances of different orderings.

Note also that the purely Komi variant of the example sentence would still include two lexical items of Russian origin, namely *school* and *tundra*. Although the basic sentence structure may look similar, Komi and Russian have rather different syntactic structures overall. For instance, Komi uses cases extensively along with postpositions, whereas Russian uses predominantly prepositions.

(2) *Школа-ö* *пыр-тödз* *ветл-і* *бать-мам-кöд* *тундра-ын.*
school-ILL enter-GER.DUR go-1SG.PST parents-PL-COMIT tundra-INES

'Until the school (age) I went to the tundra together with my parents.'

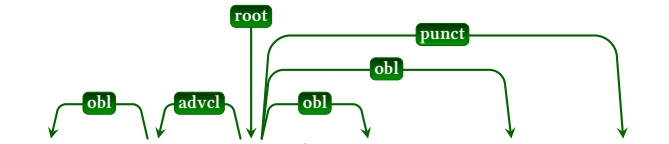
For the sake of thoroughness, it is also worth looking into one possible way to express the utterance entirely in Russian.

(3) *До* *школьн-ого* *возраста* *езди-л* *с* *родител-ями* *в* *тундр-у.*
until school-GEN age-GEN go-PST with parents-PL.INSTR to tundra-LOC

'Until the school age I went to the tundra together with my parents.'

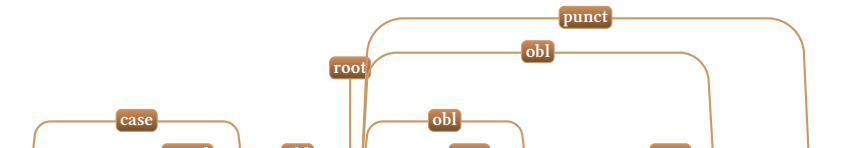
[3]Thanks to Vasili Chuprov and Sergei Gabov.

Based on different Universal Dependency (UD) corpora, the dependency structure of the Komi variant should be analyzed as in 4.



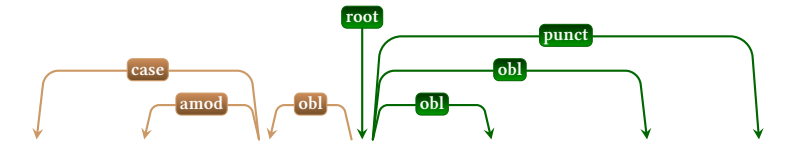
(4)

The Russian tree, on the other hand, is 5.



(5)

Based on these examples, we can conclude that a correctly analyzed dependency structure for the mixed utterance would be as presented in 6, as it effectively combines the relevant parts of the Russian and Komi annotations. As the applied annotation model is the same, the monolingual dependencies should not differ from multilingual ones.



(6)

Although the Multilingual BIST-parser is trained with multilingual material, the goal has been primarily to parse the lesser resourced language. All earlier tests have been conducted using strictly monolingual data, although different assumptions can be made about the parallel structures in the languages included in the model. Applying the parser to data that truly contains syntactic constructions specific to only the individual languages within the same utterances reveals about the parser's ability to correctly identify structures of this type. If both distinctly Russian and Komi constructions can be parsed successfully within the same sentence, this indicates that the model is able to learn and deduce language-specific structures even when they co-occur. This would open up new possibilities for automatic analysis of such kind of data.

3 Related Studies

Multilingual dependency parsing aims at building a dependency tree for several languages using one and the same model. Three major approaches have been suggested

for tackling such a task: 1) the cross-lingual annotation projection approach, 2) the joint modeling approach, and 3) the cross-lingual representation learning approach (cf. Guo et al., 2015). The main idea of the cross-lingual annotation projection approach is to project the syntactic annotations trough word alignments from a source language onto a target language (Mann and Yarowsky, 2001; Tiedemann, 2014). In a similar way, the joint modeling approach is carried out using projected dependency information for grammar inductions (Liu et al., 2013) and rule-based work (Naseem et al., 2010, 2012).

