Journal of Artificial Intelligence Research 55 (2016) 209-248 Submitted 03/15; published 01/16

Synthetic Treebanking for Cross-Lingual Dependency Parsing

Jörg Tiedemann JORG.TIEDEMANN@HELSINKI.FI
Department of Modern Languages, University of Helsinki
P.O. Box 24, FI-00014 University of Helsinki, Finland

Željko Agić ZELJKO.AGIC@HUM.KU.DK
Center for Language Technology, University of Copenhagen
Njalsgade 140, 2300 Copenhagen S, Denmark

Abstract

How do we parse the languages for which no treebanks are available? This contribution addresses the cross-lingual viewpoint on statistical dependency parsing, in which we attempt to make use of resource-rich *source* language treebanks to build and adapt models for the under-resourced *target* languages. We outline the benefits, and indicate the drawbacks of the current major approaches. We emphasize *synthetic treebanking*: the automatic creation of target language treebanks by means of annotation projection and machine translation. We present competitive results in cross-lingual dependency parsing using a combination of various techniques that contribute to the overall success of the method. We further include a detailed discussion about the impact of part-of-speech label accuracy on parsing results that provide guidance in practical applications of cross-lingual methods for truly under-resourced languages.

1. Introduction

"Languages are dialects with an army and a navy" is a famous saying popularized by the sociolinguist Max Weinreich. In modern times, this quote could be rephrased—and languages defined—as "dialects with a part-of-speech tagger, a treebank, and a machine translation system." Even though this proposition would disqualify most languages of the world, it is true that the existence of many languages is threatened due to insufficient resources and technical support. Natural language processing (NLP) becomes increasingly important in people's everyday life if we look, for example, at the success of word prediction, spelling correction, and instant on-line translation. Building linguistic resources and tools, however, is expensive and time-consuming, and one of the great challenges in computational linguistics is to port existing models to new languages and domains.

Modern NLP requires data, often annotated with explicit linguistic information, and tools that can learn from them. However, sufficient quantities of electronic data sources are available only for a handful of languages whereas most other languages do not have the privilege to draw from such resources (Bender, 2011; Uszkoreit & Rehm, 2012; Bender, 2013). Speakers of low-density languages and the countries they live in are not able to invest in large data collection and time-consuming annotation efforts, and the goal of cross-lingual

NLP is to share the rich linguistic information with poorly supported languages, making it possible to build tools and resources without starting from scratch.

In this paper, we consider the task of statistical dependency parsing (Kübler, McDonald, & Nivre, 2009). Top-performing dependency parsers are typically trained on dependency treebanks that include several thousands of manually annotated sentences. These statistical parsing models are known to be robust and very efficient, yielding high accuracy on unseen texts. However, even moderately-sized treebanks take a lot of time and resources to produce (Abeillé, 2003), and at this point, they are unavailable or scarce even for major languages.

Thus, similar to other areas of NLP research, we face the challenge posed in our abstract: How do we parse the languages for which no dependency treebanks are available? Without annotated training data we basically have four options in data-driven NLP:

1. We can build parsing models that can learn from raw data using unsupervised machine learning techniques.

2. If manually annotated data is scarcely available, we can resort to various approaches to semi-supervised learning, leveraging the various sources of fortuitous data (Søgaard, 2013).

3. We can transfer existing models and tools to new languages.

4. We can transfer data from resource-rich languages to resource-poor languages and build tools on those data sets.

All four viewpoints are studied intensively not only in connection with dependency parsing but in NLP in general. For parsing, the first option is especially difficult and unsupervised approaches still fall far behind the rest of the field (Søgaard, 2012). Unsupervised models are also difficult to evaluate and applications that build on labeled information have problems in making use of the structures produced by those models. Semi-supervised learning either augments well-resourced environments for improved cross-domain robustness, or largely coincides with the cross-lingual approaches as it is very loosely defined (Søgaard, 2013). Therefore, it is not surprising that the final two options have attracted quite some popularity and gained a lot of merit in enabling parsing for low-resource languages. In this paper, we exclusively look at those techniques.

The basic idea behind transfer approaches is that tools and resources that exist for resource-rich *source languages* are used to build corresponding tools and resources in under-resourced *target languages* by means of adaptation. For statistical dependency parsing such a *cross-lingual approach* essentially means that we either take a parsing model and apply it to another language or use treebanks to train parsers for the new language with target language adaptation taking place in any of the workflow stages. We can, thus, divide the main approaches in cross-lingual dependency parsing into two categories: *model transfer* and *data transfer*.

Model transfer methods have the appealing property that they focus on language universals and structures that can be identified in various languages without side-stepping to the (semi-)automatic creation of annotated data in the target language. There is a strong line of research looking at the identification of cross-lingual features that can be used to port models and tools to new languages. One of their biggest drawbacks is the extreme

abstraction to generic features that cannot cover all language-specific properties of natural languages. Therefore, these methods are often restricted to closely related languages and their performance is usually far below fully supervised target-specific parsing models.

Data transfer methods, on the other hand, emphasize the creation of artificial training data that can be used with standard machine learning techniques to build models in the target language. Most of the work is focused on annotation projection and the use of parallel data, that is, documents that are translated to other languages. Statistical alignment techniques make it possible to map linguistic annotation from one language to another. Another recent approach proposes the translation of treebanks (Tiedemann, Agić, & Nivre, 2014) which enables the projection of annotation without parsing unrelated parallel corpora. Both methods create synthetic data sets without manual intervention and, therefore, we group these techniques under the general term *synthetic treebanking*, which is the main focus of our paper.

The structure of our paper is as follows. After a brief outlook on the contributions of our work, we first provide an overview of cross-lingual dependency parsing approaches. After that, we discuss in depth our experiments with synthetic treebanks, where we inspect annotation projection with parallel data sets and with translated treebanks. We also include a thorough study on the impact of part-of-speech (PoS) tagging in cross-lingual parsing. Before concluding with final remarks and prospects for future work, we discuss the impact of our contribution in comparison with selected recent approaches, both in terms of empirical assessment and the underlying requirements imposed on truly under-resourced languages.

1.1 Our Contributions

The paper addresses annotation projection and treebank translation with a detailed and systematic investigation of various techniques and strategies. We build on our previous work on cross-lingual parsing (Tiedemann et al., 2014; Tiedemann, 2014, 2015) but extend our study with detailed discussions of advantages and drawbacks of each method. We also include a new idea of back-projection that integrates machine translation in the parsing workflow. Our main contributions are the following:

1. We provide an overview of the various approaches to cross-lingual dependency parsing with detailed discussions about the properties of the utilized techniques.

2. We present new competitive cross-lingual parsing results using synthetic treebanks. We ground our results through a discussion on related work and implications for truly under-resourced languages.

3. We provide a thorough study on the impact of PoS tagging in cross-lingual dependency parsing.

Before delving into more details let us first review the selected current approaches to cross-lingual dependency parsing to connect the work presented in this paper with related research.

2. Current Approaches to Cross-Lingual Dependency Parsing

This section provides an overview of cross-lingual dependency parsing. We discuss the previously outlined annotation projection and model transfer approaches in more depth including recent developments in the field. Cross-lingual parsing combines many efforts in dependency treebanking, and in creating standards for PoS and syntactic annotations. We start off by outlining the current practices in empirical evaluation of cross-lingual parsers, and the linguistic resources used for benchmarking.

2.1 Treebanks and Evaluation

In a supervised setting, cross-lingual dependency parsing amounts to training a parser on a treebank, and applying it on the target text. However, the empirical quality assessment for such a parser on the target data introduces certain additional constraints. To evaluate supervised cross-lingual parsers, we require at least the following three components:

1. parser generators: trainable, language-independent dependency parsing systems,
2. dependency treebanks for the source languages, and
3. held-out evaluation sets for the target languages.

In the years following the venerable CoNLL 2006 and 2007 shared task campaigns in dependency parsing (Buchholz & Marsi, 2006; Nivre, Hall, Kübler, McDonald, Nilsson, Riedel, & Yuret, 2007), many mature parsers were made publicly available across the different parsing paradigms. This resolves the first point from our list, as choosing to apply—and comparing between—different approaches to parsing in a cross-lingual setup is nowadays made trivial by abundant parser availability. We can now easily benchmark a respectable number of parsers for accuracy, processing speed, and memory requirements.

Experimental setup for cross-lingual parsing thus amounts to choosing the training and testing data, and to defining the evaluation metrics.

2.1.1 Intrinsic and Extrinsic Evaluation

We can perform intrinsic or extrinsic evaluation of dependency parsing. In *intrinsic evaluation*, we typically apply evaluation metrics to gauge the various aspects of parsing accuracy on held-out data, while in *extrinsic evaluation*, parsers are scored by the gains yielded in subsequent—or *downstream*—tasks which make use of dependency parses as additional input.

Dependency parsers are intrinsically evaluated for labeled (LAS) and unlabeled (UAS) attachment scores: the portions of correctly paired heads and dependents in dependency trees, with or without keeping track of the edge labels, respectively. Sometimes we also evaluate for labeled (LEM) and unlabeled (UEM) exact match scores, to determine how often the parsers correctly parse entire sentences. For a more detailed exposition of dependency parser evaluation, see the work of Nivre (2006) and Kübler et al. (2009), and also note that Plank et al. (2015) provide detailed insight into the correlations between these and various other dependency parsing metrics and human judgements on the quality of parses.

In a monolingual intrinsic evaluation scenario, we either have predefined held-out test data at our disposal, or we cross-validate by slicing the treebank into training and test sets. In both cases, the treebank and the test sets belong to the same resource, and are created using the same annotation scheme, which in turn typically stems from the same underlying syntactic theory. However, given the heterogenous development of syntactic theories, and subsequently of treebanks for different languages (Abeillé, 2003), this does not necessarily hold in a cross-lingual setup. Moreover, excluding the very recent treebanking developments—which we discuss a bit further in this section—prior to 2013, the odds of randomly sampling from a pool of all publicly available treebanks and drawing a source-target pair annotated in the same (or even similar) scheme are virtually non-existent.

