# Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets

**Željko Agić**
zagic@uni-potsdam.de

**Jörg Tiedemann**
jorg.tiedemann@lingfil.uu.se

**Kaja Dobrovoljc**
kaja.dobrovoljc@trojina.si

**Simon Krek**
simon.krek@ijs.si

**Danijela Merkler**
dmerkler@ffzg.hr

**Sara Može**
s.moze@wlv.ac.uk

## Abstract

This paper addresses cross-lingual dependency parsing using rich morphosyntactic tagsets. In our case study, we experiment with three related Slavic languages: Croatian, Serbian and Slovene. Four different dependency treebanks are used for monolingual parsing, direct cross-lingual parsing, and a recently introduced cross-lingual parsing approach that utilizes statistical machine translation and annotation projection. We argue for the benefits of using rich morphosyntactic tagsets in cross-lingual parsing and empirically support the claim by showing large improvements over an impoverished common feature representation in form of a reduced part-of-speech tagset. In the process, we improve over the previous state-of-the-art scores in dependency parsing for all three languages.

## 1 Introduction

A large majority of human languages are under-resourced in terms of text corpora and tools available for applications in natural language processing (NLP). According to recent surveys (Bender, 2011; Uszkoreit and Rehm, 2012; Bender, 2013), this is especially apparent with syntactically annotated corpora, i.e., treebanks – both dependency-based ones and others. In this paper, we focus on dependency parsing (Kübler et al., 2009), but the claims should hold in general. The lack of dependency treebanks is due to the fact that they are expensive and time-consuming to construct (Abeillé, 2003). Since dependency parsing of under-resourced languages nonetheless draws substantial interest in the NLP research community, over time, we have seen a number of research efforts directed towards their processing despite the absence of training data for supervised learning of parsing models. We give a brief overview of the major research directions in the following subsection. Here, we focus on supervised learning of dependency parsers, as the performance of unsupervised approaches still falls far behind the state of the art in supervised parser induction.

### 1.1 Related Work

There are two basic strategies for data-driven parsing of languages with no dependency treebanks: annotation projection and model transfer. Both fall into the general category of cross-lingual dependency parsing as they attempt to utilize existing dependency treebanks or parsers from a resource-rich language *(source)* for parsing the under-resourced *(target)* language.

**Annotation projection:** In this approach, dependency trees are projected from a source language to a target language using word alignments in parallel corpora. It is based on a presumption that source-target parallel corpora are more readily available than dependency treebanks. The approach comes in two varieties. In the first one, parallel corpora are exploited by applying the available state-of-the-art parsers on the source side and subsequent projection to the target side using word alignments and heuristics for resolving possible link ambiguities (Yarowsky et al., 2001; Hwa et al., 2005). Since dependency parsers typically make heavy use of various morphological and other features, the apparent benefit of this approach is the possibility of straightforward projection of these features, resulting in a feature-rich representation for the target language. On the downside, the annotation projection noise adds up to dependency parsing noise and errors in word alignment, influencing the quality of the resulting target language parser.

The other variety is rare, since it relies on parallel corpora in which the source side is a depen-



*Language Technology for Closely Related Languages and Language Variants (LT4CloseLang)*, pages 13–24,
October 29, 2014, Doha, Qatar. 

dency treebank, i.e., it is already manually annotated for syntactic dependencies (Agić et al., 2012). This removes the automatic parsing noise, while the issues with word alignment and annotation heuristics still remain.

**Model transfer:** In its simplest form, transferring a model amounts to training a source language parser and running it directly on the target language. It is usually coupled with delexicalization, i.e., removing all lexical features from the source treebank for training the parser (Zeman and Resnik, 2008; McDonald et al., 2013). This in turn relies on the same underlying feature model, typically drawing from a shared part-of-speech (POS) representation such as the Universal POS Tagset of Petrov et al. (2012). Negative effects of using such an impoverished shared representation are typically addressed by adapting the model to better fit the target language. This includes selecting source language data points appropriate for the target language (Søgaard, 2011; Täckström et al., 2013), transferring from multiple sources (McDonald et al., 2011) and using cross-lingual word clusters (Täckström et al., 2012). These approaches need no projection and enable the usage of source-side gold standard annotations, but they all rely on a shared feature representation across languages, which can be seen as a strong bottleneck. Also, while most of the earlier research made use of heterogenous treebanks and thus yielded linguistically implausible observations, research stemming from an uniform dependency scheme across languages (De Marneffe and Manning, 2008; McDonald et al., 2013) made it possible to perform more consistent experiments and to assess the accuracy of dependency labels.

**Other approaches:** More recently, Durrett et al. (2012) suggested a hybrid approach that involves bilingual lexica in cross-lingual phrase-based parsing. In their approach, a source-side treebank is adapted to a target language by "translating" the source words to target words through a bilingual lexicon. This approach is advanced by Tiedemann et al. (2014), who utilize full-scale statistical machine translation (SMT) systems for generating synthetic target language treebanks. This approach relates to annotation projection, while bypassing the issue of dependency parsing noise as gold standard annotations are projected. The SMT noise is in turn mitigated by better word alignment quality for synthetic data. The influence of various projection algorithms in this approach is further investigated by Tiedemann (2014). This line of cross-lingual parsing research substantially improves over previous work.

### 1.2 Paper Overview

All lines of previous cross-lingual parsing research left the topics of related languages and shared rich feature representations largely unaddressed, with the exception of Zeman and Resnik (2008), who deal with phrase-based parsing test-cased on Danish and Swedish treebanks, utilizing a mapping over relatively small POS tagsets.