The cross-lingual representation learning method is focused on learning cross-lingual features by aligning (or mapping) feature representations (e.g. embedding) between the source and target languages. In general, cross-lingual representation learning can be divided into two approaches depending on whether or not the parser uses lexicalized features (e.g. word embedding). Since it is relatively easy to train a parser using supervised learning, many existing cross-lingual representation learning studies have been conducted with the delexicalized approach using POS tag-sets and word sequences (McDonald et al., 2011, 2013; Dozat et al., 2017). Such an approach includes training a dependency model with the source language (e.g. English), then processes the target language (e.g. French) using the model trained according to the source language. On the other hand, the lexicalized approach is able to adapt diverse lexical features while in training. The features adapted for the dependency parsing include cross-lingual word cluster features (Täckström et al., 2012), multilingual word embeddings (Guo et al., 2015, 2016; Ammar et al., 2016b,a) and language identification embeddings (Naseem et al., 2012; Ammar et al., 2016a).

From the perspective of code-switching, conversational code-switching problems have been studied mainly with regard to language identification (e.g. Solorio et al., 2014; Barman et al., 2014) and information extraction (e.g. Sharma et al., 2014) problems. This is because in order to process cross-lingual dependency parsing, language identification and morphological analysis for those languages must precede the processing. Ammar et al. (2016b) suggested that his multilingual model-transfer parser could be used to parse input with code-switching but were not able to conduct the experiment due to the lack appropriate test corpora.

4 Cross-Lingual Dependency Parsing

In this study, we invested our effort in developing the cross-lingual representation learning method with lexicalized features for the dependency parsing of code-switching scenarios. All the cross-lingual approaches discussed in Section 3, can be applied for our study, but in terms of the availability of language resources, cross-lingual representation learning is considered the best choice because of the lack of annotated corpora. Also in regard to the performance, existing studies have already shown that representation learning with lexical features performs better than the other models (Ammar et al., 2016a; Lim and Poibeau, 2017).

In this section, we describe two main ideas for parsing code-switching data using the cross-lingual representation learning approach. One of the main goals of our research is to build cross-lingual word embeddings based on supervised learning. The other is to find a way to address adapting cross-lingual word embedding in order to build a dependency parsing model.

4.1 Cross-Lingual Word Representations

As discussed in Section 3, adding lexical information for feature representations can improve performance in cross-lingual parsing. Various approaches have been investigated for the training of cross-lingual word embeddings mainly for resource-rich languages. Moreover, most of these approaches relied on the existence of a parallel corpus, especially for languages from the Indo-European family (cf. Ammar et al., 2016a; Guo et al., 2016). As we discussed earlier, however, this study focuses on code-switching scenarios in low-resource language data. Thus, we are constrained by the fact that there is no parallel corpus and no larger annotated dataset for training a dependency parser for the (low-resource) target language Komi. However, it must be noted that even for low-resource languages, we need raw texts as the minimum resource to train a word embedding. In this study, we trained a monolingual embedding for Komi by using raw text available in the public domain. The Komi texts used have been taken from the National Library of Finland's Fenno-Ugrica collection[4], and proofread versions of those Public Domain texts are available in FU-Lab's portal *Komi Nebögain*[5]. Niko Partanen has created a list of books included both in Fenno-Ugrica and FU-Lab[6], and the currently available data adds up to one million tokens. For the contact language Russian we have used pre-trained Wikipedia word embeddings published by Facebook and described in Bojanowski et al. (2016).