The syntactic annotation schemes generally differ in: (a) rules for attaching dependents to heads, and (b) dependency relation labels, that is, the syntactic tagsets. Given two treebanks with incompatible syntactic annotations, without performing any conversions, it is more likely to expect similarities in head attachment rules, than in the syntactic tagsets. This fact is present in all the initial cross-lingual parsing experiments (Zeman & Resnik, 2008; McDonald, Petrov, & Hall, 2011; Søgaard, 2011). Such initial efforts in charting cross-lingual dependency parsing mainly used the CoNLL shared task datasets, and they all evaluated for UAS. The rare exceptions are, for example, the generally under-resourced Slavic languages (Agić, Merkler, & Berović, 2012) subscribing to (slightly modified versions of) the Prague Dependency Treebank scheme (Böhmová, Hajič, Hajičová, & Hladká, 2003).

Very recently, a substantial effort was undertaken in bridging the annotation scheme gap in dependency treebanking to facilitate uniform syntactic processing of world's languages. The effort resulted in two editions of Google Universal Treebanks (UDT) (McDonald et al., 2013), which were in turn recently superseded by the Universal Dependencies project (UD) (Nivre et al., 2015). In these projects, the Stanford typed dependencies (SD) (De Marneffe, MacCartney, & Manning, 2006) were used as the adaptable basis for designing the underlying annotation scheme, and for applying it by using human expert annotators on several languages. These datasets made possible the first reliable cross-lingual dependency parsing experiments, namely the ones by McDonald et al. (2013), and also enabled the use of LAS as the default evaluation metric, just like in monolingual parsing. For these reasons, UDT and UD are the *de facto* standard datasets for benchmarking cross-lingual parsers today, while the CoNLL datasets are still used mainly for backward compatibility with previous research. In another effort, the HamleDT dataset (Zeman et al., 2014), 30 treebanks were automatically converted to the Prague scheme, and then to SD, and are also frequently used in evaluation campaigns. We do currently note a preference for UDT and UD, since they were produced through manual annotation.

Given our short exposition of dependency treebanking in relation with cross-lingual parsing, in this paper, we opt for using UDT in our experiments. As for the choice of sources and targets, we do a Cartesian product of the dataset: we treat all the available languages as both sources and targets. This is the more common approach in cross-lingual parsing, even if there is research that uses English as a source-only language, and treats the other languages as targets.

The extrinsic evaluation of cross-lingual parsing is much less developed, although the arguments to its favor are very convincing. Namely, the underlying goal of cross-lingual parsing is enabling the processing of actual under-resourced languages. For these languages,

even the parsing test sets may not be readily available. For conducting empirical evaluations in such extreme cases, we might resort to downstream applications (Elming et al., 2013). The choice of downstream tasks might pose a separate challenge in this case, and devising feasible (and representative) tasks for extrinsic evaluation of cross-lingual dependency parsing remains largely unaddressed. In this paper, we deal only with intrinsic evaluation.

2.1.2 Part-of-Speech Tagging

As noted in our brief introduction to model transfer, dependency parsers make heavy use of PoS features. As with the syntactic annotations, sources and targets may or may not have shared PoS annotation layers, and moreover, PoS taggers may or may not be available for the target languages.

The issue of PoS compatibility is arguably less difficult to resolve than the structural or labeling differences in dependency trees, as PoS tags are more or less straightforwardly mapped to one another. At this point, we also note the recent approaches to learning PoS tag conversions (Zhang, Reichart, Barzilay, & Globerson, 2012), which systematically facilitate the conversions. Furthermore, efforts such as UDT/UD also build on a shared PoS representation, the so-called Universal PoS (UPoS) (Petrov et al., 2012). UD extends the UPoS specification by introducing additional PoS tags—17 instead of the initial 12—and by providing the support for standardized morphological features such as noun gender and case, or verb tense. That said, these added features are not yet readily available, and the shared representation in UDT/UD amounts to a 12- or 17-tag-strong PoS tagset. As for the treatment of source languages with respect to PoS tagging, most of the work in cross-lingual parsing presumes the existence of taggers, or even tests on gold standard PoS input. Recently, Petrov (2014) argued strongly for the use of predicted PoS in cross-lingual parsing, which does make for a more realistic testing environment, especially with increased availability of weakly supervised PoS taggers (Li et al., 2012; Garrette et al., 2013). In this paper, we experiment both with gold standard and predicted PoS features in order to stress the impact of tagging accuracy on parsing performance. We also discuss the implications of these choices in enabling the processing of truly under-resourced languages.

2.2 Model Transfer

We now proceed to sketch the main approaches to cross-lingual dependency parsing: model transfer, annotation projection, and treebank translation. We also reflect on the usage of cross-lingual word representations in cross-lingual parsing, while we particularly emphasize the annotation projection and treebank translation approaches.

Simplistic model transfer amounts to applying the source models to the targets with no adaptation, which can still be rather successful for closely related languages (Agić et al., 2014). However, the flavor of model transfer that has recently attracted a fair amount of interest owes to the availability of cross-lingually harmonized annotation (Petrov et al., 2012) that makes it possible to use shared PoS features across languages. The most straightforward technique is to train delexicalized parsers that heavily rely on UPoS tags. Figure 1 illustrates the basic idea behind these models. This simple technique has shown some success for closely related languages (McDonald et al., 2013). Several improvements can be achieved by using multiple source languages (McDonald et al., 2011; Naseem, Barzilay, & Globerson,

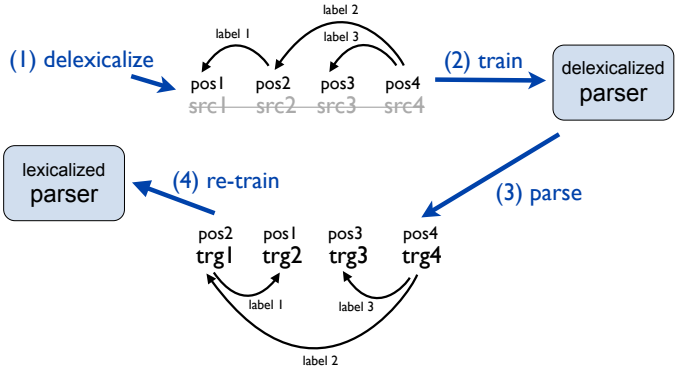

Figure 1: An illustration of the delexicalized model transfer, with an implication of the lexicalization option through self-training.

2012), and additional cross-lingual features that can be used to transfer models to a new language, such as cross-lingual word clusters (Täckström, McDonald, & Uszkoreit, 2012) or word-typology information (Täckström, McDonald, & Nivre, 2013b). There are ways to re-lexicalize models as well. Figure 1 suggests a self-learning procedure that adds lexical information from data sets that have automatically been annotated using delexicalized models. Various data selection techniques can be used to focus on reliable cases to improve the value of the induced lexical features.

The advantage of transferred models is that they do not require parallel data, at least not in their most generic form. However, reasonable models require some kind of target language adaptation and parallel or comparable data sets are usually necessary to perform such adaptations. The largest drawback of model transfer is the strong abstraction from language-specific features to the universal properties. For many fine-grained linguistic differences, this kind of coarse-grained universal knowledge is often not informative enough (Agić et al., 2014). Consequently, a large majority of recent approaches aim at bridging this representational deficiency.

2.3 Cross-Lingual Word Representations

Model transfer requires abstract features to capture the universal properties of languages. The use of cross-lingual word clusters was already mentioned in the previous section, and the benefits of monolingual clustering for dependency parsing are well-known (Koo, Carreras, & Collins, 2008). Recently, distributed word representations have entered NLP in various models (Collobert et al., 2011). The so-called *word embeddings* capture the distributional properties of words in continuous vector representations that can be used to measure syntactic and semantic relations even across languages (Mikolov, Le, & Sutskever, 2013). Their monolingual variety has found many applications in NLP. Distributed word representations for cross-lingual dependency parsing were first applied just recently by Xiao and Guo (2014). They explore word embeddings as another useful abstraction that enables more robust model transfer across languages. However, they apply their techniques to the

old CoNLL data sets and cannot provide labeled attachment scores and comparable results to our settings.

Several recent publications show that bilingual word embeddings learned from aligned bitexts improve semantic representations. Faruqui and Dyer (2014) use canonical correlation analysis to find cross-lingual projections of monolingual vector space models. Zou, Socher, Cer, and Manning (2013) learn bilingual word embeddings with fixed word alignments. Klementiev, Titov, and Bhattarai (2012) treat cross-lingual representation learning as a multitask learning problem in which cross-lingual interactions are based on word alignments and word embeddings are shared across the various tasks. All of these techniques have significant value in improved model transfer and may act as the necessary target language adaptation to move beyond language universals as the only feature in transfer models.

In cross-lingual parsing, we can envision the word representations as a valuable addition to model transfer in the direction of regularization. That said, their usage maintains the previously listed advantages and drawbacks of model transfer, and adds another prerequisite: the availability of parallel texts for inducing the embeddings. There have been some very recent developments in creating cross-lingual embeddings without parallel text (Gouws & Søgaard, 2015) but their applicability in dependency parsing is yet to be verified. Here, we note a very recent contribution by Søgaard et al. (2015), who use inverted indexing on cross-lingually overlapping Wikipedia articles to produce truly inter-lingual word embeddings. As they show competitive scores in cross-lingual dependency parsing, we further address their contribution in our related work discussion.

2.4 Annotation Projection

The use of parallel corpora and automatic word alignment for transferring linguistic annotation from a source language to a new target language has quite a long tradition in NLP. The pioneering work of Yarowsky, Ngai, and Wicentowski (2001) was followed by a number of researchers, and for various tasks, the transfer of dependency annotation among others (Hwa et al., 2005). The basic idea is to use existing tools and models to annotate the source side of a parallel corpus and then to use alignment to guide the mapping of that annotation to the target side of the corpus. Assuming that the source language annotation is sufficiently correct and that the aligned target language reflects the same syntactic patterns, we can train parsers on the projected data to bootstrap tools for languages without explicit linguistic resources such as syntactically annotated treebanks. Figure 2 illustrates the general idea of annotation projection for the case of syntactic dependencies and parser model induction. Note that PoS labels are typically projected as well along with the dependency relations.