In our contribution, the goal is to observe the properties of cross-lingual parsing in an environment of relatively free-word-order languages, which are related and characterized by rich morphology and very large morphosyntactic tagsets. We experiment with four different small- and medium-size dependency treebanks of Croatian and Slovene, and cross-lingually parse into Croatian, Serbian and Slovene. Along with monolingual and direct transfer parsing, we make use of the SMT framework of Tiedemann et al. (2014). We are motivated by:

- observing the performance of various approaches to cross-lingual dependency parsing for closely related languages, including the very recent treebank translation approach by Tiedemann et al. (2014);
- doing so by using rich morphosyntactic tagsets, in contrast to virtually all other recent cross-lingual dependency parsing experiments, which mainly utilize the Universal POS tagset of Petrov et al. (2012);
- reliably testing for labeled parsing accuracy in an environment with heterogenous dependency annotation schemes; and
- improving the state of the art for Croatian, Slovene and Serbian dependency parsing across these heterogenous schemes.

In Section 2, we describe the language resources used: treebanks, tagsets and test sets. Section 3 describes the experimental setup, which includes a description of parsing, machine translation and annotation projection. In Section 4, we discuss the results of the experiments, and we conclude the discussion by sketching the possible directions for future research in Section 5.

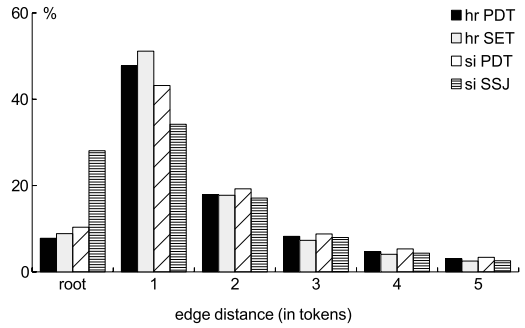


Figure 1: Histogram of edge distances in the treebanks. Edge distance is measured in tokens between heads and dependents. Distance of 1 denotes adjacent tokens.

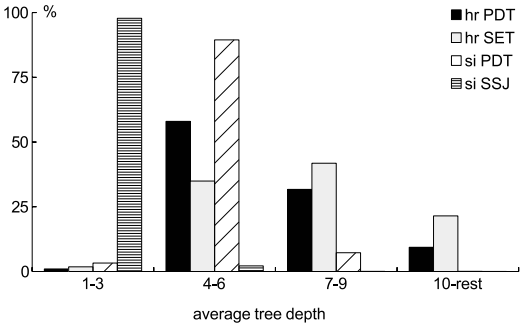


Figure 2: Histogram of average tree depths.

## 2 Resources

We make use of the publicly available language resources for Croatian, Serbian and Slovene. These include dependency treebanks, test sets annotated for morphology and dependency syntax, and a morphosyntactic feature representation drawing from the Multext East project (Erjavec, 2012). A detailed assessment of the current state of development for morphosyntactic and syntactic processing of these languages is given by Agić et al. (2013) and Uszkoreit and Rehm (2012). Here, we provide only a short description.

### 2.1 Treebanks

We use two Croatian and two Slovene dependency treebanks.[1] One for each language is based on the Prague Dependency Treebank (PDT) (Böhmová et al., 2003) annotation scheme, while the other two introduced novel and more simplified syntactic tagsets. All four treebanks use adaptations of the Multext East version 4 tagset (Erjavec, 2012) for the underlying morphological annotation layer, which we shortly describe further down. Basic statistics for the treebanks are given in Table 1.

| Feature | *hr* PDT | *hr* SET | *sl* PDT | *sl* SSJ |
|---|---|---|---|---|
| Sentences | 4,626 | 8,655 | 1,534 | 11,217 |
| Tokens | 117,369 | 192,924 | 28,750 | 232,241 |
| Types | 25,038 | 37,749 | 7,128 | 48,234 |
| Parts of speech | 13 | 13 | 12 | 13 |
| MSDs | 821 | 685 | 725 | 1,142 |
| Syntactic tags | 26 | 15 | 26 | 10 |

Table 1: Basic treebank statistics.

***hr* PDT:** This treebank is natively referred to as the Croatian Dependency Treebank (HOBS) (Tadić, 2007; Berović et al., 2012). Its most recent instance, HOBS 2.0 (Agić et al., 2014) slightly departs from the PDT scheme. Thus, in this experiment, we use the older version, HOBS 1.0, and henceforth refer to it as *hr* PDT for consistency and more clear reference to its annotation.[2]

***hr* SET:** The SETIMES.HR dependency treebank of Croatian has a 15-tag scheme. It is targeted towards high parsing accuracy, while maintaining a clear distinction between all basic grammatical categories of Croatian. Its publicly available 1.0 release consists of approximately 2,500 sentences (Agić and Merkler, 2013), while release 2.0 has just under 4,000 sentences (Agić and Ljubešić, 2014) of newspaper text. Here, we use an even newer, recently developed version with more than 8,500 sentences from multiple domains.[3]

***sl* PDT:** The PDT-based Slovene Dependency Treebank (Džeroski et al., 2006) is built on top of a rather small portion of Orwell's novel *1984* from the Multext East project (Erjavec, 2012). Even if the project was discontinued, it is still heavily used as part of the venerable CoNLL 2006 and 2007 shared task datasets (Buchholz and Marsi, 2006; Nivre et al., 2007).[4]

***sl* SSJ:** The Slovene take on simplifying syntactic annotations resulted in the 10-tag strong JOS Corpus of Slovene (Erjavec et al., 2010). Similar to *hr* SET, this new annotation scheme is loosely

[1] No treebanks of Serbian were publicly available at the time of conducting this experiment.

[2] HOBS is available through META-SHARE (Tadić and Váradi, 2012).