In a similar manner to the low-resource constraints, Artetxe et al. (2017) suggested a powerful method for projecting two monolingual embeddings in a single vector space with almost no bilingual data. Traditionally, the projection (or mapping) method for word embeddings requires a large parallel corpus or a bilingual dictionary in order to map two different word embeddings in a distributional space (Artetxe et al., 2016; Guo et al., 2015). However, Artetxe et al. (2017) showed a possible method for mapping two different embeddings based on the reinforcement learning approach with just 25 pairs of vocabularies but with almost no degradation of performance. The main idea in this method is to project two embeddings trained by different languages based on the linear transformation with bilingual word pairs.

The projection method can be described as follows. Let X and Y be the source and target word embedding matrix so that x_i refers to ith word embedding of X and y_j refers to jth word embedding of Y. And let D is a binary matrix, where D_{ij} = 1, if x_i and y_j are aligned. Our goal is then to find a transformation matrix W such that Wx approximates y. This is done by minimizing the sum of squared errors (following Artetxe et al., 2017), cf. 7.

(7)
$$\arg\min_{W} \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \|x_i W - y_i\|^2$$

The method is relatively simple to apply in our case because once we have a bilingual dictionary available, converting the dictionary as D is not a problem. We followed Artetxe's 2017 mapping idea to train a bilingual word embedding for Komi-Russian using a bilingual dictionary. The size of the dictionary used for training is 7,642 pairs, and the projected word embedding is 5.9G. Those dictionaries and projected word

[4]https://fennougrica.kansalliskirjasto.fi
[5]http://komikyv.org
[6]https://github.com/langdoc/kpv-lit

embedding are accessible in a public repository.[7] Dictionary is extracted from Jack Rueter's Komi-Zyrian dictionaries that have translations to several languages.[8]

4.2 Cross-Lingual Dependency Parsing Model

As discussed in Section 3, the major idea of the cross-lingual representation learning method is to take aligned features, especially syntactic and lexical features. Since the Universal Dependencies (UD) (Nivre et al., 2017) model provides cross-linguistically consistent grammatical annotation, we do not need to consider aligning syntactic features among the languages (i.e., POS tags, dependency tags). However, in terms of the semantic point of view, ignoring lexical features may lead to a lack of semantic information not only in monolingual but also in multilingual dependency parsing.

A recent multilingual parsing experiment, the CoNLL 2017 shared task, has addressed dependency parsing for low-resource languages using a multilingual approach (Zeman et al., 2017). The main approach was cross-lingual representation learning, and most teams applied the delexicalized model to process the low-resource languages with around 20 samples of annotated sentences. However, the LATTICE team (Lim and Poibeau, 2017) suggested concatenating a bilingual word embedding as a lexicalized feature, which is mapped by a bilingual dictionary taken from Swadesh lists. In practice larger dictionaries would improve the result, and this has been done later, but the shared task had strictly specified resources. On the other hand, a small dictionary seems to be enough to align the embeddings reasonably well. All features, including lexicalized ones, are then fed into a bidirectional Long Short-Term Memory (LSTM) to take concatenated feature representations for each token. By using the concatenated features (vectors) as an input, Lim and Poibeau applied graph-based parsing, which views the parsing problem as a search for the best-scored tree graph.

As Lim and Poibeau (2017) suggested, the BiLSTM feature representation with lexicalized features is crucial for multilingual dependency parsing, particularly in low-resource scenarios. Since we assume that there are no UD corpora for low-resource languages, one common alternative approach is to take a training corpus from another language. Once we find a grammatically related language, we then simply train a dependency model with the mapped bilingual word embedding and a UD corpus of the related language. Although the training corpus is written in the related language, the system is possible to replace tokens with ones from the low-resource language by using pre-trained bilingual word embeddings, in which vocabulary items with the same meaning are mapped between two languages. LSTM is a specific type of Recurrent Neural Network (RNN), so it also has hidden layer with hidden vectors for each sequence, $h = (h_1,h_2,...,h_n)$. If we look at the hidden layer in a sequence i, it is defined as in 8.