The first attempts to directly map dependency information coming from diverse treebanks resulted in rather poor performance. In their work, Hwa et al. (2005) had to rely on additional post-processing rules to transform the results into reasonable structures. As we argued in the previous subsection, one of the main problems in the early work was the incompatibility of treebanks that have individually been developed for various languages following different guidelines and using different label sets. The latter is also the reason why no labeled attachment scores could be reported in that work, which makes it difficult to place these cross-lingual approaches in relation to standard models trained for the target language.

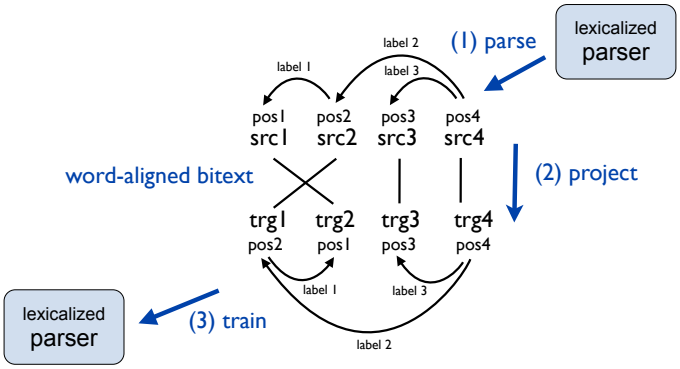

Figure 2: An illustration of the syntactic annotation projection system for cross-lingual dependency parsing.

Less frequent, but also possible, is the scenario where the source side of the parallel corpus contains manual annotation (Agić et al., 2012). This addresses the problem created by projecting noisy annotations, but it presupposes parallel corpora with manual annotation, which are rarely available. Additionally, the problem of incompatible annotation still remains.

The introduction of cross-lingually harmonized treebanks changed the situation significantly (McDonald et al., 2013). These data sets use identical labels and adhere similar annotation guidelines that make it possible to directly compare structures when projected from other languages. In the work of Tiedemann (2014), we explore projection strategies and discuss the success of annotation projection in comparison to other cross-lingual approaches. Our work builds on the direct correspondence assumption (DCA) proposed by Hwa et al. (2005). They define several projection heuristics that make it possible to project any dependency structure through given word alignments to a target language sentence. The basic procedures cover different types of word alignments. One-to-one alignments are the most straightforward case in which dependency relations can simply be copied. Unaligned source language tokens are covered by additional DUMMY nodes that capture all relations that are connected to that token in the source language (see the left-most graph in Figure 3). Many-to-one links are resolved by only keeping the link to the head of the aligned source language tokens and deleting all other links (see the graph in the middle). One-to-many alignments are handled by introducing additional DUMMY nodes that act as the immediate parent in the target language, and which will capture the dependency relation of the source side annotation (see the right-most graph in Figure 3). Many-to-many alignments are treated in two steps. First we apply the rule for one-to-many alignments and after that the many-to-one rule. Finally, unaligned target language tokens are simply dropped and will be removed from the target sentence.

Some issues are not explicitly covered by the original publication of the algorithm. For example, it is not entirely clear in what sequence these rules should be applied and how labels should be projected. Some of the rules, for example, change the alignment structure and may cause additional unaligned source tokens that need to be handled by other rules. In our implementation, we first apply the one-to-many rule for all cases in the sentence

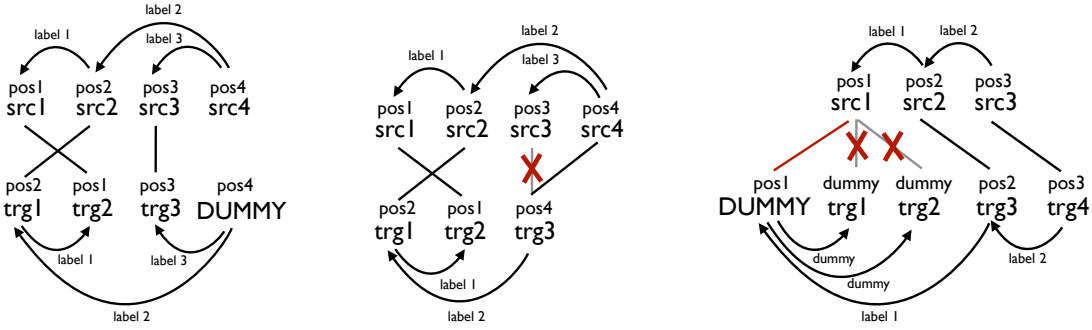

Figure 3: Annotation projection heuristics for special alignment types: Unaligned source words (left graph), many-to-one alignments (center), one-to-many alignments (right graph).

before applying the many-to-one rule and, thereafter, resolving unaligned source tokens. The final step includes the mapping of dependency relations through the remaining one-to-one alignments. For one-to-many alignments, we transfer the PoS and dependency labels to the newly created DUMMY node (following the rule for one-to-one alignments after resolving the one-to-many link) and the previously aligned target language tokens will obtain DUMMY PoS labels and their dependency relation to the governing DUMMY node will also be labeled as DUMMY (see Figure 3).

Projecting syntactic dependency annotation creates several other problems as well. First of all, crossing word alignments cause a large amount of non-projectivity in the projected data. The percentage of non-projective structures goes up to over 50% for the UDT data (Tiedemann et al., 2014). Furthermore, projection heuristics can lead to conflicting annotation as it is shown in the authentic example illustrated in Figure 4. These issues put an additional burden on the learning algorithms and many cross-lingual errors are caused by such complex and ambiguous cases.

Nevertheless, Tiedemann (2014) demonstrates that annotation projection is competitive to other cross-lingual methods and its merits are further explored by Tiedemann (2015).

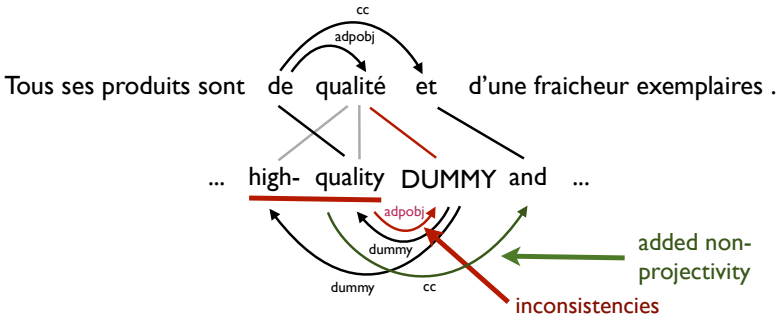

Figure 4: Issues with annotation projection illustrated on a real-life example.

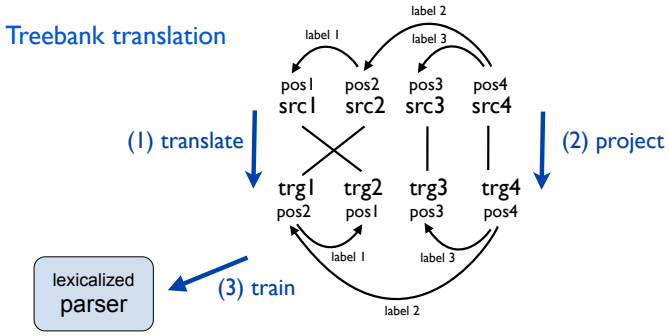

Figure 5: An illustration of the synthetic treebanking approach through translation.

2.5 Translating Treebanks

The notion of translation in cross-lingual parsing was first introduced by Zhao, Song, Kit, and Zhou (2009), who use a bilingual lexicon for lookup-based target adaptation. A similar method is also adopted by Durrett et al. (2012). This simplistic lookup approach is used by Agić et al. (2012), who exploit the availability of a parallel corpus for two closely related languages, one side of the corpus being a dependency treebank. The former evaluates for UAS on 9 languages from the CoNLL datasets, while the latter research deals only with Croatian and Slovene and is of a smaller scale.

Tiedemann et al. (2014) are the first to use full-scale statistical machine translation (SMT) to synthesize treebanks as SMT-facilitated target language adaptations for cross-lingual parsing. They use UDT for LAS evaluation, while also performing a subset of experiments with the CoNLL 2007 data for backward compatibility. In this paper, we often refer to, and we build on that work. Figure 5 illustrates the general idea of this technique, and we proceed to discuss its implications.

As sketched in the introduction, at the core of the synthetic treebanking idea is the concept of automatic source-to-target treebank translation. Its workflow consists of the following steps:

1. Take a source-target parallel corpus and a large monolingual target language corpus to train an (ideally top-performing) SMT system, or—if available—apply an existing source-target machine translation system.

2. Given a source language treebank, translate it into the target language. Word-align the original sentence and its translation, or preserve the phrase alignments provided by the SMT system.

3. Use the alignments to project the dependency annotations from the source treebank to the target translation, in turn creating an artificial (or synthetic) treebank for the target language.

4. Train a target language parser on the synthesized treebank, and apply (or evaluate) it on target language data.

This sketch of treebank translation opens up a large parameter tuning search space, and also outlines the various properties of the approach. We discuss them briefly, and defer the reader to the detailed expositions of the many intricacies in these papers (Tiedemann et al., 2014; Tiedemann, 2014, 2015).

2.5.1 Components

The prerequisites for building an SMT-supported cross-lingual parsing system are: (a) the availability of parallel corpora, (b) a platform for building state-of-the-art SMT systems, (c) algorithms for robust annotation projection, and (d) the previously listed resources needed for cross-lingual parsing in general: treebanks and parsers.

Parallel corpora are now available for a very large number of language pairs, even outside the benchmarking frameworks of CoNLL and UDT. The size and domains of the parallel data influences the quality of SMT, and subsequently of the cross-lingual parsers. The SMT community typically experiments with the Europarl dataset (Koehn, 2005), while many other datasets are also freely available and cover many more languages, such as the OPUS collection (Tiedemann, 2012). Ideally, the parallel corpora used in SMT are very large, but for some source-target pairs, this may not necessarily be the case. Moreover, the corpora might not be spread across the domains of interest, leading to decreased performance. Domain dependence is thus inherent in the choice of parallel corpora for training SMT systems. Here, we note a recent contribution by Agić et al. (2015), who learn a hundred PoS taggers for truly under-resourced languages by using label propagation on a multi-parallel Bible corpus, indicating the possibility of bootstrapping NLP tools in even the most hostile environments, and the subsequent applicability of such tools across domains.

