Error Analysis of Cross-lingual Tagging and Parsing

Rudolf Rosa and **Zdeněk Žabokrtský**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{rosa,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We thoroughly analyse the performance of cross-lingual tagger and parser transfer from English into 32 languages. We suggest potential remedies for identified issues and evaluate some of them.

1 Introduction

In this case study, we try to answer several questions one might have about the performance of cross-lingual tagging and parsing. We do that by extensively evaluating a state-of-the-art cross-lingual setup, with a single source language (English) and 32 target languages.

A researcher in cross-lingual parsing might ask what the strengths and weaknesses of the system are, which information is transferred well from the input knowledge, which information is lost in the transfer, and which information is already missing or confusing on the input – and why that probably is and how this might potentially be addressed.

Furthermore, a user of the cross-lingual parsing, such as a computational linguist interested in utilising the outputs of the cross-lingual parsing in subsequent automatic processing, or a formal linguist interested in the syntax of low-resource languages, may still ask a somewhat different set of questions, such as how trustworthy the outputs of the system are, and how likely to be correct which parts of the outputs are.

We try to answer questions of both of these kinds, analysing errors in cross-lingual parsing along various dimensions. We focus on a state-of-the-art cross-lingual parsing setup, based on translating training data with a 1:1 machine translation (MT) system – this is the approach used in SFNW (Rosa et al., 2017), the winning system of the VarDial cross-lingual parsing shared task (Zampieri et al., 2017).

We make sure our setup is realistic for the supposed low-resource scenario, by only requiring a dependency treebank for a source language (we use English) and source-target parallel data to perform the cross-lingual parser transfer; in particular, we do not assume the availability of a target language tagger (or data to train one), contrary to a lot of previous work in the field.

In practice, significantly better results can be achieved by carefully selecting one or more appropriate source languages for each target language, but this would add too much complexity to our analysis, and we thus leave this for future work. Using a fixed source language makes it easier to generalise in our observations over some or all of the target languages. Moreover, choosing English specifically, which we understand well both theoretically and practically, allows us to perform a more in-depth analysis than with a source language we do only have a limited knowledge of.

Note that we do require supervised target language treebanks to be able to perform the error analysis. However, we hope that our observations can be used to provide a more general insight into the mechanisms of cross-lingual processing, driving intuitions and seeding expectations valid even for languages that we did not cover, thus facilitating a researcher to informedly choose a particular setup for this scenario, knowing what to be careful about and what to expect. We hope this to be especially useful with truly under-resourced target languages, where performing an error analysis of the outputs is costly.

We review previous work in Section 2 and describe our setup in Section 3. We then proceed with error analysis of cross-lingual tagging (Section 4) and parsing (Section 5), evaluate some of our suggested remedies in Section 6, and conclude with Section 7.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 106–118,
Prague, Czech Republic, January 23–24, 2018.

Keywords: cross-lingual parsing, cross-lingual tagging, Universal Dependencies

2 Cross-lingual parsing

Cross-lingual parsing is the task of performing syntactic analysis of a *target* language with no treebank available for that language by using annotated data for a different *source* language and a method for transferring the knowledge about syntactic structures from that source language into the target language. It has already been studied for over a decade, starting with the works of Hwa et al. (2005) and Zeman and Resnik (2008), and then continued by many others, such as McDonald et al. (2011), Täckström et al. (2012), Georgi et al. (2013), Agić et al. (2015), Søgaard et al. (2015), and Duong et al. (2015).

A thorough overview, analysis and comparison of existing methods can be found in (Tiedemann et al., 2016). The authors also include a detailed analysis of the performance of the systems based on various factors, such as part-of-speech (POS) labelling accuracy or size of training data. Another work dealing with error analysis of cross-lingual parsing systems is that of Ramasamy et al. (2014).

The system evaluated in this paper is a new version of the aforementioned SFNW (Rosa et al., 2017), improved and generalised according to our experiments and findings of other researchers, such as Tiedemann (2014). The core of our approach is to translate the source treebank into the target language by a word-by-word statistical MT system (Moses in an adapted setup), resulting in a pseudo-target treebank, which is then used to train a standard tagger and parser. Limiting the MT system in this way leads to a lower quality of the translations, but allows us to use an extremely simple 1:1 cross-lingual transfer strategy. This approach has been shown to achieve results competitive to high quality phrase-to-phrase translation followed by complex many-to-many transfer strategies, as usually done by other authors.

For simplicity, we use a setup with a fixed source language (English) in this work. This allows us to keep the experimental space at a manageable scale, as well as to provide a more in-depth analysis thanks to our knowledge of the shared English source. However, we admit that this also significantly reduces the achieved scores – in practice, one should always carefully select appropriate source language(s) for each target language, as shown e.g. by Rosa and Žabokrtský (2015), or more recently and comprehensively by Agić (2017). Admittedly, the value of our analysis is thus somewhat limited from that perspective.

3 Setup

3.1 Cross-lingual tagger and parser transfer

We use the following approach to obtain a tagger and a parser for the target language t, assuming the availability of a treebank for a source language s (English), and s-t sentence-aligned parallel data:

1. Train a word-based MT system on the parallel data
2. Obtain a synthetic t treebank by translating the words in the s treebank
3. Train a tagger on the t treebank
4. Re-tag the t treebank with the tagger
5. Train a parser on the re-tagged treebank

As the cross-lingual transfer happens already in the training phase, the prediction phase is then trivial:

1. Tag the t text with the t tagger
2. Parse the tagged text with the t parser

We only use the word forms and the POS tags predicted by the tagger, as the other features (lemma, morphological features) are usually too specific for each language and do not transfer well cross-lingually, typically bringing only very moderate improvements or even deteriorations.