(8) $h_i = (W_{th}t_i + W_{hh}h_{i-1} + b_h)$

The basic LSTM model is able to make use of the previous context based on the computation $W_{hh}h_{i-1}$ (to put it simply, we can think of W_{hh} as a hidden input weight matrix and h_{i-1} as the previous value of the hidden layer). Thus, BiLSTM can store contexts from the LSTM both in regular order ($LSTM_{forward}$) and inverse order ($LSTM_{backward}$). For further details on the LSTM model, see Huang et al. (2015) and Cho (2015).

[7] https://github.com/jujbob/multilingual-models
[8] https://victorio.uit.no/langtech/trunk/words/dicts/kpv2X/src

For the current study, we have extended the parser by Lim and Poibeau (2017) using the multilingual word embeddings proposed in Section 4.1. The bilingual dictionaries used in the word embedding alignment contained several thousands of word pairs, and the recent study by Artetxe et al. (2017) shows that the dictionary size we operate with should be large enough to reach a high level of alignment accuracy.

5 Experiment Design

The following section discusses in more detail the creation and use of the corpora used for training and testing.

5.1 Training Corpora

The applied tools have been developed specifically for parsing low-resource languages, and this study originates from the same background. The main part of training data consists of a Russian UD v2.0 corpus with 3,850 sentences[9], while the Komi part, which we have prepared, is only 40 sentences.

5.2 Testing Corpora

The early-stage Komi-Zyrian Universal Dependency corpus was used for the model training[10], as this makes the results comparable with our earlier studies and was readily available. All in all, the tests in this study were performed on three different subsets or variants of test corpora, which are described below. All of the data used is publicly available in the *IWCLUL* branch of the repository.

1. Monolingual written Komi test corpus: 80 sentences
2. Multilingual written Komi-Russian test corpus, based on the monolingual corpus but adapted to contain constructions comparable to those that occur in spoken data: 80 sentences
3. Spoken Komi test corpus, contains spontaneous code-switching and code-mixing: 25 sentences

Similar to our earlier research, the monolingual Komi testing corpus was used as one method for evaluating the baseline for the results. Another Komi corpus currently being built will eventually include more spoken data, however it is not directly comparable with the monolingual testing corpus as the examples are entirely different. The spoken language data, although dialectal, is still phonologically and morphologically close to the written language, and in this case the data were slightly normalized in order to harmonize the transcription conventions with the orthographic representation in the written corpora used.

As the kinds of constructions we were interested in analyzing tend to occur only in spontaneous spoken language, it was not possible to use a parallel corpus of written texts to compare the performance as such. Instead, another approach was adopted in which the code-switching-like elements were inserted into an originally monolingual testing corpus. It must be stressed that the Russian elements were not inserted

[9]https://github.com/UniversalDependencies/UD_Russian/releases/tag/r2.0
[10]https://github.com/langdoc/UD_Komi-Zyrian

randomly, but were carefully crafted to follow patterns observed in the real spoken data. The creation of mixed test corpus was helped by the large number of available translations for the texts used. To illustrate this, we can take one sentence that is part of the testing corpus:

(9) *Шофер-ыс, том зонка на, дзик-ӧдз растеряйтч-ис .*
driver-3SG young boy still totally-TERM get_confused-PST.3SG

'Driver, still a young boy, got totally confused.'

As the same source book has translations into multiple minority languages of Russia, with the original Russian version, there is always access to multilingual versions of the same text segments. In this case the Russian version is as presented below:

(10) *Водитель машин-ы, еще молодой парнишка, совсем растеря-л-ся.*
driver car-GEN still young boy totally get_confused-PST.REFL

'Driver, still a young boy, got totally confused.'

With translations available, it is possible to compare the examples into occurrences that there are in spoken language corpus that naturally contains intermixed Russian. Although the details vary, we can at least add pointers into example sentences in spoken corpus that contain *comparable* occurrences, although they naturally would never be identical, or comparable from only one point of view. The example sentence above has been restructured in following way, Russian in bold:

(11) *Шофер-ыс, том зонка на,* ***совсем растеря-л-ся****.*
driver-3SG young boy still totally get_confused-PST-REFL

'Driver, still a young boy, got totally confused.'