In this paper, we opt for using Moses (Koehn et al., 2007) as the *de facto* standard platform for conducting SMT research. In summary, since our approach to SMT goes beyond the dictionary lookup of Durrett et al. (2012), we mainly experiment with phrase-based models, gaining the target language adaptations in the form of both the lexical features and the reordering. The projection algorithms for synthetic treebanking can in whole be transferred from the annotation projection approaches. We do, however, consider their various parametrizations, while Tiedemann et al. (2014) previously proposed a novel algorithm, and Tiedemann (2014) thoroughly compared various approaches to annotation projection.

2.5.2 Advantages and Drawbacks

Automatic translation has the advantage that we can use the manually verified annotation of the source language treebank and the given word alignment, which is an integral part of the translation model. Recent advances in statistical machine translation (SMT) combined with the ever-growing availability of parallel corpora are now making this a realistic alternative. The relation to annotation projection is obvious as both involve parallel data with one side being annotated. However, the use of direct translation brings two important advantages.

First of all, using SMT, we do not accumulate errors from two sources: the tool—tagger or parser—used to annotate the source language of a bilingual corpus, and the noise coming from alignment and projection. Instead, we use the gold standard annotation of the source language which can safely be assumed to be of much higher quality than any automatic

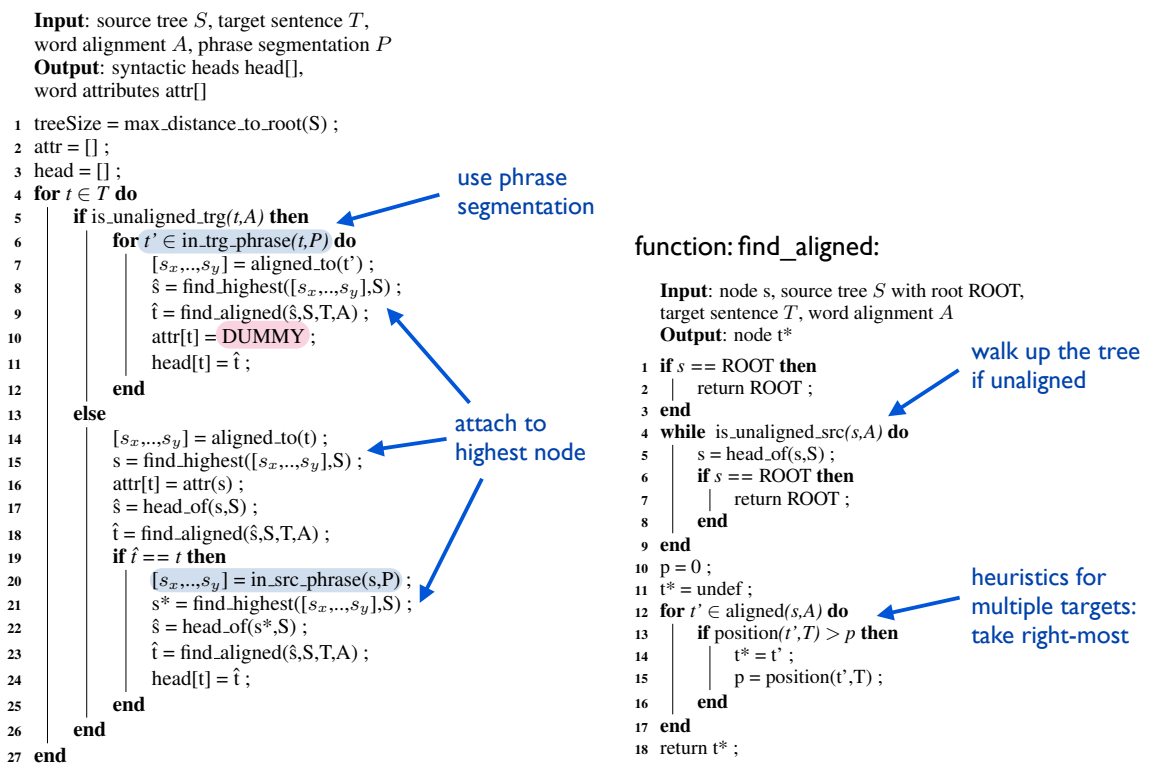

Figure 6: Annotation projection without DUMMY nodes proposed by Tiedemann et al. (2014).

annotation obtained by using a tool trained on that data, especially in light of cross-domain accuracy drops. Moreover, using SMT may help in bypassing domain shift problems, which are common when applying tools trained (and evaluated) on one resource to text from another domain.

Secondly, we can assume that SMT will produce output that is much closer to the input than manual translations in parallel texts usually are. Even if this may seem like a shortcoming in general, in the case of annotation projection it should rather be an advantage, because it makes it more straightforward and less error-prone to transfer annotation from source to target. Furthermore, the alignment between words and phrases is inherently provided as an output of all common SMT models. Hence, no additional procedures have to be performed on top of the translated corpus. Recent research (Zhao et al., 2009; Durrett et al., 2012) has attempted to address synthetic data creation for syntactic parsing via bilingual lexica. Tiedemann et al. (2014) extend this idea by proposing three different models for automatic translation based on induced bilingual lexica and phrase-based translation models. In that work, the authors propose a new projection algorithm that avoids the creation of DUMMY nodes in the target language that we have discussed in section 2.4. The procedure is summarized in the pseudo-code shown in Figure 6.

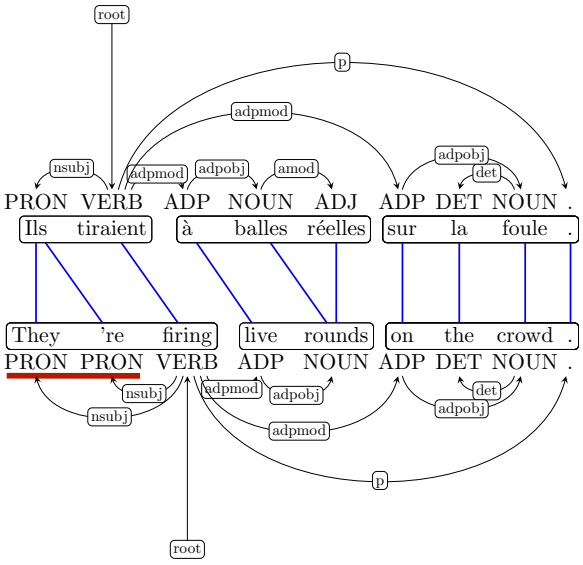

Figure 7: An example sentence translated from French to English with projections using the algorithm shown in Figure 6. The boxes indicate the segmentation used by the phrase-based translation model.

The key feature of this algorithm is that it makes use of the segmentation of sentences into "phrases" together with their counterparts in the other language that are applied by the underlying translation model. We can use this information to handle unaligned tokens without creating additional DUMMY nodes as described in Figure 6. However, contrary to our expectations, this algorithm does not work very well in practice and Tiedemann et al. (2014) show empirically that a simple word-to-word translation model outperforms the phrase-based systems with this projection algorithm in most cases. Part of the problem is the ambiguous projection of PoS labels when handling one-to-many and many-to-one alignments. An example is shown in Figure 7. Both *They* and *'re* are assigned to be pronouns due to the links to the French *Ils* which certainly confuses the model trained on such projected data.

The treebank translation approach using phrase-based SMT is further explored by Tiedemann (2014). Tiedemann (2015) introduces the use of syntax-based SMT for cross-lingual dependency parsing. In that work, the authors propose several improvements of the DCA-based projection heuristics originally developed by Hwa et al. (2005). Simple techniques that reduce the number of DUMMY elements in the projected data help to significantly improve the results in cross-lingual parsing. We also realized that the placement of DUMMY nodes is crucial. Strategies that choose positions where that minimize the risk of additional non-projectivity are useful to improve parser model induction. We will mainly use the techniques developed in that work in the experiments described in section 3.

The drawbacks of the synthetic treebanking approach are related to its hybrid nature: a) it inherits the syntax projection risks from the annotation projection approach as its success is bound by the projection quality, and b) it critically depends on the quality of SMT, which in turn depends on the size and quality of the underlying parallel corpora.

As for the latter point, the experiments by Tiedemann et al. (2014) reveal a compelling robustness of cross-lingual parsing to SMT noise in this framework, and in that paper we also argue that projection into synthetic texts is simpler than projection between actual parallel text. Another important drawback is the need for large parallel data sets to train reasonable translation models for the languages under consideration. Alternatively, any handcrafted rule-based system could be applied as well. However, such systems and data sets are rarely available for low-resource languages. On the other hand, there are techniques that can improve machine translation via bridge languages. Tiedemann and Nakov (2013) demonstrate how small amounts of parallel data can successfully been used for building translation models for truly under-resourced languages. Their approach of creating synthetic training data for statistical machine translation with low resource languages fits very well in the spirit of synthetic treebanking.

2.6 What About Truly Under-Resourced Languages?

Up to this point, we have outlined the underlying concepts for the major approaches to cross-lingual dependency parsing today. We have also discussed some intricacies of enabling cross-lingual parser evaluation. Here, we proceed to discuss how these two outlooks—namely, the way we implement cross-lingual parsers, and the way we evaluate them for parsing accuracy—reflect on dependency parsing of *truly* under-resourced languages.

What makes a language under-resourced? Following Uszkoreit and Rehm (2012), we acknowledge the many facets involved in attempting to address this question. Generally, however, an under-resourced language is distinguished by lacking the basic NLP-enabling linguistic resources, such as PoS-tagged corpora or treebanks. In this paper, we take dependency parsing-oriented viewpoint, which allows for casting the issue of under-resourcedness in the specific terms of dependency parsing enablement for a given language. Thus, a language is under-resourced if we cannot build a dependency parser for it or, otherwise said, if no dependency treebank exists for that language. Since statistical dependency parsing critically depends on the availability of PoS tagging, we make this an additional requirement, which in turn implies the following three levels of resource availability. Note that this list is a parsing-oriented specialization of the general discussion on low-resource languages from the introduction.