We also trained fully supervised monolingual taggers and parsers to provide reference scores; these were trained with the same settings, but using existing target treebanks instead of the synthetic ones.

3.2 Languages and dataset

We used the Universal Dependencies v1.4 treebanks[1] (Nivre et al., 2016) – *train* for training and *dev* for evaluation – and parallel OpenSubtitles2016 data from the Opus collection[2] (Tiedemann, 2012). We used all UD 1.4 languages except for those that had no or too small parallel data (*cop*, *cu*, *ga*, *got*, *grc*, *kk*,

[1] http://universaldependencies.org/docsv1/index.html
[2] http://opus.lingfil.uu.se/

la, *sa*, *swl*, *ta*, *ug*) and those that do not use spaces to separate words (*ja*, *zh*), thus limiting ourselves to 32 target languages.[3] For the analysis, we sorted and grouped the languages into three groups according to cross-lingual tagging accuracy. A detailed overview of the languages and datasets can be found in Table 4 in the Appendix; a brief overview of the emergent language groups follows:

High *pt*, *no*, *it*, *fr*, *da*, *de*, *sv*
European languages closely related to English, from the Germanic and Romance language families, with sufficient parallel data to provide high-quality machine translation, and thus high accuracy in cross-lingual tagging and parsing.

Med *bg*, *ca*, *gl*, *nl*, *sk*, *cs*, *ru*, *id*, *el*, *hr*, *ro*, *pl*, *et*, *lv*, *sl*
Mostly European languages from the Indo-European family (with the exception of *id* and *et*) which are more distant from English and/or lower on parallel data, but still achieving competitive translation quality and mediocre accuracy of cross-lingual methods.

Low *fi*, *he*, *hi*, *uk*, *tr*, *ar*, *fa*, *vi*, *eu*, *hi*
Distant non-European or non-Indo-European languages (with the exception of *uk*, which is extremely low on parallel data), achieving very low quality of both MT and cross-lingual methods.

3.3 Tools

We used the following tools in the cross-lingual analysis pipeline in the following ways:

- a rule-based Treex tokenizer[4] (Popel and Žabokrtský, 2010) to tokenize the parallel data,
- UDPipe tagger and parser bundle[5] (Straka et al., 2016) to train the taggers and parsers,
- word2vec[6] (Mikolov et al., 2013) to pre-compute target word embeddings for the parser,
- MGiza[7] to compute intersection-symmetrized word alignment links (`-alignment intersect`),
- Moses SMT system[8] (Koehn et al., 2007) to translate the treebank data, constrained to perform word-to-word translation with no reordering (`-max-phrase-length 1 -dl 0`),
- KenLM language model (Heafield et al., 2013) as a component of Moses.

Our source codes are freely available on GitHub,[9] containing both the cross-lingual parsing pipeline, as well as evaluation scripts which can produce detailed accuracy breakdowns along various dimensions for both tagging and parsing and which provided data for the tables in this paper.

To manually inspect the CoNLLU files, we used the `conll_view` tool (Rosa, 2017).

4 Tagging error analysis

As parsing heavily depends on the UPOS tags, we will first analyse errors in tagging. Cross-lingual Universal POS (UPOS) tagging accuracies for several most frequent UPOS tags are shown in Table 1. For an interested reader, a larger table can be found in the Appendix (Table 5), showing UPOS tagging accuracies for all UPOS tags, as well as most common errors in cross-lingual tagging together with their frequencies. However, the presented analysis is also based on other, more detailed numbers, which are not shown here for space reasons, as well as on direct inspection of the inputs and outputs in some cases.

Note that we are mainly interested in tagging as a pre-processing step for parsing – achieving high-quality tagging is expected to improve the parsing quality, but is not our primary goal in itself.

4.1 Nouns

Cross-lingual tagging of both common nouns (*NOUN*) and proper nouns (*PROPN*) is very successful, with accuracies usually notably above the average across all language groups – a noun in one language seems to usually correspond 1:1 to a noun in the other language, making nouns highly suitable for the selected lexical transfer method.

[3] This was done mostly for simplicity – *ja* and *zh* tokenizers do exists and/or can be trained, and *some* parallel data could presumably be found even for the omitted languages; we leave re-including those languages for future work.
[4] https://github.com/ufal/treex/blob/master/lib/Treex/Block/W2A/Tokenize.pm
[5] http://ufal.mff.cuni.cz/udpipe
[6] https://code.google.com/archive/p/word2vec/
[7] https://github.com/moses-smt/mgiza
[8] http://www.statmt.org/moses/
[9] https://github.com/ptakopysk/crosssynt

Setup	Languages	all	*NOUN*	*VERB*	*PRON*	*ADP*	*DET*	*PROPN*	*ADJ*	*ADV*	*AUX*
Cross-lingual	Low	58%	63%	55%	57%	57%	59%	61%	34%	39%	38%
	Med	73%	79%	74%	57%	75%	61%	78%	51%	56%	52%
	High	82%	86%	81%	73%	87%	80%	75%	62%	60%	70%
Supervised	English	94%	93%	95%	98%	97%	99%	85%	89%	88%	97%

Table 1: Macro-averaged tagging accuracy of the cross-lingual tagging, factored along gold UPOS tags (only several most frequent shown) and language groups; also listing the fully supervised English tagger.

The most common error in tagging of nouns is mistaking one of the types for the other (*NOUN* for *PROPN* or *PROPN* for *NOUN*) – specifically, 30% of words predicted to be *PROPN*s are actually *NOUN*s, which is a rather high error rate. Many of these errors happen at the sentence-initial word, in parts of titles, and at nouns that are capitalised in English (months, days of the week, titles) – these could probably be at least partially avoided by truecasing the data.