The acceptability of the adapted sentences can be justified at least partly by the test corpus design, of which up to 35% originates from texts that have parallel variants in Russian. This has been very useful in order to examine how similar the sentences would be in different languages. Additionally, 40% of the testing corpus has been translated from Komi into Russian. In the majority of cases, the basic structure has indeed been so similar that the Russian and Komi versions should, to a large extent, display identical dependency structures with core relations, although the details still differ substantially. It is left outside the current investigation whether the translations that are present are the most natural ways to express these ideas in either of the languages, as the goal was primarily evaluate how the parser behaves in this kind of scenario.

In order to make these decisions explicit, the mixed corpus version has an additional metadata field *spoken_comparison*, which contains a link to the IKPD corpus of spoken language recordings that exhibits comparable Russian constructions. In this case we have pointed into examples where Russian adverb *совсем* is used on place of native Komi *дзик*, as well as recordings that exhibit insertions of Russian verb forms. The examples are not supposed to be identical, but illustrate that the modification bears some connection to what can be observed in real data. Some of the observed phenomena are relatively rare (although present) in Komi, but are described as common in other Uralic languages, such as Erzya, by Janurik (2017). The presence of

Russian items and their different types in either natural or artificial test corpora does not reflect the frequencies with which they occur larger spoken corpus, as no studies have been conducted that would provide metrics that could be used.

It has to be emphasized that the goal of this exercise has not been to create new data that would be directly useful for any other purposes, but to have a dataset that is comparable to the monolingual test corpus, and would be close enough to realistic phenomena that we observe in spoken data that we can use it to evaluate the parser's behaviour.

One of the available metrics comes from on-going research in which the items of Russian origin have been tagged in different text types. This examination shows that the rate of Russian items was, depending on the speaker, somewhere between 20%-40%, which is similar to the proportions used here.

6 Evaluation Strategy

The results are evaluated according to their LAS and UAS scores, but in order to analyze more precisely how the parser interacts with the constructions specific to Komi and Russian, we have examined some of these constructions in further detail. The recognition accuracy is also calculated separately for each dependency relation type. For evaluation purposes, the languages have been tagged into the misc-field of CONLL-U files, but the parser has not been aware of this information, and it is used only for evaluation.

There is a small portion of tokens occurring in the Komi corpus that are identical in form and function with corresponding Russian items. These are mainly particles and conjunctions. In the misc-field of the test corpus, these items have been classified with the tag "mixed", as their form and function are nearly identical in both languages. In addition to this, the "mixed" category also contains tokens that cannot be clearly defined as lexical items of either Komi or Russian, such as non-adapted Russian verb stems with Komi inflections.

Note that our analysis of a "mixed" category is also in line with the recent sociolinguistic description of similar contact-induced phenomena in Erzya (Janurik, 2017, 64, 89). According to this study, distinguishing between borrowing and code-switching is often very difficult in the case of Erzya and Russian. The same criteria seem to apply with regard to Komi-Russian language contact as well. Recent borrowings not displaying clear Russian morphology have therefore also been tagged as mixed, as the lack of phonological adaptation often makes them identical to the Russian alternatives, and using the Russian origin as the main criteria seems perfectly sensible.

The tokens that are unambiguously Russian and exhibit Russian morphology are tagged as Russian, so that it is possible to compare these parts of the corpus. The percentages of different languages across the testing corpora is as follows in Table 1.

The accuracy is also evaluated independently for a few grammatical structures in which the constituent order or relation type would differ in the two languages. In Komi noun phrases, nouns modify other nouns directly in the nominative, whereas in Russian, this would be accomplished using derived adjectives. In possessive constructions, the languages employ opposite strategies: possessor–possessed in Komi and possessed–possessor in Russian. Due the restrictions on the training data, it would be assumed that the parser would be more sensitive towards the Russian strategies, as the exposure to the Komi patterns has been minimal.