1. There is a PoS-tagged corpus and a treebank available for a given language, and by virtue of those, we have at hands a PoS tagger and a dependency parser for that language. We call such languages well-resourced or resource-rich languages from a dependency parsing viewpoint, as we can use the dedicated native language resources to parse texts written in that language.

2. For a given language, there are no PoS tagging or parsing resources available. This includes both the annotated corpora and the NLP tools. We address such languages as under-resourced or low-resource languages, as we cannot natively parse them for syntactic dependencies, neither can we annotate them for PoS tags.

3. We have a PoS-tagged corpus or PoS tagger available for a given language, but no treebanks or parsers exist for it. Even if there is some NLP support for such languages

through PoS annotation, we still approach them as under-resourced from the viewpoint of dependency parsing.

If we want to parse the languages from group 2 for syntactic dependencies, we must address both issues—the unavailability of supporting resources for PoS tagging and dependency parsing—and often even more basic processing facilities such as sentence splitters or tokenizers. In NLP, we often call such languages *truly* under-resourced. Group 3 is somewhat easier, as we presumably only address the dependency-syntactic processing layer.

In the recent years, the field has dealt extensively—and by and large, separately—with providing low-resource languages with PoS taggers and dependency parsers. Taking two examples into account, Das and Petrov (2011) show how to bootstrap accurate taggers using parallel corpora, while Agić et al. (2015) take under-resourcedness to the extreme by presuming severe data sparsity and still manage to yield very reasonable PoS taggers for a large number of low-resource languages. We are thus safe to conclude that even for the most severely under-resourced languages, reasonable PoS taggers can be made available using one of these techniques, if not already available off-the-shelf.

This reasoning underlies all current approaches to cross-lingual dependency parsing, in that we presume the availability of PoS annotations, natively or through publicly available related research. Since we are also required to at least intrinsically evaluate the resulting parsers, we conduct our empirical assessments in an exclusive group of languages with at least some syntactically annotated test data available. In effect, we are evaluating *by proxy*, as the truly under-resourced languages do not enjoy even the basic test set availability. On top of all that, the various top-performing approaches to cross-lingual parsing—such as the previously discussed annotation projection, treebank translation, or word representation-supported model transfer—introduce additional constraints or requirements. Most often, we presume the availability of large source-target parallel corpora. One might argue accordingly that we make a poor case for low-resource languages by amassing the prerequisites for our methods to work, thus departing from the very definition of a low-resource language. In turn, and in favor of the current approaches, we argue the following.

- The current research in enabling PoS tagging for under-resourced languages justifies the separate handling of cross-lingual dependency parsing by presuming the availability of PoS tagging. We refer the reader to the work by Täckström et al. (2013a) for a detailed exposition and state-of-the-art results, together with the previously mentioned work on bootstrapping taggers.

- McDonald et al. (2013) validate the evaluation by proxy by showing how a uniform syntactic representation partially enables inferential reasoning about the performance of ported parsers on truly under-resourced languages. Namely, they show that typological similarity plays an important role in predicting the quality of transferred parsers. This is built on by, for example, Rosa and Zabokrtsky (2015), who use a data-driven language similarity metric to actually predict the best sources for the given targets in cross-lingual parsing.

- The remaining prerequisites for top-level cross-lingual parsing, such as the treebank translation approach we argue for in this paper, amount to source-target parallel

corpora and possibly also monolingual target corpora. While this may at first seem as a substantial added requirement, we note that text corpora are more readily available than expert-annotated linguistic resources, and the collections such as OPUS (Tiedemann, 2012) provide large quantities of cross-domain data for many languages. To further the claim, Agić et al. (2015) illustrate how annotation projection could be applied to learn PoS taggers for hundreds, possibly even thousands of languages using nothing but translations of (parts of) the Bible in a very simple setup.

Before concluding, we duly note the perceived disconnect between *evaluating* cross-lingual parsers and actually *enabling* dependency parsing for languages that lack the respective resources. We argue here that the former constitutes empirical research, while the latter is primarily an engineering feat, and we are thus obliged to follow the field in adhering to the former in this contribution. However, we do note that devising multiple systematic downstream evaluation scenarios for truly under-resourced languages is sorely needed at this point in the field's development, and would resolve an important disconnect in cross-lingual NLP research.

We now proceed to discuss the core of our paper: the empirical validation of the synthetic treebanking approach to cross-lingual parsing. We reflect once more on the prerequisites and truly under-resourced languages in the related work discussion that follows our exposition of synthetic treebanking.

3. Synthetic Treebanking Experiments

In this section, we will discuss a series of experiments that systematically explore various cross-lingual parsing models based on annotation projection and treebank translation. Here, we only assess the properties of the specific approach, and we compare them intrinsically or to the baseline. We provide a comparison to selected more recent work in section 4.

In our setup, we always use the test sets provided by the Universal Dependency Treebank version 1 (UDT) (McDonald et al., 2013) with their cross-lingually harmonized annotation that makes it possible to perform fair evaluations across languages including labeled attachment scores (LAS), which we will use as our primary evaluation metric. Similar to previous literature, we include punctuation in the calculation of LAS to ensure comparability to related literature (Tiedemann, 2014). In all our experiments, we apply `mate-tools` (Bohnet, 2010) to train graph-based dependency parsers, which gives us very competitive performance in all settings. We leave out Korean in our experiments due to the fact that we do not have bitexts from the same domain as for the other languages, which we need for annotation projection and SMT training. Thus, we experiment using five languages: English (EN), French (FR), German (DE), Spanish (ES), and Swedish (SV).

3.1 Baseline

Our initial baseline is a delexicalized model which is straightforward to train on the provided training data of the UDT. Table 1 lists the attachment scores achieved by applying these models across languages. Our scores confirm the results of McDonald et al. (2013); minor differences are due to the different choices of the training algorithms. Note that we always use columns to represent the target languages that we test and rows refer to source languages

used in training, projection or translation. We also always report the scores for all source-target pairs, as reporting on averages or highest per-target scores might arguably make for a biased insight into the methods.

	target language ⟶				
LAS	DE	EN	ES	FR	SV
DE	**70.84**	45.28	48.90	49.09	52.24
EN	48.60	**82.44**	56.25	58.47	59.42
ES	47.16	47.31	**71.45**	62.39	54.63
FR	46.77	47.94	62.66	**73.71**	54.89
SV	52.53	48.24	52.95	55.02	**74.55**
`mate-tools` (coarse)	78.38	91.46	82.30	82.30	84.52
`mate-tools` (full)	80.34	92.11	83.65	82.17	85.97

Table 1: Results for the delexicalized models. For comparison there are also LAS's of lexicalized models at the bottom of the table. *coarse* uses coarse-grained PoS labels only and *full* adds even fine-grained PoS information.

As we can see, the results are around 10 LAS points below the fully lexicalized models and significant drops can be observed when training on other languages even though they are all quite closely related. This is all but unexpected considering the naive approach of using coarse-grained PoS label sequences without modification as the only type of information in training these models. We do note, however, that the decrease in accuracy is not so drastic for the typologically closest language pair (French-Spanish). In the following section, we discuss various ways of adapting cross-lingual models to the target language, and we will start with annotation projection in aligned parallel corpora.

3.2 Improved Annotation Projection

Annotation projection is used in connection with word-aligned bilingual parallel corpora (bitexts). In our experiments, we use Europarl (Koehn, 2005) for each language pair following the basic setup of Tiedemann (2014). The baseline model applies the DCA projection heuristics as presented by Hwa et al. (2005) and the first 40,000 sentences of each bitext in the corpus (repetitions of sentences included). Word alignments are produced using IBM model 4 as implemented in GIZA++ (Och & Ney, 2003) trained in the typical pipeline as it is common in statistical machine translation using the Moses toolbox (Koehn et al., 2007). We use the entire Europarl corpus version 7 to train the alignment models to obtain proper statistics and reliable parameter estimates. The asymmetric alignments are symmetrized with the intersection and the grow-diag-final-and heuristics. The results of our baseline projection model is given in Table 2.

The value of word-aligned bitext can clearly be seen in the performance of the cross-lingual parser models. They outperform the naive delexicalized models by a large margin. However, they are still pretty far away from the supervised monolingual models even for these related language pairs. Tiedemann (2015) discusses various improvements of the projection algorithm with significant effects on the performance of the trained models. One

	DE	EN	ES	FR	SV
DE	–	53.27	57.69	60.49	65.25
EN	62.28	–	62.29	65.54	66.97
ES	60.46	49.34	–	68.10	64.67
FR	61.27	53.46	66.51	–	62.75
SV	62.96	51.07	61.82	64.99	–

Table 2: Baseline performance in LAS of a DCA-based annotation projection with 40,000 parallel sentences tested on target language test sets.

problem of the DCA algorithm is the creation of DUMMY nodes and labels that disturb the training procedures. Many of these nodes can easily be removed without loosing much information. Figure 8 illustrates our approach that deletes DUMMY leaf nodes and collapses dependency relations that run via internal DUMMY nodes with single out-going edges.

Adding this modification to the DCA projection heuristics we can achieve significant improvements for various language pairs. Table 3 summarizes the LAS's for all models with the new treatment of DUMMY nodes.

Tiedemann (2015) also introduces a new procedure for treating one-to-many word alignments. In the original algorithm, they cause additional DUMMY nodes that act as parents for the other aligned target language tokens. The new approach takes advantage of different alignment symmetrization algorithms and uses the high-precision links coming from the intersection of asymmetric word alignments to find the head of a multi-word unit, whereas links from the high-recall symmetrization are used to attach the words to that head word. Figure 9 illustrates this procedure by means of a sentence pair from Europarl.