As the capitalisation of *PROPN*s is an important feature for the tagger, we saw a huge drop in *PROPN* tagging accuracy for German (capitalises all nouns) and Hindi (does not capitalise anything). For such languages, it might make sense to abandon the *NOUN*/*PROPN* distinction (as is common in other tagsets), leading to a less granular but more accurate tagging which the parser could better rely on; a new feature could be added to the parser input capturing information about the casing of the word (e.g. lowercase/uppercase/capitalised/mixed) so that this information is not lost.

4.2 Adjectives

The overall most frequent error is an adjective (*ADJ*) confused for *NOUN*. This seems to be mostly caused by the fact that in English, *NOUN*s are often used as adjectives – as in e.g. "fruit salad", where the noun "fruit" in this context would be expressed by an adjective in many languages. Because of that, the translation of the treebank often contains much noise in the form of adjectives labelled as nouns, hence the error.

Other than choosing a different source language which does not have this property, one could try to alleviate this problem by e.g. identifying such cases in the source data and forcibly relabelling them with the UPOS of the expected translation; or, more straightforwardly, by simply removing all sentences containing such trap words. As suggested by Reviewer 2, even a more fine-grained approach could be used, by only deleting the confusing adjective-like nouns but keeping the modified sentences in the training data. We note that although this problem seems to be rather specific for English, similar confusing situations with words of unclear POS exist in other languages.

Moreover, *ADJ*s perform particularly badly in target languages with the *NOUN ADJ* word order, with all Romance languages (*pt*, *it*, *fr*, *ca*, *gl*, *ro*) constituting a prominent example – if the error distribution is computed only on Romance languages, only 40% of *ADJ* labels get actually assigned to *ADJ*s, while 45% of words predicted to be *ADJ*s are actually *NOUN*s or *PROPN*s. This shows the tremendous importance of word order for tagging. Primarily, one should try to use a source language with similar word order to the target language. Otherwise, it may be possible to handle these cases by employing a reordering model within the MT system (which we explicitly disallowed in our setup), or by pre-reordering the source sentences to resemble more closely the target word order, as done e.g. by Aufrant et al. (2016). A simpler but potentially interesting approach could also be to modify the word order randomly, by locally shuffling parts of the sentences, thus making the tagger more robust to differences in word order.

4.3 Verbs

Auxiliaries (*AUX*) are often confused with verbs (*VERB*), with the accuracy on *AUX* quite low even for many of the High group languages (with the exception of the Romance languages), and falling quickly for the other language groups. As different languages use different verbs as auxiliaries and in different ways, they get very easily mistranslated by the MT system.

Of course, as always, one should choose a source language that uses *AUX*es in the same way as the target language. However, if this is not possible, it may help to discard the *VERB*/*AUX* distinction and label everything as *VERB*s. This theoretically means loosing some information, but, looking at the accuracies of *AUX* tagging, in many cases the information is already lost anyway. On the other hand, it could make the subsequent parser more robust and thus more successful than a parser that learns to trust the *AUX* labels.

Furthermore, some languages do not seem to use auxiliaries much (or at all). In such cases (as in all cases where a source data label is not relevant for the target language), the cross-lingual parsing might be improved by deleting the *AUX* tokens from the source data altogether.

4.4 Pronouns, Determiners and Adpositions

Pronouns (*PRON*) seem to be rather difficult, with a very low accuracy even in the High languages, as even similar languages tend to use pronouns differently (this may still partially be due to unresolved inter-lingual annotation inconsistencies).

A common error is confusing *PRON*s with determiners (*DET*) both ways, especially in languages where the same word form can be used both as a *DET* and as a *PRON* (e.g. *fr*, *it*). We believe that it may help to relabel all *DET*s as *PRON*s in such cases, thus postponing the decision to parsing.

Another frequent error is related to reflexive pronouns, which are very common in many languages but not very prominent in English, leading to misalignments, mistranslations, and then mistaggings – e.g. the reflexive pronoun in the target language gets often aligned to an *AUX* in English (which may or may not be appropriate). We have also noticed frequent mistranslations of English *PRON*s with pro-drop target languages; again, this time the source *PRON* gets typically aligned to some other word, such as an *AUX* (which, again, might be the best thing to do in some cases, but not always).

If a source language matching in the aforementioned characteristics cannot be used, it may be possible to modify the source to correspond better to the target. However, these cases clearly show the limitations of the selected word-by-word MT approach, in contrast to the classical phrase-based one, which inherently learns to add/remove words that do not have a proper counterpart in the other language by using variable-length phrases, and thus should suffer from such problems much less.

Tagging of adpositions (*ADP*) is relatively accurate, but they are sometimes confused for *DET*s; this happens more often in languages that are low on *DET*s (e.g. Russian), where the word aligner is likely to misalign one of the *DET*s that are abundant in English onto a target *ADP*. In such cases, it might be beneficial to remove some of the *DET*s from the source data – e.g. “a” and “the” if the target language does not use similar determiners – but keep the other *DET*s (“this”, “some”, etc.). Still, in some target languages, *DET*s seem to be so rare (or possibly even non-existent) that the cross-lingual parsing might by improved by simply deleting all *DET* tokens from the source data.

5 Parsing errors analysis

Labelled Attachment Scores (LAS) for several most frequent dependency relation labels are shown in Table 2. For an interested reader, a larger table can be found in the Appendix (Table 6), showing accuracies for more labels, as well as most common labelling errors together with their frequencies.