When evaluating the results, the possibility of mistakes remaining in training and

file	corpus	kpv	mixed	rus
kpv-ud-test.conllu	written monolingual	96.2%	3.8%	-
kpv-ud-test-mixed.conllu	written artificially mixed	70.2%	2.3%	27.5%
kpv-ud-ikdp.conllu	spoken	50.2%	9.9%	39.9%

Table 1: The compositional ratio of corpora between Komi (kpv), Russian (rus) and mixed.

testing data itself cannot be excluded. As there are very few annotated datasets for Komi, it is not always perfectly clear what would be the most adequate annotation or relation in every case. Further work with Universal Dependencies on smaller Uralic languages will certainly shed light also into best ways to analyze Komi data.

7 Results

The LAS and UAS scores of the tested corpora are presented in Table 2. The results varied significantly by epoch, and all tests were run for 10 iterations. The differences were particularly large within the spoken corpus, as the parsing accuracy of individual sentences had direct relation to the scores as whole, just because the number of analyzed tokens was so small. Addition of individual sentences would make scores fluctuate very much, whereas other corpora behave more consistently, which indicates that test corpus of approximately hundred sentences in the test corpus seems to be enough for consistency in results.

The test corpora containing more Russian produce slightly better results. The reason seems to be that the parser is more sensitive towards recognizing Russian, as both Russian training corpus and the word embedding used are significantly larger. Indeed, when the parser is run on the identical settings to Russian test corpus, the LAS score is almost 70,00. This happens even under scenario where the parser is specifically targeted to parse Komi, and will first try to look for tokens from Komi part of word embedding. The examination of language-tagged tokens showed that the dependency relation types were analyzed correctly on average 10% more often on Russian tokens than with Komi tokens. The difference in recognizing heads was even higher in favor of Russian. This seems to reflect the generally higher accuracy in respect to Russian, which is explainable by the larger resource portions used in training. On the other hand, preliminary tests done after the research for this paper was conducted indicated that simply building the Komi word embeddings from larger text corpus would improve the monolingual Komi score and bring those closer to one another.

One way to test the cross-linguistic applicability of the parser is to look into constructions that are specific only to one of the languages. Earlier mentioned uses of prepositions and postpositions in Russian and Komi seem to be properly recognized. In the manually mixed test corpus half of the adpositions were in Komi and half in Russian (13/13), and in the best epochs they contained only individual errors. The roots were located correctly 80% if the time. Table 3 presents the accuracy percentages for different dependencies in monolingual and mixed test corpora in . The spoken corpora is not presented here due to its small size and thereby sporadic number of different relations.

Corpus	LAS	UAS
Written corpus	51.34	67.73
Artificially mixed corpus	53.61	65.74
Spoken corpus	54.77	68.20

Table 2: The results of Labeled attachment scores (LAS) and unlabeled attachment scores (UAS) for Komi-Russian code-switching data (Artificially mixed corpus and Spoken corpus) and the regular scenario (only Komi). Komi word embedding size 1,0 million tokens.

deprel	count kpv	correct in kpv	count mixed	correct in mixed
amod	24	95.8%	24	91.7%
case	15	93.3%	25	96%
advmod	91	85.7%	93	89.2%
root	80	80%	80	78.8%
conj	10	70%	10	80%
acl	14	7.14%	14	7.14%
xcomp	21	66.7%	21	71.4%
obj	20	60%	20	60%
cc	23	60.9%	23	56.5%
nsubj	47	57.4%	47	55.3%
mark	7	57.1%	7	71.4%
discourse	9	55.6%	9	66.7%
aux	14	42.9%	12	33.3%
nmod	33	30.3%	38	39.5%
advcl	5	20%	3	0%
ccomp	1	100%	1	100%
appos	6	0%	6	0%
cop	3	0%	3	0%
det	5	0%	5	0%
flat	2	0%	2	0%
iobj	5	0%	4	0%
obl	47	0%	46	19.6%
parataxis	3	0%	3	0%
vocative	1	0%	1	0%
fixed	0	0%	1	0%

Table 3: The comparison to processed results between regular Komi corpus (Komi only) and the code-switching corpus (Artificially mixed corpus) for each dependency relations.