[3] http://nlp.ffzg.hr/resources/corpora/setimes-hr/

[4] http://nl.ijs.si/sdt/

PDT-based, but considerably reduced to facilitate manual annotation. The initial 100,000 token corpus has recently doubled in size, as described by Dobrovoljc et al. (2012). We use the latter version in our experiment.[5]

The statistics in Table 1 show a variety of treebank sizes and annotations. Figure 1 illustrates the structural complexity of the treebanks by providing a histogram of egdes by token distance. While adjacent edges expectedly dominate the distributions, it is interesting to see that almost 30% of all edges in *sl* SSJ attach to root, resulting in an easily parsable flattened tree structure. Knowing that relations denoting attributes account for more than one third of all non-root dependents in the remainder, one can expect dependency parsing performance comparable to CoNLL-style chunking (Tjong Kim Sang and Buchholz, 2000). This is further supported by the distributions of sentences in the four treebanks by average tree depth in Figure 2. We can see that virtually all *sl* SSJ trees have average depths of 1 to 3, while the other treebanks exhibit the more common structural properties of dependency trees.

In these terms of complexity, the Croatian treebanks are richer than their Slovene counterparts. In *sl* SSJ, attributes and edges to root account for more than 60% of all dependencies. Even in the other three treebanks, 20-30% of the edges are labeled as attributes, while the rest is spread more evenly between the basic syntactic categories such as predicates, subject and objects. More detailed and more linguistically motivated comparisons of the three annotation guidelines fall outside the scope of our paper. Instead, we refer to the previously noted publications on the respective treebanks, and to (Agić and Merkler, 2013; Agić et al., 2013) for comparisons between PDT and SET in parsing Croatian and Serbian.

### 2.2 Morphosyntactic Tagset

All four treebanks were manually created: they are sentence- and token-split, lemmatized, morphosyntactically tagged and syntactically annotated. In morphosyntactic annotation, they all make use of the Multext East version 4 (MTE 4) guidelines (Erjavec, 2012).[6] MTE 4 is a positional tagset in which morphosyntactic descriptors of word forms are captured by a morphosyntactic tag (MSD) created by merging atomic attributes in the predefined positions. This is illustrated in Table 2 through an example verb tag. The first character of the tag denotes the part of speech (POS), while each of the following characters encodes a specific attribute in a specific position. Both the positions and the attributes are language-dependent in MTE 4, but the attributes are still largely shared between these three languages due to their relatedness.

The Slovene treebanks closely adhere to the specification, while each of the Croatian treebanks implements slight adaptations of the tagset towards Croatian specifics. In *hr* PDT, the adaptation is governed by and documented in the Croatian Morphological Lexicon (Tadić and Fulgosi, 2003), and the modifications in *hr* SET were targeted to more closely match the ones for Slovene.[7]

### 2.3 Test Sets

Recent research by McDonald et al. (2013) has uncovered the downsides of experimenting with parsing using heterogenous dependency annotations, while at the same time providing possibly the first reliable results in cross-lingual parsing. They did so by creating the uniformly annotated Universal Dependency Treebanks collection based on Stanford Typed Dependencies (De Marneffe and Manning, 2008), which in turn also enabled measuring both labeled (LAS) and unlabeled (UAS) parsing accuracy.

Having four treebanks with three different annotation schemes, we seek to enable reliable experimentation through our test sets. Along with Croatian and Slovene, which are represented in the training sets, we introduce Serbian as a target-only language in the test data. Following the CoNLL shared tasks setup (Buchholz and Marsi, 2006; Nivre et al., 2007), our test sets have 200 sentences (approx. 5,000 tokens) per language, split 50:50 between newswire and Wikipedia text. Each test set is manually annotated for morphosyntax, following the MTE 4 guidelines for the respective languages, and checked by native speakers for validity. On top of that, all test sets are annotated with all three dependency schemes: PDT, SET and SSJ. This enables observing LAS in a heterogenous experimental environment, as we test each monolingual and cross-lingual parser on an anno-

[5] http://eng.slovenscina.eu/tehnologije/ucni-korpus

[6] http://nl.ijs.si/ME/V4/

[7] http://nlp.ffzg.hr/data/tagging/msd-hr.html

| Language | MSD tag | Attribute-value pairs |
|---|---|---|
| *hr* | `Vmn` | Category = **V**erb, Type = **m**ain, Vform = i**n**finitive |
| *sl* | `Vmen` | Category = **V**erb, Type = **m**ain, Aspect = p**e**rfective, VForm = i**n**finitive |
| *sr* | `Vmn----an-n---e` | Category = **V**erb, Type = **m**ain, VForm = i**n**finitive, Voice = **a**ctive, Negative = **n**o, Clitic = **n**o, Aspect = p**e**rfective |

Table 2: Illustration of the Multext East version 4 tagset for Croatian, Serbian and Slovene. The attributes are language-dependent, as well as their positions in the tag, which are also dependent on the part of speech, denoted by position zero in the tag.

tation layer matching its training set. In contrast, the MTE 4 tagsets are not adjusted, i.e., each test set only has a single language-specific MTE 4 annotation. We rely on their underlying similarities in feature representations to suffice for improved cross-lingual parsing performance.

## 3 Experiment Setup

This section describes the experiment settings. We list the general workflow of the experiment and then provide the details on the parser setup and the more advanced approaches used for target language adaptation of the models.

### 3.1 Workflow

The experiment consists of three work packages: (1) monolingual parsing, (2) direct cross-lingual parsing, and (3) cross-lingual parsing using synthetic training data from SMT. In the first one, we train dependency parsers on the four treebanks and test them on the corresponding languages, thus assessing the monolingual parsing performance. The second stage observes the effects of directly applying the parsers from the first stage across the languages. Finaly, in the third work package, we use four different approaches to automatic translation to create synthetic training data. We translate the Croatian treebanks to Slovene and vice versa, project the annotations using two different projection algorithms, and train and apply the adapted parsers across the languages. The details are included in the two following subsections.