The least frequent dependency relations are not included in any of the tables, as the evaluation results have little meaning there – mostly the scores are computed over very small numbers of instances, and the measured accuracies are thus rather random numbers. A general remark regarding the low-frequency labels is that they mostly should not be trusted, as even the parser has very little training support for them. It is definitely worth considering to remove them altogether from the training data in the cross-lingual scenario, replacing them by some more general relations (even *dep*), as with the accuracy of the cross-lingual parsing as low as it is, these come out mostly as random noise.

5.1 Nouns

With nouns, the dependency relation (usually *nmod*, *compound*, *nsubj*, or *dobj*) is often incorrectly distinguished. It should be noted that for other parts of speech, it is usually easier to correctly identify the relation label than the head – the label is often determined by the POS already, sometimes including some simple local context. For nouns, however, the situation is different, as there are 4 very common

Setup	Langs	ALL	*punct*	*nmod*	*case*	*nsubj*	*det*	*root*	*dobj*	*compound*	*advmod*	*amod*
Cross-lingual	Low	20%	28%	8%	21%	21%	36%	35%	10%	16%	17%	21%
	Med	34%	38%	22%	48%	32%	46%	55%	32%	15%	33%	41%
	High	51%	49%	45%	75%	48%	66%	63%	49%	23%	41%	55%
Supervised	English	80%	75%	74%	92%	87%	95%	88%	84%	74%	74%	83%

Table 2: Macro-averaged LAS of the cross-lingual parsing, factored along gold-standard dependency relations (only several most frequent shown) and language groups; also including LAS for the fully supervised English parser.

dependency labels, and they can be very hard to correctly identify, especially in a cross-lingual setting – different languages use different means of distinguishing them (e.g. word order, adpositions, determiners, or morphology), and so they are often mislabelled even when the head is identified correctly. For languages from the High group, the problem is not that severe, since they mostly use similar distinguishing features as English; however, we observe a huge drop in accuracy when moving to the Med group, and we even see low results for some of the High languages, such as German.

Detailed investigation showed that the most frequent mistake is mislabelling an *nmod* relation as a *compound*. *Nmod*s in English are nearly always marked by adpositions (as in "the house **of** the lady"), while a sequence of nouns without a preposition is typically a *compound* (as in "investment firm"). However, many languages (e.g. German) use case marking for *nmod*s, where the case may be expressed e.g. by a determiner (as in "das Haus **der** Frau") – which, due to an adposition not being present, the parser usually mislabels as a *compound*. Most of the noun labelling errors are actually *compound*s mislabelled as other relations, or other relations mislabelled as *compound*s. What hugely adds to this is the fact that the *compound* relation is much more frequent in English than in most of the target languages, where it is usually rare or not present (again, this may partially be an inter-lingual inconsistency of the annotation). Due to this, it may be sensible to either relabel the *compound*s as other relations (presumably *nmod*s, which they are on average most frequently confused with), or delete the *compound* tokens from the source data altogether. While this would inevitably cut the *compound*-labelling accuracy to zero, it may still increase the overall parsing accuracy thanks to the rareness of this label in most target languages.

Other labels get frequently confused as well, such as switching *nsubj* and *dobj*, especially in languages which mark the subject and object morphologically rather than with word order.

Thus, it seems highly important when choosing a source language for a given target language to observe the way they mark noun-based relations and the way they join together chains of nouns, as the mismatches in this aspect led to the largest number of errors on our dataset.

Moreover, *amod*s also get often mislabelled as *compound*s, due to the difficulty in correctly identifying the *NOUN* or *ADJ* category when translating from English, as explained in Section 4.2.

Furthermore, the parsing of *PROPN*s also shows very low accuracies across all of the language groups. However, this seems to be at least partially caused by inter-treebank annotation inconsistencies, as the v1 of the UD guidelines seems not to have been explicit enough in the correct way of annotating names (later noting e.g. that "The *name* label is another one that has led to confusion."). Therefore, UD decided to redesign name annotation in UD v2, as explained online,[10] which will hopefully suppress this problem significantly.

However, a real problem with *PROPN*s in the source data remains that they are necessarily often unknown to the MT system and thus remain untranslated in the training treebank, which may confuse the subsequent tools. It is therefore probably worth considering to pre-process the data in some way. One option would be to replace the specific names (which are bound to be unknown to all the tools) by some generic placeholders (which the tools can be trained to be able to process), provided this can be done on the target side as well (e.g. using cross-lingual or language-independent named entity recognisers). A slightly different approach could be to replace uncommon names with more common ones (so e.g. we could rename "Pervaiz Musharraf" and "Velupillai Prabhakaran" to "John Smith" and "Martin Jones").

[10] http://universaldependencies.org/v2/semantic-categories.html

Experiment	Low group		Med group		High group		All languages	
Base		19.6%		34.1%		51.2%		33.3%
NOUN+PROPN	4/10	-0.6%	6/15	-0.2%	2/7	-0.4%	12/32	-0.4%
VERB+AUX	7/10	0.0%	10/15	0.3%	2/7	0.0%	19/32	0.1%
PRON+DET	6/10	-0.3%	9/15	0.1%	3/7	-0.2%	18/32	-0.1%
nmod+compound	5/10	0.8%	9/15	0.8%	4/7	-0.1%	18/32	0.6%
Reordering	6/10	1.0%	2/15	-3.7%	0/7	-10.4%	8/32	-3.7%

Table 3: Number of target languages for which improvement was observed and absolute improvement in macro-averaged LAS when various modifications are applied, as compared to Base (Table 2).