It seems that the rarer dependencies are also generally poorer in their accuracy, with many never being recognized correctly. As this is evaluation of just the best epoch, the accuracy of zero doesn't mean that the parser would never recognize this relation, but the poor accuracy seems to be consistent across tests. This may be connected to the small size of Komi training corpus which contained only 40 sentences, and thereby there are lots of relations which occur only sporadically there as well. However, the table Table 3 also shows that with some relations the accuracy is much

better than for others.

One reason for high accuracy with adpositions could be explained by their very high frequency and relatively small number of distinct forms. Some attention has to be paid into the situation with obliques, which are almost uniformly parsed incorrectly in Komi test corpus. Within the Russian part of the mixed corpus the recognition accuracy increases, and this gain comes from the Russian part. In Komi part of the corpus obliques are most commonly parsed as nominal subjects. Across all training epochs this is most commonly mis-identified relation. Within Russian part the obliques are generally parsed correctly. In case of Russian the obliques are usually marked with prepositions and distinct case such as prepositional or dative.

In the monolingual Komi corpus and in Komi part of the mixed corpora very frequently mis-parsed relation was nominal subjects being analyzed as nominal modifiers. Right after this comes the analysis of nominal objects as nominal subjects.

Some of the results match fit typological differences between Komi and Russian. For example, noun modifiers in certain contexts were recognized much worse than could be expected. Even when constructions share lexical items with Russian, the parser systematically recognizes the first element as the head, probably reflecting Russian pattern where the order would be reversed, or the first component be an adjective and the relation thus *amod* instead of *nmod*.

As mentioned above, the adpositions were generally parsed correctly, irrespective of their language or direction. There were individual Komi postpositions which seemed to be often parsed incorrectly, but these were either used in non-prototypical way or were relatively rare otherwise. So rarer types were recognized worse, which may be related to the general difficulties in recognizing obliques as well, as those have hardly any prototypical form in which they appear in Komi.

8 Conclusion

According to our analysis, the Multilingual BIST-parser described in Lim and Poibeau (2017) is able to parse with comparable accuracy monolingual data and code-switching data. The analysis of parsing result of different dependency relation labels showed that some are recognized considerably more often than others, and especially with rarer relations the accuracy is suffering. There are some relations which show large differences between language pairs used in model training, such as obliques, but also cross-linguistically differently behaving categories, for example adpositions, which are recognized considerably well even when they occur in same sentences in code-switching data.

At the moment the main reason for relatively poor accuracy seems to be a lack of larger training corpus. At the moment the training has been done only with 40 sentences, which is by any standards very little. However, it is so small that comparable dataset could be easily created for virtually any language, and thereby the results are encouraging for extending this approach to new languages. Another aspect that needs more rigorous testing is the alignation and quality of word embeddings used. The currently used Komi embedding was built from one million token text corpus, and possibly an increase in the embedding size could already bring improvements to the performance. On the other hand, also Russian embeddings, although large, are from Wikipedia and could be improved by including wider variety of text types. Evaluating the minimum size that is needed for embeddings is also important in order to estimate how well suited the proposed method is for low-resource languages.

