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 informally 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.
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) – \textit{train} for training and \textit{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 (\textit{cop}, \textit{cu}, \textit{ga}, \textit{got}, \textit{grc}, \textit{kk}, \textit{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. 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
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.

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

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 PROPNs are actually NOUNs, 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 PROPNs 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, NOUNs 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, ADJs 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 ADJs, while 45% of words predicted to be ADJs are actually NOUNs or PROPNs. 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 AUXes 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 VERBS. This theoretically means losing 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 PRONS 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 DETs as PRONS 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 PRONs 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 DETs; this happens more often in languages that are low on DETs (e.g. Russian), where the word aligner is likely to misalign one of the DETs that are abundant in English onto a target ADP. In such cases, it might be beneficial to remove some of the DETs from the source data – e.g. “a” and “the” if the target language does not use similar determiners – but keep the other DETs (“this”, “some”, etc.). Still, in some target languages, DETs 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
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 lan-
guages 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. Nmods 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 nmods, 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 compounds
mislabelled as other relations, or other relations mislabelled as compounds. 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 compounds as other relations (presumably
nmods, 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, amods also get often mislabelled as compounds, 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 PROPNs 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 PROPNs 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”).

---

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

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

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 VERBs 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), \(^\text{11}\) and allowing reordering in Moses. \(^\text{12}\) 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


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

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

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.
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.