Two general remarks apply to our experiment. First, we perform cross-lingual parsing, and not cross-annotation-scheme parsing. Thus, we do not compare the dependency parsing scores between the annotation schemes, but rather just between the in-scheme parsers. Second, we use Serbian as a test-set-only language. As there are no treebanks of Serbian, we cannot use it as a source language, and we leave SMT and annotation projection into Serbian for future work.

### 3.2 Dependency Parsing

In all experiments, we use the graph-based dependency parser by Bohnet (2010) with default settings. We base our parser choice on its state-of-the-art performance across various morphologically rich languages in the SPMLR 2013 shared task (Seddah et al., 2013). While newer contributions targeted at joint morphological and syntactic analysis (Bohnet and Kuhn, 2012; Bohnet et al., 2013) report slightly higher scores, we chose the former one for speed and robustness, and because we use gold standard POS/MSD annotations. The choice of gold standard preprocessing is motivated by previous research in parsing Croatian and Serbian (Agić et al., 2013), and by insight of Seddah et al. (2013), who report a predictable linear decrease in accuracy for automatic preprocessing. This decrease amounts to approximately 3 points LAS for Croatian and Serbian across various test cases in (Agić et al., 2013).

We observe effects of (de)lexicalization and of using full MSD tagset as opposed to only POS tags in all experiments. Namely, in all work packages, we compare parsers trained with {lexicalized, delexicalized} × {MSD, POS} features. In lexicalized parsers, we use word forms and features, while we exclude lemmas from all experiments – both previous research using MSTParser (McDonald et al., 2005) and our own test runs show no use for lemmas as features in dependency parsing. Delexicalized parsers are stripped of all lexical features, i.e., word forms are omitted from training and testing data. Full MSD parsers use both the POS information and the sub-POS features in the form of atomic attribute-value pairs, while POS-only parsers are stripped of the MSD features – they use just the POS information. The delexicalized POS scenario is thus very similar to the

direct transfer by McDonald et al. (2013), since MTE 4 POS is virtually identical to Universal POS (Petrov et al., 2012).[8]

### 3.3 Treebank Translation and Annotation Projection

For machine translation, we closely adhere to the setup implemented by Tiedemann et al. (2014) in their treebank translation experiments. Namely, our translations are based on automatic word alignment and subsequent extraction of translation equivalents as common in phrase-based SMT. We perform word alignment by using GIZA++ (Och and Ney, 2003), while utilizing IBM model 4 for creating the Viterbi word alignments for parallel corpora. For the extraction of translation tables, we use the de facto standard SMT toolbox Moses (Koehn et al., 2007) with default settings. Phrase-based SMT models are tuned using minimum error rate training (Och, 2003). Our monolingual language modeling using KenLM tools[9] (Heafield, 2011) produces standard 5-gram language models using modified Kneser-Ney smoothing without pruning.

For building the translation models, we use the OpenSubtitles parallel resources from OPUS[10] (Tiedemann, 2009) for the Croatian-Slovene pair. Even if we expect this to be a rather noisy parallel resource, we justify the choice by (1) the fact that no other parallel corpora[11] of Croatian and Slovene exist, other than Orwell's *1984* from the Multext East project, which is too small for SMT training and falls into a very narrow domain, and (2) evidence from (Tiedemann et al., 2014) that the SMT-supported cross-lingual parsing approach is very robust to translation noise.

For translating Croatian treebanks into Slovene and vice versa, we implement and test four different methods of translation. They are coupled with approaches to annotation projection from the source side gold dependency trees to the target translations via the word alignment information available from SMT.

**LOOKUP:** The first approach to translation in our experiment is the dictionary lookup approach. We simply select the most reliable translations of single words in the source language into the target language by looking up the phrase translation tables extracted from the parallel corpus. This is very similar to what Agić et al. (2012) did for the Croatian-Slovene pair. However, their approach involved both translating and testing on the same small corpus (Orwell's novel), while here we extract the translations from full-blown SMT phrase tables on a much larger scale. The trees projection from source to target is trivial since the number and the ordering of words between them does not change. Thus, the dependencies are simply copied.

**CHAR:** By this acronym, we refer to an approach known as character-based statistical machine translation. It is shown to perform very well for closely related languages (Vilar et al., 2007; Tiedemann, 2012; Tiedemann and Nakov, 2013). The motivation for character-level translation is the ability of such models to better generalize the mapping between similar languages especially in cases of rich productive morphology and limited amounts of training data. With this, character-level models largely reduce the number of out-of-vocabulary words. In a nutshell, our character-based model performs word-to-word translation using character-level modeling. Similar to LOOKUP, this is also a word-to-word translation model, which also requires no adaptation of the source dependency trees – they are once again simply copied to target sentences.

**WORD:** Our third take on SMT is slightly more elaborate but still restricts the translation model to one-to-one word mappings. In particular, we extract all single word translation pairs from the phrase tables and apply the standard beam-search decoder implemented in Moses to translate the original treebanks to all target languages. Thus, we allow word reordering and use a language model while still keeping the projection of annotated data as simple as possible. The language model may influence not only the word order but also the lexical choice as we now allow multiple translation options in our phrase table. Also note that this approach may introduce additional non-projectivity in the projected trees. This system is the overall top-performer in (Tiedemann et al.,

[8] A mapping from Slovene MTE 4 to Universal POS is available at `https://code.google.com/p/universal-pos-tags/` as an example.