One additional concern is that as all training data and word embeddings are based on written data. Thereby there are many features of spoken language, such as discourse particles, which occur rarely if at all in any of these sources, even when the corpora would be relatively large. Although the discourse particles were in this case analyzed better than majority of the relations, there are still certainly numerous constructions that tend to occur mainly in spoken data. One of these are particular mixed forms which are likely never found in monolingual resources of these two languages, and thereby cannot directly benefit from the method tested in this paper. The IKDP Komi corpus counts approximately 300,000 tokens at present and training new word embeddings from this data alone doesn't seem reasonable right now. However, as regular transcription work increases the corpus size over time, reaching a million or more tokens should be reasonable in the foreseeable future. Meanwhile further experiments should be conducted on building word embeddings that mix spoken and written varieties, and thereby also contain spoken data with code-switching. Naturally, increasing the sizes of training and test corpora for Komi is also a foremost priority for our own future research.

Acknowledgments

We want to express our gratitude to Marina Fedina, Jack Rueter, Trond Trosterud and Rogier Blokland for collaboration and continuous valuable feedback. Thanks are also due to two anonymous peer-reviewers for their useful comments. This work has been developed in the framework of the LAKME project funded by a grant from Paris Sciences et Lettres (IDEX PSL reference ANR-10-IDEX-0001-02). Thierry Poibeau is also partially supported by a RGNF-CNRS (grant between the LATTICE-CNRS Laboratory and the Russian State University for the Humanities in Moscow). This work has been carried out in collaboration with the project "Language Documentation meets Language Technology: The Next Step in the Description of Komi", led by Rogier Blokland, Niko Partanen and Michael Rießler and funded by Kone Foundation.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016a. One parser, many languages. *CoRR* http://arxiv.org/abs/1602.01595.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016b. Massively multilingual word embeddings. *CoRR* http://arxiv.org/abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. www.aclweb.org/anthology/D16-1250.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 451–462. aclweb.org/anthology/P17-1042.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. pages 13–23.

Rogier Blokland, Marina Fedina, Niko Partanen, and Michael Rießler. 2009–2017. Izhva Kyy. In *The Language Archive (TLA)*, Max Planck Institute for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000C-1CF6-F@view.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR* http://arxiv.org/abs/1607.04606.

Kyunghyun Cho. 2015. Natural language understanding with distributed representation. *arXiv preprint* http://arxiv.org/abs/1511.07916.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, pages 20–30. http://www.aclweb.org/anthology/K/K17/K17-3002.pdf.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Association for Computational Linguistics, ACL Anthology, pages 57–66. http://www.aclweb.org/anthology/W17-0109.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology* 4:29–47. https://doi.org/10.3384/nejlt.2000-1533.1643.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI*. pages 2734–2740.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint* http://arxiv.org/abs/1508.01991.

Boglárka Janurik. 2017. *Erzya-Russian bilingual discourse: A structural analysis of intrasentential code-switching patterns*. Ph.D. thesis, University of Szeged. http://doktori.bibl.u-szeged.hu/4097.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies* 1(1):1–127.

KyungTae Lim and Thierry Poibeau. 2017. A system for multilingual dependency parsing based on bidirectional lstm feature representations. In *Proceedings of the*

CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Vancouver, Canada, pages 63–70. http://www.aclweb.org/anthology/K/K17/K17-3006.pdf.

Kai Liu, Yajuan Lü, Wenbin Jiang, and Qun Liu. 2013. Bilingually-guided monolingual dependency grammar induction. In *ACL (1)*. pages 1063–1072.

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 on Language technologies*. Association for Computational Linguistics, pages 1–8. https://doi.org/10.3115/1073336.1073356.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 62–72. https://www.aclweb.org/anthology/D11-1006.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*. pages 92–97. https://www.aclweb.org/anthology/P13-2017.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 629–637.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1234–1244.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0, CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-2184.

Kalika Bali Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. *EMNLP 2014* page 116. https://www.aclweb.org/anthology/W14-3914.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. www.aclweb.org/anthology/W14-3907.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, pages 477–487. www.aclweb.org/anthology/N12-1052.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *COLING*. pages 1854–1864.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Hěctor Martínez Alonso, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.