5.2 Easy regular phenomena

Unsurprisingly, phenomena that behave quite regularly – *case*, *nummod*, *punct*, *det*, *amod*, *advmod* – are rather easy to parse correctly, as long as they bear the correct POS tag. As explained in Section 4, correctly tagging some of them is often tricky, especially with *amod* (*ADJ* tag), *advmod* (*ADV* tag), and *det* (*DET* tag); however, if their tagging succeeds, it is usually not difficult for the parser to identify the correct head for them, and to identify the correct dependency relation label is mostly trivial. In particular, the *amod* accuracies are quite low for Romance languages, which prefer the *NOUN ADJ* order.

As could be expected, the head assignment accuracy for the *case* relation drops near zero for target languages that strongly prefer postpositions while the source language strongly prefers prepositions. This is manifested by the relatively very low *case* accuracy for the Low language group, which contains several such languages.

As already discussed in Section 4.2, the problems related to differences in word order may be solvable by employing a reordering component, either before or during the translation.

5.3 Verbs

In general, parsing of *VERB*s is quite successful over all language groups. However, the auxiliary verbs (*aux*, *cop*) are only parsed well in the High group, i.e. in languages with sufficiently similar grammar (the ideal source language should use auxiliary verbs similarly to the target language).

Moreover, clausal relations (*advcl*, *acl*, *xcomp*, *ccomp*) are very hard to get right, even for the High languages (and often even for a fully supervised parser) – both in assigning the correct head, as they tend to form long-distance relations, as well as in assigning the correct label, as all of these are frequently confused for each other. Thus, these should not be trusted much on the output of cross-lingual parsing.

6 Preliminary experiments

Implementing, fine-tuning and evaluating all of the modifications of the base approach that we suggest would clearly be beyond the scope of this work. Nevertheless, we include at least a brief experimental part, evaluating the effects of several of the suggested modifications – merging a pair of UPOS labels (*NOUN+PROPN*, *VERB+AUX*, *PRON+DET*), merging a pair of dependency relation labels (*nmod* and *compound*),[11] and allowing reordering in Moses.[12] Note that these are rather preliminary results, without the usual several iterations of experimentation and evaluation.

Table 3 shows the number of languages for which LAS improved when the modifications were applied, and the average improvement/deterioration in LAS for each language group.

We see that even the very noisy *PROPN* signal from the tagger is useful for the parser, probably because the main distinguishing feature (capitalization) is not directly available to the parser, and it thus cannot easily make the distinction itself. We thus believe that other approaches are to be tried out, such as truecasing the data and/or explicitly including information about the casing into the parser input.

Merging the other label pairs usually behaved quite expectedly, slightly improving the results for the low and med groups, but not for the high group. The results for merging of *DET* and *PRON* are rather

[11]The labels were not merged in the test data – the parser is still "expected" by the evaluator to output the *compound* label.
[12]We used the setting recommended in the documentation (`-reordering msd-bidirectional-fe`). Moses decoding was set to output the word alignment (`-alignment-output-file file.a`), which was used to correctly transfer the annotations.

mixed, as the language groups do not sufficiently differentiate the usage of determiners in the target language; one should be more careful when deciding whether to merge these labels or not. The very frequent *compound* label, on the other hand, is something very specific for English, while in most target languages it is rare or non-existent; thus, removing it helped even for many languages in the high group.

Surprisingly, enabling reordering in Moses led to deteriorations (often large) in LAS for all languages, except for a few of the most dissimilar ones (8/32), even though the BLEU score actually improved in most cases (24/32). This clearly requires a thorough further investigation, as our previous experiments (unpublished) indicated a positive correlation between BLEU and LAS. Based on a quick inspection of the data, we currently hypothesise that disallowing reordering forces the MT system to produce more literal translations, which better preserve the sentence structure (POS and dependency relations).

7 Conclusion

We thoroughly analysed a particular cross-lingual tagging and parsing setup, investigating the behaviour of the tools factored along labels and language groups.

We found that the properties of the source and target language have a huge impact on the way the tools work and the kinds of errors we encounter. It is not surprising that best results are obtained when the source and target languages are close. However, we believe it is not straightforward to determine which aspects of the language similarity will have what effect on the analysis of which language phenomena; here, we see the value of our work.

In particular, we saw a high importance of grammatical similarity, especially in terms of word order and auxiliary words usage, such as auxiliary verbs, determiners, pronouns, and adpositions. Except for adpositions, the interlingual variation in usage of the auxiliaries often causes severe problems already in the translation step, with the auxiliaries being frequently misaligned, then necessarily mistranslated, and subsequently mishandled by the tagger and parser.

We spent much of our analyses with understanding the errors that revolve around nouns. However, it seems that the nouns themselves do not cause the problems; it is rather the words around them (especially the auxiliaries), which different languages use differently to mark the roles fulfilled by the nouns.

The question of the word order similarity is less subtle – we clearly saw well-known word order patterns, such as *ADJ NOUN* vs *NOUN ADJ*, or prepositions vs postpositions, to cause severe drops in accuracy in case of a mismatch of the preferred word order between the source and target language.

We hope that this analysis can be used to provide more insight into cross-lingual tagging and parsing, and to help develop better-performing cross-lingual tools in future.

Acknowledgments

This work has been supported by the grant No. DG16P02B048 of the Ministry of Culture of the Czech Republic, the grant No. CZ.02.1.01/0.0/0.0/16_013/0001781 of the Ministry of Education, Youth and Sports of the Czech Republic, the SVV 260 453 grant, and the grant 15-10472S of the Czech Science Foundation. This work has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

We would also like to thank the anonymous reviewers for many helpful observations and suggestions.