[9] `https://kheafield.com/code/kenlm/`

[10] `http://opus.lingfil.uu.se/`

[11] We note the Croatian-Slovene parallel corpus project described by Požgaj Hadži and Tadić (2000), but it appears that the project was not completed and the corpus itself is not publicly available.

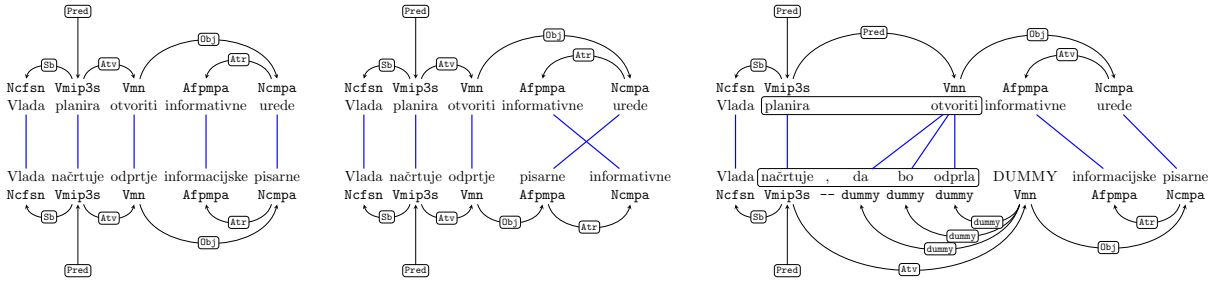


Figure 3: An illustration of the projections. Left side = CHAR, middle = WORD, right side = PHRASE. As illustrated, WORD might introduce reorderings, while PHRASE can enter dummy nodes and edges to the dependency trees. The sentence: *The government plans to open information offices.* See (Tiedemann et al., 2014; Tiedemann, 2014) for detailed insight into projection algorithms.

2014), where reordering played an important role in adapting the models to the target languages. We test whether it holds for related languages as well.

**PHRASE:** This model implements translation based on the entire phrase table using the standard approach to phrase-based SMT. We basically run the Moses decoder with default settings and the parameters and models trained on our parallel corpus. Here, we can have many-to-many word alignments, which require a more elaborate approach to the projection of the source side dependency annotations. It is important for the annotation transfer to keep track of the alignment between phrases and words of the input and output sentences. The Moses decoder provides both, phrase segmentation and word alignment. We use the annotation projection algorithm of Hwa et al. (2005). As illustrated in Figure 3, it resolves many-to-many alignments by introducing dummy nodes to the dependency trees. We use the implementation by Tiedemann (2014), which addresses certain issues with algorithm choices for ambiguous alignments which were left unaccounted for in the original work. Since this paper does not focus on the intricacies of annotation projection, but rather on applying it in an environment of related languages and rich MSD tagsets, we refer the reader to related work regarding the details.

We translate from Croatian to Slovene and vice versa using four different treebanks and these four different methods of translation and annotation projection. As we stated in the experiment overview, for each of these, we also experiment with (de)lexicalization and MSD vs. POS, and we test on all three languages. The three experimental batches – monolingual, direct and SMT-supported transfer – produce a large number of observations, all of which we assess in the following section.

## 4 Results and Discussion

We split our discussion of the parsing results into the following three subsections. We first observe the performance of monolingual parsers. Secondly, we measure the quality of these when applied directly on the other two languages. Finally, we look into the accuracy of parsers trained on SMT-generated artificial treebank data when applied across the test languages.

### 4.1 Monolingual Parsing

Accuracies of parsers trained and applied on training and testing data belonging to the same language – i.e., our monolingual parsers – are provided in the grayed out sections of Table 3.

Parsing Croatian using *hr* PDT yields a high score of 69.45 LAS, better than the former state of the art on this test set (Agić et al., 2013) simply due to applying a newer generation parser. This score is provided by a lexicalized model with the full MSD feature set. Replacing MSD with POS or delexicalizing this model results in a 3-point drop in LAS, while applying both replacements substantially decreases the score – by more than 11 points LAS. We observe virtually the same pattern for the other Croatian treebank, *hr* SET, where this latter drop is even more significant, at 14 points. Incidentally, 76.36 points LAS is also the new state of the art for *hr* SET parsing, owing to the recent enlargement of the treebank.

The Slovene parsers exhibit effectively the same behavior as the Croatian ones. The lexicalized MSD models of *sl* PDT and *sl* SSJ both record new state-of-the-art scores, although the latter one on a different test set than in previous research (Dobrovoljc et al., 2012). At over 92 points LAS, *sl* SSJ

| | | *lexicalized* | | | | | | *delexicalized* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *hr* | | *sl* | | *sr* | | *hr* | | *sl* | | *sr* | |
| | | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* |
| *hr* | PDT | 69.45 | 66.95 | 60.09 | 50.19 | 69.42 | 66.96 | 66.03 | 57.79 | 57.98 | 42.66 | 66.79 | 57.41 |
| | SET | 76.36 | 73.02 | 68.65 | 59.52 | 76.08 | 73.37 | 72.52 | 62.31 | 68.16 | 55.17 | 72.71 | 62.04 |
| *sl* | PDT | 51.19 | 47.99 | 76.46 | 73.33 | 52.46 | 49.64 | 49.58 | 42.59 | 71.96 | 62.99 | 50.41 | 44.11 |
| | SSJ | 78.50 | 74.18 | 92.38 | 88.93 | 78.94 | 75.96 | 75.23 | 66.23 | 87.19 | 77.92 | 75.25 | 67.47 |

Table 3: Monolingual and direct cross-lingual parsing accuracy, expressed by the labeled accuracy metric (LAS). Scores are split for lexicalized and delexicalized, full MSD and POS only parsers. Monolingual scores are in grey. Row indices represent source languages and treebanks.

expectedly shows to be the easiest to parse, most likely due to the relatively flat tree structure and its small label set.