Finally, Tiedemann (2015) also proposes to discard all trees that have remaining DUMMY nodes. This may remove up to 90% of the training examples but assuming the availability of large bitexts makes it possible to project additional sentences to fill the training data. Discarding projected trees with DUMMY nodes effectively removes sentence pairs with non-literal translations and complex alignment structures that are in any case less suited for

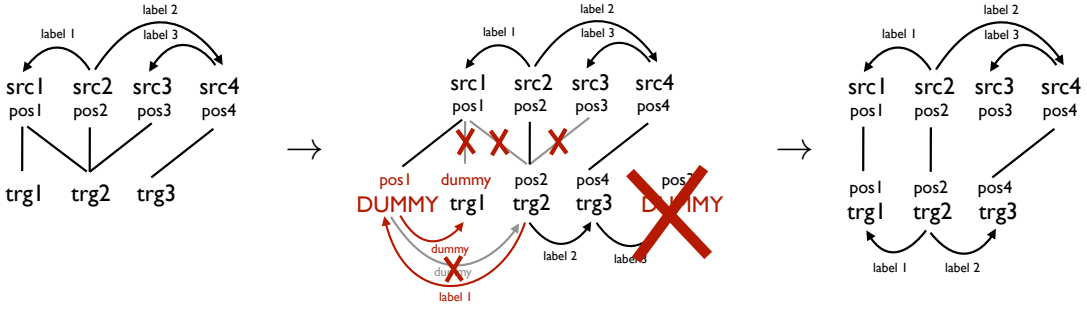

Figure 8: Removing DUMMY nodes from projected parse trees: (i) Delete DUMMY leaf nodes. (ii) Collapse unary productions over DUMMY nodes.

	DE	EN	ES	FR	SV
DE		**53.54**$^{+0.27}$	****60.17**$^{+2.48}$	****62.35**$^{+1.86}$	****66.99**$^{+1.74}$
EN	****62.97**$^{+0.69}$		****63.80**$^{+1.51}$	****66.47**$^{+0.93}$	**67.19**$^{+0.22}$
ES	59.88$^{-0.58}$	48.85$^{-0.49}$		**68.55**$^{+0.45}$	****65.33**$^{+0.66}$
FR	**61.59**$^{+0.32}$	53.12$^{-0.34}$	**67.00**$^{+0.49}$		****64.52**$^{+1.77}$
SV	62.16$^{-0.80}$	**51.31**$^{+0.24}$	***62.58**$^{+0.76}$	**65.38**$^{+0.39}$	

Table 3: Results for collapsing dependency relations over unary dummy nodes and removing dummy leaves (difference to the annotation projection baseline in superscript). Improvements marked with ** are statistically significant according to McNemar's test with $p < 0.01$ and improvements marked with * are statistically significant with $p < 0.05$.

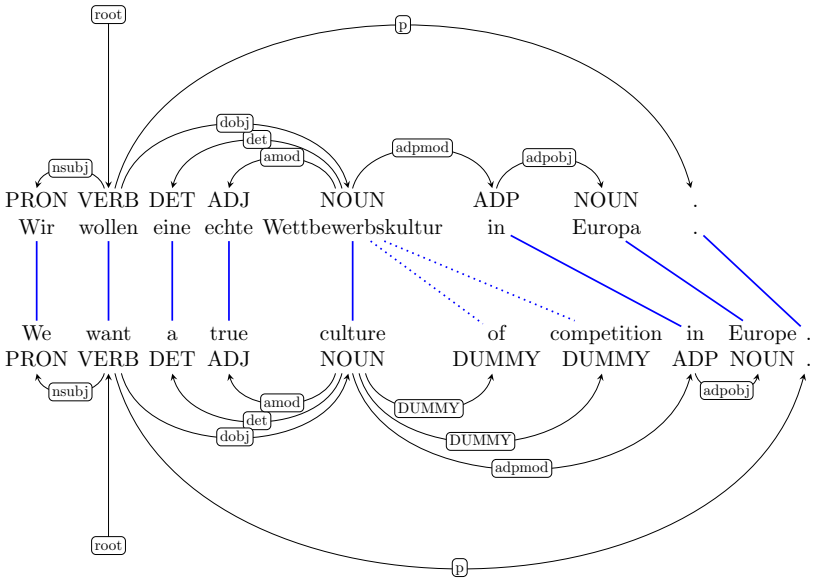

Figure 9: Projecting from German to English using an alternative treatment for one-to-many word alignments. Dotted lines are links from the grow-diag-final-and symmetrization heuristics and solid lines refer to links in the intersection of word alignments.

annotation projection. Table 4 summarizes the results of this method tested in our setup. We can observe significant improvements for all language pairs compared to the baseline approach and all but two cases are also better than the results of the previous setting shown in Table 3.

	DE	EN	ES	FR	SV
DE		$^{**}\mathbf{53.80}^{+0.53}$	$^{****}\mathbf{61.34}^{+3.65}$	$^{**}\mathbf{62.32}^{+1.83}$	$^{****}\mathbf{68.20}^{+2.95}$
EN	$^{***}\mathbf{63.52}^{+1.24}$		$^{***}\mathbf{63.18}^{+0.89}$	$^{**}\mathbf{67.04}^{+1.50}$	$^{****}\mathbf{67.74}^{+0.77}$
ES	$\mathbf{60.65}^{+0.19}$	$^{****}\mathbf{50.10}^{+0.76}$		$^{*}\mathbf{68.81}^{+0.71}$	$^{***}\mathbf{65.79}^{+1.12}$
FR	$^{****}\mathbf{62.49}^{+1.22}$	$^{***}\mathbf{53.88}^{+0.42}$	$^{****}\mathbf{68.15}^{+1.64}$		$^{**}\mathbf{64.83}^{+2.08}$
SV	$^{****}\mathbf{63.83}^{+0.87}$	$^{****}\mathbf{52.36}^{+1.29}$	$^{***}\mathbf{63.29}^{+1.47}$	$^{**}\mathbf{66.12}^{+1.13}$	

Table 4: Discarding trees that include DUMMY nodes; results with 40,000 accepted trees. Results marked with ** and * are significantly better than the projection baseline (with $p < 0.01$ and $p < 0.05$, respectively) and results marked **** and *** are also significantly better than the ones in Table 3 (with $p < 0.01$ and $p < 0.05$, respectively).

3.3 Phrase-Based Treebank Translation

Treebank translation is an interesting alternative to annotation projection. The main advantage is that we can skip noisy source-side annotation of an out-of-domain bitext to be able to project information from source to target language. Furthermore, word alignment is tightly coupled with most statistical translation models which makes it straightforward to use these links for projection. Finally, it is an advantage for projection that machine translation prefers literal translations in similar syntactic structures. Unrestricted human translations are much more varied and a proper alignment between translation equivalents is not necessarily straightforward. In machine translation, the mapping between tokens and token n-grams is essential which favors successful annotation projection. The largest drawback is, of course, translation quality. Machine translation is a difficult task on its own and its use in annotation projection requires at least some level of quality even though we are not necessarily interested in semantically adequate translations.

Our first approach applies the model proposed by Tiedemann et al. (2014), using a standard phrase-based SMT model to translate source language treebanks to a target language. The projection is based on the DCA heuristics similar to the ones applied to annotation projection described in the previous section. We also apply the modification of DUMMY node handling as introduced before. However, we cannot apply the alternative treatment of one-to-many alignments as we do not have different types of word alignment in our translation model. We also do not filter out trees with remaining DUMMY nodes as this would cause a serious reduction of the already small-sized treebanks. In contrast to projection with bitexts we cannot add more data to fill up the training data.

In all the experiments, our MT setup is very generic and uses the Moses toolbox for training, tuning and decoding (Koehn et al., 2007). The translation models are trained on the entire Europarl corpus version 7 without language-pair-specific optimization. Word alignments are essentially the same that we have used for our experiments with annotation projection in section 3.2. For tuning we use MERT (Och, 2003) and the newstest2011 data provided by the annual workshop on statistical machine translation (WMT).[1] For Swedish

1. http://www.statmt.org/wmt14.

we use a sample from the OpenSubtitles2012 corpus (Tiedemann, 2012). The language model is a standard 5-gram model and is based on a combination of Europarl and News data provided from the same source. We apply modified Kneser-Ney smoothing without pruning, applying KenLM tools (Heafield, Pouzyrevsky, Clark, & Koehn, 2013) for estimating the LM parameters.

	DE	EN	ES	FR	SV
DE		$^{**}\mathbf{56.24}^{+2.70}$	$^{**}57.65^{-2.52}$	$^{**}59.06^{-3.29}$	$^{**}64.62^{-2.37}$
EN	$^{**}59.41^{-3.56}$	–	$63.76^{-0.04}$	$^{**}\mathbf{67.99}^{+1.52}$	$\mathbf{67.52}^{+0.33}$
ES	$^{**}53.94^{-5.94}$	$^{**}\mathbf{50.65}^{+1.80}$		$^{**}\mathbf{69.70}^{+1.15}$	$^{**}62.73^{-2.60}$
FR	$^{**}57.05^{-4.54}$	$^{**}\mathbf{55.69}^{+2.57}$	$^{**}\mathbf{68.66}^{+1.66}$		$^{**}62.77^{-1.75}$
SV	$^{**}58.57^{-3.59}$	$^{**}\mathbf{53.01}^{+1.70}$	$\mathbf{62.69}^{+0.11}$	$64.76^{-0.62}$	

Table 5: Results for phrase-based treebank translation (difference to the corresponding annotation projection model with DUMMY node removal from Table 3 in superscript). Results marked with ** are significantly different from the projection results (with $p < 0.01$).

The results of our experiments with phrase-based SMT is summarized in Table 5. To a large extent, we can confirm the findings of Tiedemann (2014) that the translation approach has some advantages over the projection of automatically annotated parallel corpora. For some language pairs, the labeled attachment scores are significantly above the projection results even though the parsers are trained on much smaller data sets (the treebanks are typically much smaller than 40,000 sentences for most language pairs). Very striking is also the outcome for German as a target language, which seems to be the hardest language to translate to in this data set. This is not very surprising as German is in general considered to be a difficult target language in the setup of languages that are, for example, supported by WMT. This also applies to the use of German as a source language with a surprising exception when translating to English. Overall, the good results for English may be influenced by the strong impact of the language model that can draw from the large monolingual resources.

3.4 Syntax-Based Treebank Translation

Tiedemann (2015) introduces the use of syntax-based SMT as another alternative to treebank translation. The standard syntax-based MT models supported by Moses are based on synchronous phrase-structure grammars which are induced from word-aligned parallel data. Several modes are available. In our case, we are mostly interested in the tree-to-string models that use synchronous tree substitution grammars (STSGs). Our assumption is that the structural relations that are induced from the parallel corpus with a fixed given source-side analysis improve the projection of syntactic relations when used in combination with syntax-based translation.