References

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 *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*. Hrvatska znanstvena bibliografija i MZOS-Svibor. http://aclweb.org/anthology/P15-2044.

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*. Linköping University Electronic Press, Linköpings universitet, 135, pages 1–10. http://aclweb.org/anthology/W17-0401.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 119–130. http://aclweb.org/anthology/C16-1012.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Beijing, China, pages 113–122. http://www.aclweb.org/anthology/K15-1012.

Ryan Georgi, Fei Xia, and William D. Lewis. 2013. Enhanced and portable dependency projection algorithms using interlinear glossed text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 306–311. http://www.aclweb.org/anthology/P13-2055.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 690–696. http://www.aclweb.org/anthology/P13-2121.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering* 11:11–311. https://doi.org/10.1017/S1351324905003840.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/P07-2045.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. https://www.aclweb.org/anthology/D11-1006.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*. https://arxiv.org/abs/1301.3781.

Joakim Nivre et al. 2016. Universal dependencies 1.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1827.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Proceedings of IceTAL, 7th International Conference on Natural Language Processing, Reykjavík, Iceland, August 17, 2010*. Springer, pages 293–304. https://doi.org/10.1007/978-3-642-14770-8_33.

Loganathan Ramasamy, David Mareček, and Zdeněk Žabokrtský. 2014. Multilingual dependency parsing: Using machine translated texts instead of parallel corpora. *The Prague Bulletin of Mathematical Linguistics* 102:93–104. http://ufal.mff.cuni.cz/pbml/102/art-ramasamy-marecek-zabokrtsky.pdf.

Rudolf Rosa. 2017. Terminal-based CoNLL-file viewer, v2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2514.

Rudolf Rosa, Daniel Zeman, David Mareček, and Zdeněk Žabokrtský. 2017. Slavic Forest, Norwegian Wood. In Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali, editors, *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial4)*. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pages 210–219. http://www.aclweb.org/anthology/W17-1226.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KL_{cpos^3} - a language similarity measure for delexicalized parser transfer. 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)*. Association for Computational Linguistics, Beijing, China, pages 243–249. http://www.aclweb.org/anthology/P15-2040.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*. http://www.aclweb.org/anthology/P15-1165.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Paris, France. http://www.lrec-conf.org/proceedings/lrec2016/pdf/873_Paper.pdf.

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, Stroudsburg, PA, USA, NAACL HLT '12, pages 477–487. http://www.aclweb.org/anthology/N12-1052.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*. pages 2214–2218. http://lrec.elra.info/proceedings/lrec2012/pdf/463_Paper.pdf.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pages 1854–1864. http://www.aclweb.org/anthology/C14-1175.

Jörg Tiedemann, Željko Agić, et al. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* 55:209–248. http://dx.doi.org/10.1613/jair.4785.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain. http://aclweb.org/anthology/W17-1201.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Workshop on NLP for Less-Privileged Languages, IJCNLP*. Hyderabad, India. http://www.aclweb.org/anthology/I08-3008.

A Detailed Evaluation Results

Target group	Target language iso	Target language name	Para data (en tokens)	MT BLEU	Treebank tokens train	Treebank tokens dtest	UPOS acc x-ling	UPOS acc sup	LAS x-ling	LAS sup
Low	hi	Hindi	321,339	7.3%	281,057	35,217	48.5%	96.4%	10.4%	86.7%
	eu	Basque	1,082,072	6.1%	72,974	24,095	49.5%	94.1%	10.2%	63.4%
	vi	Vietnamese	13,582,467	8.8%	31,799	6,093	51.3%	88.2%	21.5%	52.0%
	fa	Farsi	23,653,954	1.3%	121,064	15,832	55.0%	96.7%	14.1%	79.2%
	ar	Arabic	149,458,897	3.7%	225,853	28,263	56.5%	95.7%	14.0%	72.5%
	tr	Turkish	234,219,925	3.6%	40,617	8,852	59.8%	93.4%	13.8%	69.5%
	uk	Ukrainian	3,797,579	1.4%	1,281	241	62.7%	68.0%	35.7%	31.5%
	hu	Hungarian	215,222,322	8.1%	33,016	4,781	63.0%	94.2%	21.3%	72.4%
	he	Hebrew	156,340,612	22.0%	135,496	11,234	66.0%	95.4%	28.5%	76.4%
	fi	Finnish	133,830,769	4.0%	162,721	9,161	66.3%	94.5%	26.8%	73.1%
		Average	93,150,994	6.6%	110,588	14,377	57.9%	91.7%	19.6%	67.7%
		Std. dev.	94,197,934	6.0%	92,061	11,313	6.7%	8.6%	8.6%	15.7%
Med	sl	Slovenian	106,842,127	11.5%	112,334	14,021	68.8%	95.0%	33.5%	80.3%
	lv	Latvian	2,548,465	7.3%	13,083	3,640	70.7%	91.2%	24.1%	57.4%
	et	Estonian	64,034,502	10.3%	187,814	22,867	71.6%	94.6%	29.4%	72.8%
	pl	Polish	183,401,406	8.7%	69,499	6,887	71.9%	95.3%	37.9%	80.0%
	ro	Romanian	249,781,321	16.3%	163,262	27,965	72.0%	96.8%	32.3%	76.1%
	hr	Croatian	174,234,575	18.6%	127,894	4,823	72.8%	98.0%	34.4%	78.9%
	el	Greek	205,382,482	13.4%	47,449	6,039	73.1%	97.9%	46.4%	77.5%
	id	Indonesian	31,382,075	16.5%	97,531	12,612	73.7%	93.3%	24.3%	72.0%
	ru	Russian	117,951,946	10.2%	79,772	10,044	73.9%	95.7%	30.4%	74.3%
	cs	Czech	217,464,167	10.2%	1,173,282	159,284	74.1%	98.3%	32.6%	79.7%
	sk	Slovak	44,334,287	11.5%	80,575	12,440	74.1%	94.1%	39.4%	75.6%
	nl	Dutch	197,441,086	20.5%	197,134	6,434	74.8%	94.3%	41.5%	74.1%
	gl	Galician	1,106,922	12.1%	79,329	29,777	75.2%	97.2%	18.9%	77.6%
	ca	Catalan	2,513,413	11.9%	429,157	58,020	75.6%	98.0%	41.2%	80.1%
	bg	Bulgarian	214,756,441	11.2%	124,474	16,111	76.3%	97.7%	45.2%	82.8%
		Average	120,878,348	12.7%	198,839	26,064	73.2%	95.8%	34.1%	75.9%
		Std. dev.	90,308,798	3.7%	286,467	39,435	2.0%	2.1%	8.0%	6.0%
High	sv	Swedish	81,231,502	12.7%	66,645	9,797	79.1%	95.0%	47.5%	72.9%
	de	German	88,261,445	15.9%	269,626	12,348	80.6%	90.1%	47.4%	76.2%
	da	Danish	73,620,273	15.0%	88,979	5,870	81.2%	95.5%	50.7%	74.1%
	fr	French	221,712,167	18.3%	356,419	38,758	81.2%	97.1%	51.8%	83.8%
	it	Italian	172,151,250	13.0%	270,598	10,921	81.8%	97.3%	51.4%	83.7%
	no	Norwegian	37,362,647	22.0%	243,887	36,369	83.3%	97.0%	58.6%	82.3%
	pt	Portuguese	160,033,555	14.7%	216,001	5,124	83.4%	96.7%	51.0%	81.9%
		Average	119,196,120	15.9%	216,022	17,027	81.5%	95.5%	51.2%	79.3%
		Std. dev.	66,022,248	3.2%	103,918	14,283	1.5%	2.5%	3.7%	4.7%
All		Average	111,845,562	11.5%	175,019	20,435	70.2%	94.5%	33.3%	74.1%
		Std. dev.	85,249,035	5.6%	208,818	28,439	9.9%	5.3%	13.6%	10.6%
Source	en	English			204,586	25,148		94.3%		79.6%