We note the following general pattern of feature importance. Dropping MSD features seems to carry the most weight in all models, followed by lexicalization. Dropping MSD is compensated in part by lexical features paired with POS, while dropping both MSD and word forms severely degrades all models. At this point, it is very important to note that at 60-70 points LAS, these decreased scores closely resemble those of McDonald et al. (2013) for the six languages in the Universal Treebanks. This observation is taken further in the next subsection.

### 4.2 Direct Cross-lingual Parsing

The models used for monolingual parsing are here directly applied on all languages but the treebank source language, thus constituting a direct cross-lingual parsing scenario. Its scores are also given in Table 3, but now in the non-grey parts.

Croatian models are applied to Slovene and Serbian test sets. For *hr* PDT, the highest score is 60.09 LAS on Slovene and 69.42 LAS on Serbian, the latter noted as the state of the art for Serbian PDT parsing. Comparing the cross-lingual score to monolingual Slovene, the difference is substantial as expected and comparable to the drops observed by McDonald et al. (2013) in their experiments. Our ranking of feature significance established in the monolingual experiments holds here as well, or rather, the absolute differences are even more pronounced. Most notably, the difference between the lexicalized MSD model and the delexicalized POS model is 17 points LAS in favor of the former one on Slovene. *hr* SET appears to be more resilient to delexicalization and tagset reduction when applied on Slovene and Serbian, most likely due to the treebank's size, well-balanced dependency label set and closer conformance with the official MTE 4 guidelines. That said, the feature patterns still hold. Also, 76.08 LAS for Serbian is the new state of the art for SET parsing.

Slovene PDT is an outlier due to its small size, as its training set is just over 1,500 sentences. Still, the scores maintain the level of those in related research, and the feature rankings hold. Performance of parsing Croatian and Serbian using *sl* SSJ is high, arguably up to the level of usability in down-stream applications. These are the first recorded scores in parsing the two languages using SSJ, and they reach above 78 points LAS for both. Even if the scores are not comparable across the annotation schemes due to their differences, it still holds that the SSJ scores are the highest absolute parsing scores recorded in the experiment. This might hold significance in applications that require robust parsing for shallow syntax.

Generally, the best transfer scores are quite high in comparison with those on Universal Treebanks (McDonald et al., 2013; Tiedemann et al., 2014). This is surely due to the relatedness of the three languages. However, even for these arguably closely related languages, the performance of delexicalized models that rely only on POS features – averaging at around 55 points LAS – is virtually identical to that on more distant languages test-cased in related work. We see this as a very strong indicator of fundamental limitations of using linguistically impoverished shared feature representations in cross-lingual parsing.

### 4.3 Cross-lingual Parsing with Treebank Translation

Finally, we discuss what happens to parsing performance when we replace direct cross-lingual application of parsers with training models on translated treebanks. We take a treebank, Croatian or Slovene, and translate it into the other language.

| Target | Approach | PDT | SET | SSJ |
|---|---|---|---|---|
| *hr* | monolingual | 69.45 | 76.36 | – |
| | direct | 51.19 | – | 78.50 |
| | translated | 67.55 ♡ | 74.68 ♢ | 79.51 ♣ |
| *sl* | monolingual | 76.46 | – | 92.38 |
| | direct | 60.09 | 68.65 | – |
| | translated | 72.35 ♠ | 70.52 ♠ | 88.71 ♣ |
| *sr* | monolingual | – | – | – |
| | direct | 69.42 | 76.08 | 78.94 |
| | translated | 68.11 ♣ | 74.31 ♢ | 79.81 ♡♣ |
| Legend: | ♠ CHAR ♡ LOOKUP ♢ PHRASE ♣ WORD | | | |

Table 4: Parsing score (LAS) summary for the top-performing systems with respect to language and approach to parser induction. All models are MSD + lexicalized.

We then train a parser on the translation and apply it on all three target test sets. We do this for all the treebanks, and in all variations regarding translation and projection methods, morphological features and lexicalization.

All scores for this evaluation stage are given in Table 5 for completeness. The table contains 192 different LAS scores, possibly constituting a tedious read. Thus, in Table 4 we provide a summary of information on the top-performing parsers from all three experimental stages, which includes treebank translation.

We can see that the best models based on translating the treebanks predominantly stem from word-to-word SMT, i.e., from WORD translation models that basically enrich the lexical feature space and perform word reordering, enabling straightforward copying of syntactic structures from translation sources to translation targets. Following them are the CHAR and LOOKUP models, expectedly leaving – although not too far behind – PHRASE behind given the similarities of the language pair. Since Croatian and Slovene are related languages, the differences between the models are not as substantial as in (Tiedemann et al., 2014), but WORD models still turn out to be the most robust ones, even if word reordering might not be so frequent in this language pair as in the data from (McDonald et al., 2013). Further, when comparing the best SMT-supported models to monolingual parsers, we see that the models with translation come really close to monolingual performance. In comparison with direct transfer, models trained on translated treebanks manage to outperform them in most cases, especially for the more distant language pairs. For example, the *sl* $\mapsto$ *hr* SSJ WORD model is 1 point LAS better on Croatian than the directly applied Slovene model, and the same holds for testing on Serbian with the same dataset. On the other side, directly applied models from Croatian SET outperform the translated ones for Serbian. For PDT, the translated models are substantially better between Croatian and Slovene since *sl* PDT is an outlier in terms of size and dataset selection, while direct transfer from Croatian seems to work better for Serbian than the translated models.