In order to make it possible to use dependency information in the framework of synchronous STSGs we convert projective dependency trees to the bracketing structure that can be used to train tree-to-string models with Moses. We use the yield of each word to

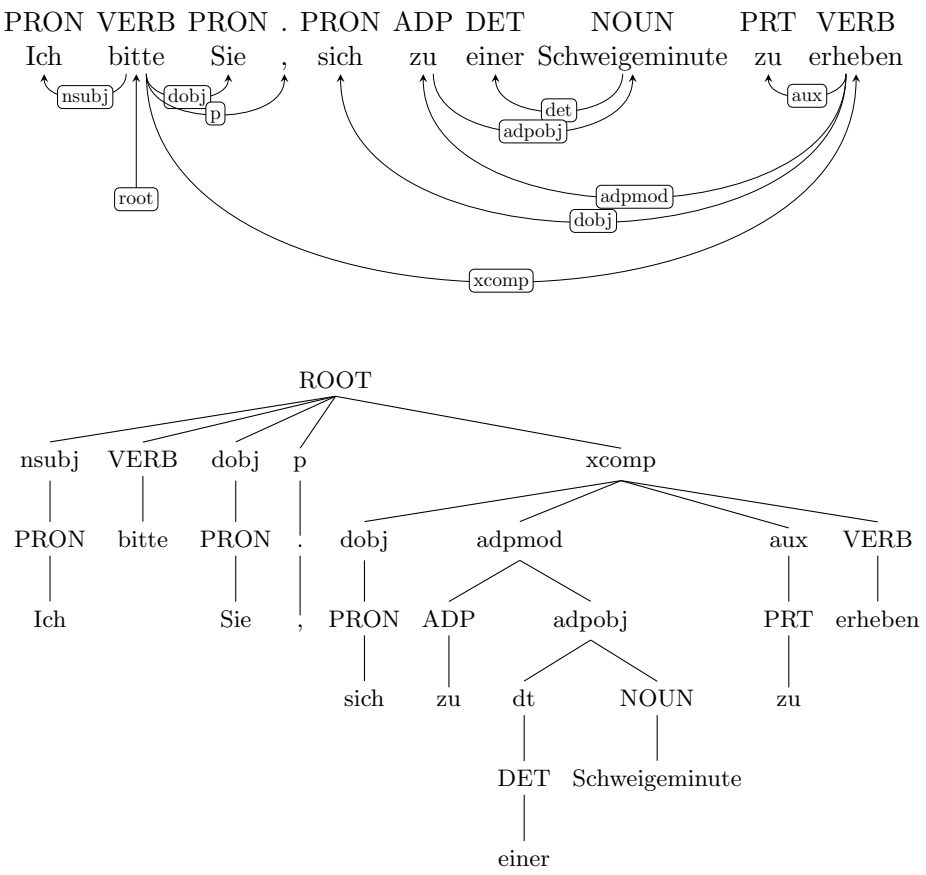

Figure 10: A dependency tree taken from the automatically annotated parallel data and its lossy conversion to a constituency representation.

define a span over the sentence which forms a constituent with the label taken from the relation of that word to its head.

Dependency trees are certainly not optimal for this kind of constituency-based SMT model as they are usually very flat and do not provide the deep hierarchical structures that are common in phrase-structure trees. However, our previous research has shown that valuable syntactic information can be pushed into the model in this way that can be beneficial for projecting dependency relations. Note that we use PoS tags as additional pre-terminal nodes to enrich the information given to the system.

For training the models we used the same data sets and word alignments as we have used for phrase-based SMT. However, we require a number of additional steps listed below:

- We tag the source side of a parallel corpus with a PoS tagger trained on the UDT training data using HunPos (Halácsy, Kornai, & Oravecz, 2007).

- We parse the tagged corpus using a MaltParser model trained on the UDT with a feature model optimized with MaltOptimizer (Ballesteros & Nivre, 2012).[2]
- We projectivize all trees using MaltParser and convert to nested tree annotations as explained above (Tiedemann, 2015).
- We extract synchronous rule tables from the word aligned bitext with source side syntax and score rules using Good Turing discounting. We do not use any size limit for replacing sub-phrases with non-terminals at the source side and restrict the number of non-terminals on the right-hand side of extracted rules to three. Furthermore, we allow consecutive non-terminals on the source side to increase coverage, which is not allowed in the default settings of the hierarchical rule extractor in Moses.
- We tune the model using MERT and the same data sets as before.
- Finally, we convert the training data of the UDT in the source language and translate it to the target language using the tree-to-string model created above.

The results of our approach are listed in Table 6. We can see that syntax-based models are superior to phrase-based models in almost all cases. For the majority of language pairs we can also see an improvement over the annotation projection approach even though the training data is much smaller. This confirms the findings of Tiedemann (2015) but outperforms their results by a large margin due to the parsing model used in our experiments.

	DE	EN	ES	FR	SV
DE		$^{**\dagger\dagger}\mathbf{58.60}^{+5.06}$	$^{**}\mathbf{61.00}^{+0.83}$	$^{**\dagger}\mathbf{63.45}^{+1.10}$	$^{**\dagger\dagger}\mathbf{67.88}^{+0.89}$
EN	$^{**}\mathbf{62.67}^{-0.30}$		$^{**\dagger}\mathbf{64.58}^{+0.78}$	$^{\dagger\dagger}\mathbf{68.45}^{+1.98}$	$^{**\dagger\dagger}\mathbf{68.16}^{+0.97}$
ES	$^{**\dagger\dagger}\mathbf{57.13}^{-2.75}$	$^{**\dagger\dagger}\mathbf{52.65}^{+3.80}$		$^{\dagger}69.37^{+0.82}$	$^{**\dagger\dagger}\mathbf{63.55}^{-1.78}$
FR	$^{**}\mathbf{61.41}^{-0.18}$	$^{**\dagger\dagger}\mathbf{56.83}^{+3.71}$	$^{\dagger}\mathbf{68.97}^{+1.97}$		$^{\dagger\dagger}62.56^{-1.96}$
SV	$^{**}\mathbf{61.73}^{-0.43}$	$^{**\dagger\dagger}52.13^{+0.82}$	$62.34^{-0.24}$	$^{\dagger}64.50^{-0.88}$	

Table 6: Results for syntax-based treebank translation (difference to the corresponding annotation projection model from Table 5 in superscript). Numbers in bold face are better than the corresponding phrase-based SMT model. Results marked with ** are significantly different from the phrase-based translation results ($p < 0.01$); † and †† are significantly different from the projection model ($p < 0.01$ and $p < 0.05$, respectively).

3.5 Translation and Back-Projection

Another possibility for cross-lingual parsing is the integration of translation in the actual parsing pipeline. The basic idea is to use tools in other languages, such as dependency parsers, without modification by adjusting the input to match the expectations of the tool,

2. We use MaltParser here for efficiency reasons. The parsing performance is slightly below the baseline models trained with `mate-tools` but parsing is very fast which we require for parsing all bitexts.

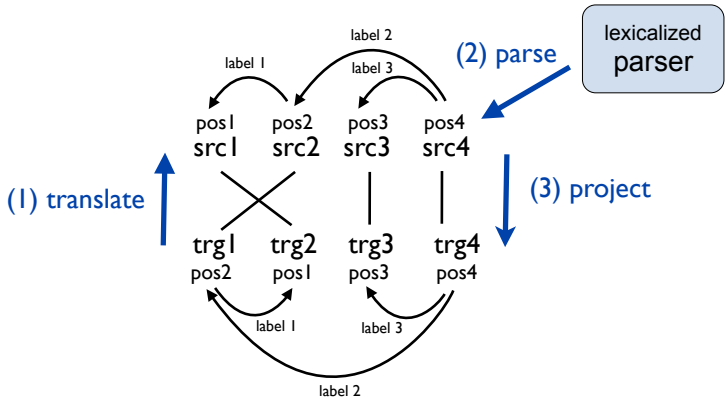

Figure 11: Translation and back-projection: Input data is translated to a source language with existing parsers (step 1), parsed in the source language (step 2) and, finally, the parse tree is projected back to the original target language.

for example, by translating it to the language that a parser accepts. This is very much in spirit of text normalization approaches that are frequently used in NLP for historical documents and user-generated content in which the input is modified in such a way that existing tools for standard language can be applied. Figure 11 illustrates the approach applied to dependency parsing.

The advantage of this approach is that we can rely on optimized parsers that are trained on manually corrected treebanks. However, there are several significant drawbacks. First of all, we loose efficiency due to the additional translation step that is required at parsing time. This is a crucial disadvantage that rules out this approach for many applications which require parsed information of large scale data sets or real-time responses. Another important drawback is the noise coming from translation leading to some kind of input, which a parser is usually not trained for and, therefore, has a hard time to handle correctly. Finally, there is also the problem of back-projection. Unfortunately, it is not straightforward to reverse the projection heuristics discussed earlier. We cannot introduce DUMMY nodes to fill gaps that are required for projecting the entire structure and DUMMY labels are not useful either. The projection heuristics discussed in section 3.2 help to avoid DUMMY nodes and, therefore, we apply these extensions in our experiments. Another problem is related to unaligned target words. In the DCA algorithm (including all modified versions discussed so far), these tokens are simply deleted and will not be attached to the dependency tree at all. This method, however, is not possible for back-projection in which all tokens need to be attached. For this reason, we implement a new rule that attaches each unaligned token to either its preceding or consecutive word if they are attached to the tree themselves. If this is not the case then we simply attach them to ROOT. Another problem is the label that should be added to that dependency and due to the lack of further knowledge we set the label to DUMMY. In this way, we do not get any credit in LAS but may at least improve our UASs. We test this approach using syntax-based SMT as our translation model. The results are listed in table 7.

	DE	EN	ES	FR	SV
DE	–	$35.92^{-17.35}$	$32.90^{-24.79}$	$36.68^{-23.81}$	$45.56^{-19.69}$
EN	$44.86^{-17.42}$	–	$48.08^{-14.21}$	$48.19^{-17.35}$	$51.74^{-15.23}$
ES	$36.69^{-23.77}$	$41.91^{-7.43}$	–	$54.78^{-13.32}$	$43.23^{-21.44}$
FR	$37.44^{-23.83}$	$42.00^{-11.46}$	$55.54^{-10.97}$	–	$42.39^{-20.36}$
SV	$36.84^{-26.12}$	$35.23^{-15.84}$	$31.96^{-29.86}$	$33.74^{-31.25}$	–

Table 7: Back-projection results in comparison to the annotation projection baseline from section 3.2 (Table 3).