Table 4: List of all target languages divided into the three groups, reporting their source-target parallel data size (number of tokens in the English side of the parallel data), treebank size (number of tokens in training and development test set of the treebank), translation quality (BLEU measured on the last 10,000 sentences held out from the parallel data), UPOS accuracy and Labelled Attachment Score (for both cross-lingual and fully supervised monolingual tagging and parsing).
Averages are also included, together with standard deviations to illustrate the variance in the data.
The last line lists some of this information for the source language (English).

Gold tag	**Actual predicted tag**							
NOUN	75.5%	NOUN	8.0%	PROPN	6.7%	VERB	4.6%	ADJ
VERB	69.6%	VERB	12.0%	NOUN	6.2%	AUX	3.6%	ADJ
PUNCT	94.7%	PUNCT	2.2%	CONJ	0.9%	DET	0.6%	SYM
PRON	60.3%	PRON	9.9%	DET	4.4%	AUX	4.2%	VERB
ADP	72.0%	ADP	7.8%	DET	3.6%	PART	3.5%	NOUN
DET	65.2%	DET	16.3%	PRON	6.3%	ADJ	3.9%	ADP
PROPN	72.2%	PROPN	16.0%	NOUN	2.8%	PRON	2.6%	ADJ
ADJ	48.4%	ADJ	25.3%	NOUN	8.7%	VERB	6.8%	PROPN
ADV	52.3%	ADV	8.9%	NOUN	8.8%	ADJ	6.3%	VERB
AUX	52.3%	AUX	20.7%	VERB	8.9%	PRON	4.4%	NOUN
CONJ	78.0%	CONJ	4.5%	ADV	3.8%	SCONJ	2.6%	ADP
PART	32.3%	PART	17.7%	ADV	11.9%	PRON	9.2%	DET
NUM	79.1%	NUM	5.9%	DET	5.5%	NOUN	3.6%	ADJ
SCONJ	39.3%	SCONJ	14.7%	PRON	10.5%	ADP	8.8%	DET
X	33.3%	NOUN	27.1%	PROPN	7.4%	X	6.5%	ADP
INTJ	29.9%	INTJ	20.8%	NOUN	16.9%	ADV	11.0%	PROPN
SYM	36.7%	SYM	29.2%	PUNCT	25.0%	NOUN	3.0%	PROPN

Predicted tag	**Actual gold tag**							
NOUN	75.7%	NOUN	7.8%	ADJ	6.7%	VERB	4.1%	PROPN
VERB	66.8%	VERB	13.4%	NOUN	5.4%	ADJ	4.7%	AUX
PUNCT	96.7%	PUNCT	0.6%	AUX	0.5%	ADP	0.5%	VERB
PRON	56.2%	PRON	11.8%	DET	5.5%	SCONJ	4.9%	AUX
ADP	74.8%	ADP	3.7%	ADV	3.6%	VERB	3.3%	DET
DET	45.9%	DET	16.1%	ADP	10.0%	PRON	3.9%	VERB
PROPN	54.5%	PROPN	29.6%	NOUN	7.4%	ADJ	2.3%	VERB
ADJ	56.1%	ADJ	18.3%	NOUN	7.3%	VERB	6.0%	ADV
ADV	53.1%	ADV	8.7%	NOUN	7.7%	ADJ	5.3%	PART
AUX	34.1%	AUX	33.9%	VERB	8.9%	PRON	7.0%	PART
CONJ	88.0%	CONJ	3.9%	PUNCT	2.1%	SCONJ	2.0%	ADV
PART	31.2%	ADP	23.9%	PART	11.4%	ADV	9.1%	VERB
NUM	77.1%	NUM	6.5%	ADJ	6.3%	NOUN	3.4%	PROPN
SCONJ	38.4%	SCONJ	21.7%	ADP	10.3%	CONJ	8.2%	ADV
X	31.1%	NOUN	16.3%	NUM	12.8%	PROPN	9.1%	VERB
INTJ	19.7%	ADV	15.0%	NOUN	13.7%	PROPN	13.6%	VERB
SYM	35.1%	PUNCT	22.1%	NOUN	19.5%	SYM	4.3%	PRON