Reflecting on the summary in Table 4 more generally, by and large, we see high parsing accuracies. Averages across the formalisms reach well beyond 70 points LAS. We attribute this to the relatedness of the languages selected for this case study, as well as to the quality of the underlying language resources. From another viewpoint, the table clearly shows the prominence of lexical and especially rich morphosyntactic tagset features throughout the experiment. Across our monolingual, direct and SMT-supported parsing experiments, these features are represented in the best systems, and dropping them incurs significant decreases in accuracy.

## 5 Conclusions and Future Work

In this contribution, we addressed the topic of cross-lingual dependency parsing, i.e., applying dependency parsers from typically resource-rich source languages to under-resourced target languages. We used three Slavic languages – Croatian, Slovene and Serbian – as a test case for related languages in different stages of language resource development. As these are relatively free-word-order languages with rich morphology, we were able to test the cross-lingual parsers for performance when using training features drawing from large morphosyntactic tagsets – typically consisting of over 1,000 different tags – in contrast to impoverished common part-of-speech representations. We tested monolingual parsing, direct cross-lingual parsing and a very recent promising approach with artificial creation of training data via machine translation. In the experiments, we observed state-of-the-art results in dependency parsing for all three languages. We strongly argued and supported the case for using common rich representations of morphology in dependency parsing

| | | | *lexicalized* | | | | | | *delexicalized* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *hr* | | *sl* | | *sr* | | *hr* | | *sl* | | *sr* | |
| | | | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* | *MSD* | *POS* |
| CHAR | *hr* ↦ *sl* | PDT | 66.92 | 60.25 | 61.49 | 55.57 | 67.83 | 62.04 | 66.56 | 57.63 | 58.34 | 43.04 | 66.89 | 57.65 |
| | | SET | 73.65 | 64.64 | 70.52 | 66.11 | 72.95 | 64.44 | 72.98 | 62.98 | 69.03 | 54.81 | 72.74 | 62.73 |
| | *sl* ↦ *hr* | PDT | 51.96 | 48.14 | 72.35 | 63.71 | 53.11 | 49.47 | 49.58 | 42.59 | 71.96 | 62.99 | 50.41 | 44.11 |
| | | SSJ | 78.69 | 75.45 | 88.21 | 78.88 | 79.25 | 77.09 | 75.23 | 66.23 | 87.19 | 77.92 | 75.25 | 67.47 |
| LOOKUP | *hr* ↦ *sl* | PDT | 67.55 | 59.96 | 60.81 | 56.54 | 67.78 | 61.41 | 66.56 | 57.63 | 58.34 | 43.04 | 66.89 | 57.65 |
| | | SET | 73.58 | 64.98 | 69.93 | 68.09 | 73.70 | 64.25 | 72.52 | 62.72 | 68.47 | 55.27 | 72.71 | 62.73 |
| | *sl* ↦ *hr* | PDT | 51.74 | 49.15 | 72.02 | 63.08 | 53.49 | 51.33 | 49.58 | 42.59 | 71.96 | 62.99 | 50.41 | 44.11 |
| | | SSJ | 79.25 | 77.06 | 88.10 | 78.53 | 79.81 | 77.23 | 75.23 | 66.23 | 87.19 | 77.92 | 75.25 | 67.47 |
| WORD | *hr* ↦ *sl* | PDT | 67.33 | 59.24 | 61.80 | 57.14 | 68.11 | 61.13 | 65.84 | 57.12 | 58.17 | 42.99 | 67.12 | 57.70 |
| | | SET | 73.26 | 65.87 | 69.98 | 68.98 | 73.63 | 65.85 | 72.71 | 62.29 | 68.50 | 55.06 | 73.14 | 62.40 |
| | *sl* ↦ *hr* | PDT | 51.67 | 49.58 | 71.47 | 63.51 | 54.62 | 51.82 | 50.25 | 43.17 | 71.27 | 62.79 | 50.79 | 44.07 |
| | | SSJ | 79.51 | 76.89 | 88.71 | 79.69 | 79.81 | 78.03 | 75.95 | 67.19 | 86.92 | 77.28 | 75.89 | 68.18 |
| PHRASE | *hr* ↦ *sl* | PDT | 67.28 | 58.90 | 60.53 | 56.79 | 67.92 | 61.36 | 65.77 | 55.06 | 58.18 | 45.41 | 66.16 | 55.79 |
| | | SET | 74.68 | 65.29 | 69.42 | 68.55 | 74.31 | 65.17 | 73.36 | 60.77 | 68.16 | 58.42 | 72.15 | 61.55 |
| | *sl* ↦ *hr* | PDT | 49.92 | 46.82 | 68.18 | 58.18 | 52.15 | 49.42 | 47.73 | 41.08 | 68.51 | 55.29 | 48.93 | 42.59 |
| | | SSJ | 79.29 | 78.09 | 88.24 | 78.75 | 79.32 | 78.85 | 75.33 | 68.10 | 86.59 | 75.66 | 75.91 | 68.67 |

Table 5: Parsing scores (LAS) for cross-lingual parsers trained on translated treebanks. Scores are split for lexicalized and delexicalized, full MSD and POS only parsers, and with respect to the translation/projection approaches. Row indices represent source languages and treebanks, and indicate the direction of applying SMT (e.g., *hr* ↦ *sl* denotes a Croatian treebank translated to Slovene).

for morphologically rich languages. Through our multilayered test set annotation, we also facilitated a reliable cross-lingual evaluation in a heterogenous testing environment. We list our most important observations:

- Even for closely related languages, using only the basic POS features – which are virtually identical to the widely-used Universal POS of Petrov et al. (2012) – substantially decreases parsing accuracy up to the level comparable with results of McDonald et al. (2013) across the Universal Treebanks language groups.
- Adding MSD features heavily influences all the scores in a positive way. This has obvious implications for improving over McDonald et al. (2013) on the Universal Treebanks dataset.
- Other than that, we show that it is possible to cross-lingually parse Croatian, Serbian and Slovene using all three syntactic annotation schemes, and with high accuracy. A treebank for Serbian does not exist, but we accurately parse Serbian by using PDT, SET and SSJ-style annotations. We parse Croatian using SSJ (transferred from Slovene) and Slovene using SSJ (transferred from Croatian). This clearly indicates the possibilities of uniform downstream pipelining for any of the schemes.
- We show clear benefits of using the SMT approach for transferring SSJ parsers to Croatian and SET parsers to Slovene. We observe these benefits regardless of the low-quality, out-of-domain SMT training data (OpenSubs).

Given the current interest for cross-lingual dependency parsing in the natural language processing community, we will seek to further test our observations on shared morphological features by using other pairs of languages of varying relatedness, drawing from datasets such as Google Universal Treebanks (McDonald et al., 2013) or HamleDT (Zeman et al., 2012; Rosa et al., 2014). The goal of cross-lingual processing in general is to enable improved general access to under-resourced languages. With this in mind, seeing how we introduced a test case of Serbian as a language currently without a treebank, we hope to explore other options for performing cross-lingual experiments on actual under-resourced languages, rather than in an exclusive group of resource-rich placeholders, possibly by means of down-stream evaluation.

**Acknowledgments** The second author was supported by the Swedish Research Council (Vetenskapsrådet), project 2012-916. The fifth author is funded by the EU FP7 STREP project XLike.

## References


Anne Abeillé. 2003. *Treebanks: Building and Using Parsed Corpora*. Springer.

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR Linguistically Annotated Corpus of Croatian. In *Proc. LREC*, pages 1724–1727.

Željko Agić and Danijela Merkler. 2013. Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. *LNCS*, 8082:560–567.

Željko Agić, Danijela Merkler, and Daša Berović. 2012. Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing. In *Proc. IS-LTC*, pages 5–9.

Željko Agić, Danijela Merkler, and Daša Berović. 2013. Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. In *Proc. SPMRL*, pages 22–33.

Željko Agić, Daša Berović, Danijela Merkler, and Marko Tadić. 2014. Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing. In *Proc. LREC*, pages 2313–2319.

Emily Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Emily Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers.

Daša Berović, Željko Agić, and Marko Tadić. 2012. Croatian Dependency Treebank: Recent Development and Initial Experiments. In *Proc. LREC*, pages 1902–1906.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank. In *Treebanks*, pages 103–127.

Bernd Bohnet and Jonas Kuhn. 2012. The Best of Both Worlds – A Graph-based Completion Model for Transition-based Parsers. In *Proc. EACL*, pages 77–87.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajic. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *TACL*, 1:415–428.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proc. COLING*, pages 89–97.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proc. CoNLL*, pages 149–164.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The Stanford Typed Dependencies Representation. In *Proc. COLING*, pages 1–8.

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. In *Proc. IS-LTC*, pages 42–47.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic Transfer Using a Bilingual Lexicon. In *Proc. EMNLP-CoNLL*, pages 1–11.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene Dependency Treebank. In *Proc. LREC*, pages 1388–1391.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proc. LREC*, pages 1806–1809.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1):131–142.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proc. WSMT*, pages 187–197.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311–325.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. ACL*, pages 177–180.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool Publishers.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proc. HLT-EMNLP*, pages 523–530.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proc. EMNLP*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proc. ACL*, pages 92–97.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proc. CoNLL*, pages 915–932.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pages 160–167.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proc. LREC*, pages 2089–2096.

Vesna Požgaj Hadži and Marko Tadić. 2000. Croatian-Slovene Parallel Corpus. In *Proc. IS-LTC*.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. 2014. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proc. LREC*, pages 2334–2341.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: Cross-framework Evaluation of Parsing Morphologically Rich Languages. In *Proc. SPMRL*, pages 146–182.

Anders Søgaard. 2011. Data Point Selection for Cross-language Adaptation of Dependency Parsers. In *Proc. ACL*, pages 682–686.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proc. NAACL*, pages 477–487.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target Language Adaptation of Discriminative Transfer Parsers. In *Proc. NAACL*, pages 1061–1071.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian Morphological Lexicon. In *Proc. BSNLP*, pages 41–46.

Marko Tadić and Tamás Váradi. 2012. Central and South-East European Resources in META-SHARE. *Proc. COLING*, pages 431–438.

Marko Tadić. 2007. Building the Croatian Dependency Treebank: The Initial Stages. *Suvremena lingvistika*, 63:85–92.

Jörg Tiedemann and Preslav Nakov. 2013. Analyzing the Use of Character-Level Translation with Sparse and Noisy Datasets. In *Proc. RANLP*, pages 676–684.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank Translation for Cross-Lingual Parser Induction. In *Proc. CoNLL*, pages 130–140.

Jörg Tiedemann. 2009. News from OPUS: A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proc. RANLP*, volume 5, pages 237–248.

Jörg Tiedemann. 2012. Character-Based Pivot Translations for Under-Resourced Languages and Domains. In *Proc. EACL*, pages 141–151.

Jörg Tiedemann. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proc. COLING*.

Erik F Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proc. CoNLL*, pages 127–132.

Hans Uszkoreit and Georg Rehm. 2012. *Language White Paper Series*. Springer.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can We Translate Letters? In *Proc. WMT*, pages 33–39.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In *Proc. HLT*, pages 1–8.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proc. IJCNLP*, pages 35–42.

Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Stepánek, Zdenek Zabokrtskỳ, and Jan Hajic. 2012. HamleDT: To Parse or Not to Parse? In *Proc. LREC*, pages 2735–2741.