Table 5: Error distribution in cross-lingual UPOS tagging, each row listing an UPOS tag and the four most common tags found with it (i.e. usually showing the three most common errors), macro average over all target languages. The rows are ordered by the frequency of the UPOS tags in the English treebank.

Gold label	Actual predicted label							
punct	94.6%	punct	2.3%	cc	0.9%	case	0.8%	det
nmod	43.3%	nmod	19.0%	compound	8.3%	dobj	6.0%	nsubj
case	72.3%	case	7.7%	det	5.7%	mark	2.8%	advmod
nsubj	45.8%	nsubj	12.4%	compound	9.8%	dobj	5.6%	nmod
det	61.7%	det	10.4%	nmod	7.6%	amod	4.9%	nsubj
root	50.4%	root	5.7%	nsubj	4.3%	acl	4.2%	nmod
dobj	36.4%	dobj	12.9%	nmod	12.5%	compound	10.5%	nsubj
compound	23.6%	compound	17.5%	nmod	9.4%	nummod	9.2%	case
advmod	48.9%	advmod	6.0%	amod	6.0%	case	5.3%	nmod
amod	48.1%	amod	13.0%	compound	9.7%	nmod	4.3%	dobj
conj	50.4%	conj	9.0%	compound	5.8%	acl	5.5%	amod
mark	46.2%	mark	12.8%	case	9.0%	nsubj	6.8%	det
cc	82.8%	cc	3.4%	advmod	2.2%	case	1.9%	det
aux	45.1%	aux	8.1%	nsubj	6.9%	cop	6.4%	mark
cop	52.3%	cop	7.4%	aux	5.8%	root	5.5%	auxpass
advcl	33.7%	advcl	7.9%	root	7.7%	acl	6.3%	amod
acl	34.7%	acl	10.2%	amod	7.6%	advcl	6.7%	root
xcomp	16.4%	xcomp	13.3%	root	9.4%	ccomp	8.3%	dobj
nummod	73.4%	nummod	5.8%	det	5.3%	compound	4.5%	nmod
ccomp	22.8%	ccomp	10.3%	acl	9.9%	advcl	6.9%	root
neg	69.4%	neg	11.9%	nsubj	3.8%	aux	2.8%	punct
appos	22.2%	appos	17.5%	compound	12.2%	nmod	10.8%	name
Predicted label	**Actual gold label**							
punct	96.0%	punct	0.5%	nmod	0.5%	case	0.3%	auxpass
nmod	56.9%	nmod	7.0%	dobj	4.8%	amod	4.1%	det
case	72.4%	case	3.7%	nmod	3.7%	det	3.6%	advmod
nsubj	42.1%	nsubj	14.3%	nmod	7.6%	dobj	4.4%	root
det	49.4%	det	16.1%	case	4.2%	nmod	3.8%	mark
root	54.7%	root	6.6%	nmod	5.5%	nsubj	3.0%	amod
dobj	34.2%	dobj	22.6%	nmod	11.6%	nsubj	4.7%	amod
compound	37.4%	nmod	12.9%	amod	9.5%	nsubj	8.1%	dobj
advmod	53.5%	advmod	7.1%	nmod	5.3%	case	4.8%	amod
amod	49.0%	amod	11.8%	nmod	5.0%	det	4.8%	advmod
conj	46.3%	conj	11.3%	nmod	4.7%	amod	4.3%	dobj
mark	43.4%	mark	22.7%	case	6.7%	advmod	4.5%	aux
cc	84.9%	cc	4.3%	punct	3.1%	advmod	1.3%	discourse
aux	36.5%	aux	12.3%	root	7.2%	advmod	4.7%	cop
cop	47.6%	cop	11.6%	aux	9.2%	root	3.6%	auxpass
advcl	21.6%	advcl	11.5%	nmod	10.2%	root	8.6%	acl
acl	30.5%	acl	11.8%	root	10.1%	nmod	7.8%	conj
xcomp	17.2%	xcomp	10.0%	nmod	9.7%	root	8.4%	amod
nummod	59.8%	nummod	11.5%	nmod	7.4%	amod	2.2%	conj
ccomp	24.9%	ccomp	12.1%	root	8.5%	xcomp	7.5%	acl
neg	64.6%	neg	8.1%	advmod	4.9%	aux	3.9%	cop
appos	29.3%	nmod	12.6%	appos	11.0%	name	7.0%	nsubj

Table 6: Error distribution in cross-lingual parsing, each row listing a relation label and the four most common labels found with it (i.e. usually showing the three most common errors), reporting macro average of dependency relation label assignment over all target languages (disregarding the head assignment, i.e. this is not LAS). The rows are ordered by the frequency of the relations in the English treebank, and only the most frequent are included in this table.