The scores are very low, as they even fall behind those of the baseline delexicalized models. This extreme drop in performance is actually a bit surprising but considering the strong disadvantages discussed above this may be expected as well. Another reason for the extreme differences in performance is also the fact that we need to rely on predicted PoS labels in the translated data before piping them into the source language parser. This is certainly a strong disadvantage of the procedure and the comparison to evaluations based on gold standard PoS annotation is not entirely fair. See also section 3.8 for more discussions on the impact of PoS label accuracy on parsing performance.

3.6 Annotation Projection and Translation Quality

An interesting question is whether there is a correlation between translation quality and the performance of the cross-lingual parsers based on translated treebanks. As an approximation for treebank translation quality we computed BLEU scores over well-established MT test sets from the WMT shared task, in our case the newstest from 2012.[3]

Figure 12 illustrates the correlation between BLEU scores obtained on newstest data and LAS's of the corresponding cross-lingual parsers. First of all, we can see that the MT performance of phrase-based and syntax-based models is quite comparable with some noticeable exceptions in which syntax-based SMT is significantly better (French-English and French-Spanish, which is rather surprising). However, looking at most language pairs we can see that the increased parsing performance does not seem to be due to improvements in translation but rather due to the better fit of these models for syntactic annotation projection (see German, for example). Nevertheless, we can observe a weak correlation between BLEU scores and LAS within a class of models with one notable outlier, Spanish-English. This correlation reflects the importance of the syntactic relation between languages for the success of machine translation and annotation projection. Closely related languages like French and Spanish are on the top level in both tasks whereas French and Spanish do not map well to German. Translations to English are an exception in this evaluation. Translation models often work well in this direction whereas annotation projection to English underperforms in our experiments.

3. Note that we have to leave out Swedish for this test as there is no test set available for this language.

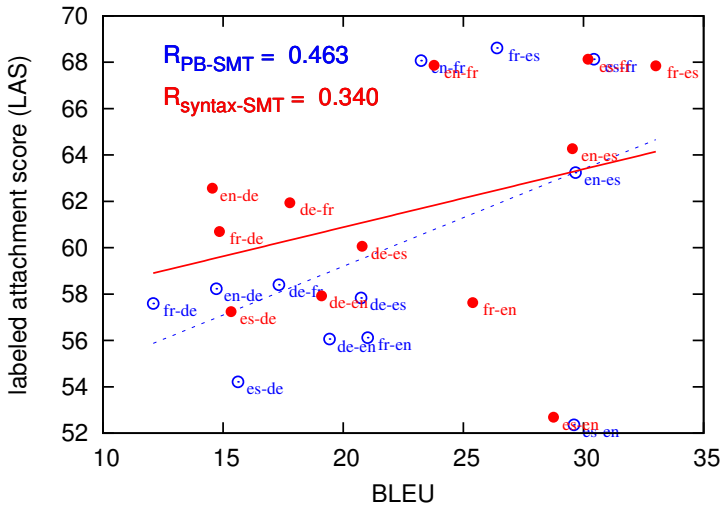

Figure 12: Correlation between BLEU scores and cross-lingual parsing accuracy (using Pearson's correlation coefficient).

3.7 System Combination and Multi-Source Models

So far, we were interested in transferring syntactic information from one source language to the target language using one specific model for cross-lingual parsing. However, the approaches above can easily be combined as they all focus on the creation of synthetic training data. There are at least two possibilities that can be explored.

1. We can combine data from several source languages to increase the amount of training data and to obtain evidence from various languages projected to the target language.

2. Several models can be combined to benefit from the various strengths of each model that may work as complementary information.

In this paper, we opt for a very simple approach to test these ideas. Here we concatenate data sets to augment our training data and train standard parsing models as usual. First, we will look at multi-source models within each paradigm. Table 8 lists the labeled attachment scores that we obtain when combining all data sets for all source languages to train target language parsers on the projected annotations.

From the table, we can see that we are able to achieve significant improvements for all languages and models except for Spanish. Furthermore, for English and for French we obtain the overall best result presented in this paper for the combined syntax-based SMT projections. In our final system combination, we now merge all data sets for all languages and models. The results of the parsers trained on these combined data sets are shown in Table 9.

4. These results are multi-source and multi-model system combinations provided by Tiedemann (2015).

LAS	DE	EN	ES	FR	SV
best published result[4]	60.94	56.58	68.45	69.15	68.95
best individual model	63.83	58.60	68.97	69.70	68.20
annotation projection	**66.76**	55.30	*67.37*	69.48	71.95
phrase-based SMT	61.85	60.94	*68.08*	71.54	71.69
syntax-based SMT	65.89	**61.56**	***68.60***	**72.78**	**72.14**

Table 8: Results for combining projected data of all source languages to train target language parsing models. Numbers in *italics* are worse than one of the models trained on data for individual language pairs.

	DE	EN	ES	FR	SV
LAS	**67.60**	57.05	**69.36**	72.03	**73.40**
UAS	75.27	64.54	76.85	79.21	81.28
ACC	81.99	72.75	82.22	83.06	83.04

Table 9: Results for combining projected data of all source languages to train target language parsing models. Additionally to LAS we also includes unlabeled attachment scores (UAS) and label accuracy (ACC) here to make it easier to compare our results with related work.

For German, French and Swedish this yields yet another significant improvement with labeled attachment scores close to 70% or even above. These results represent the highest scores that have been reported in this task so far and outperform previously published scores by a large margin. We expect that more sophisticated system combinations would push these results even further.

3.8 Gold vs. Predicted PoS Labels

It is common to evaluate results with gold PoS labels that are given in the test set of the target language treebank. This disregard for the impact of PoS quality—often present in related work—makes for a very unrealistic evaluation scenario. In the previous section, we discussed results that use gold standard annotation in order to make it possible to compare our results with the baselines and related work. In this section, we look into more details when replacing PoS labels with predicted values. Here, we report only the results for the treebank translation approach using syntax-based SMT as a test case. The other approaches show similar trends.

The first experiment looks at the case where annotated data is available for the target language for training PoS taggers. We use HunPos (Halácsy et al., 2007) to train models on the training data of each language and use them to replace the gold standard tags in all test sets with PoS labels that our models predict. The results of these experiments applied to the translated treebanks from section 3.4 are shown in Table 10.

	DE	EN	ES	FR	SV
DE	–	$56.49^{-2.11}$	$57.52^{-3.48}$	$59.99^{-3.46}$	$62.68^{-5.20}$
EN	$58.70^{-3.97}$	–	$61.25^{-3.33}$	$64.32^{-4.13}$	$63.97^{-4.19}$
ES	$53.37^{-3.76}$	$50.89^{-1.76}$	–	$65.24^{-4.13}$	$58.78^{-4.77}$
FR	$57.18^{-4.23}$	$54.94^{-1.89}$	$64.32^{-4.65}$	–	$58.22^{-4.34}$
SV	$57.63^{-4.10}$	$50.17^{-1.96}$	$59.36^{-2.98}$	$60.89^{-3.61}$	–
PoS tagger	95.24	97.56	95.37	95.08	95.86

Table 10: Results for cross-lingual parsing with predicted PoS labels coming from taggers trained on target language treebanks. The numbers in superscript give the difference to the result with gold standard labels (Table 6). The last row shows the overall accuracy of the PoS tagger.

We can see that PoS labels have a strong impact on parsing performance. For all language pairs, we can observe a significant drop in LAS even with quite accurate taggers, which proves that one need to be careful with applying models in real-life scenarios. The next experiment stresses this point even more. Here, we replace PoS labels with tags that are predicted by taggers that are trained on the noisy translated treebanks and their projected annotation. Note that we need to remove training examples with DUMMY labels to reduce errors of the tagger.

	DE	EN	ES	FR	SV
DE	–	81.32	81.23	82.41	84.29
EN	85.33	–	84.41	85.56	86.32
ES	82.39	81.05	–	89.37	83.26
FR	83.76	80.64	89.95	–	84.11
SV	84.79	81.66	86.05	84.81	–

Table 11: PoS tagging accuracy for models trained on translated treebanks.

Table 11 lists the accuracy of the taggers trained on noisy projected data. We can observe a significant drop in tagger performance which is completely plausible considering the substantial noise added through translation and projection and also considering the limited size of the data we use for training. Treebanks are considerably smaller than annotated corpora that are usually taken for training PoS classifiers. When applying these taggers to our test sets we can observe a dramatic drop in parsing performance as expected. Table 12 lists the results of these experiments.

From the above findings we can conclude that cross-lingual techniques still require a lot of improvement to become practically useful in low-resource scenarios in the real world. We have done the same experiment for the annotation projection approach and observed the same behavior even though we can rely on larger data sets for training the taggers. The performance drop of using predicted PoS labels trained on noisy data sets amounts to over

	DE	EN	ES	FR	SV
DE	–	$46.04^{-10.45}$	$48.61^{-8.91}$	$50.36^{-9.63}$	$52.73^{-9.95}$
EN	$51.89^{-6.81}$	–	$59.37^{-1.88}$	$62.37^{-1.95}$	$60.43^{-3.54}$
ES	$44.59^{-8.78}$	$47.81^{-3.08}$	–	$59.81^{-5.43}$	$52.12^{-6.66}$
FR	$49.72^{-7.46}$	$49.04^{-5.90}$	$61.30^{-3.02}$	–	$51.10^{-7.12}$
SV	$47.94^{-9.69}$	$44.23^{-5.94}$	$55.02^{-4.34}$	$52.79^{-8.10}$	–

Table 12: Results for cross-lingual parsing with predicted PoS labels coming from taggers trained on projected treebanks. The difference to the results with predicted labels from Table 10 are shown in superscript.

10 LAS points in most cases similar to what we see in the treebank translation approach. We omit the results as they do not add any new information to our discussion.

Finally, we also need to check whether system combinations and multi-source models help to improve the quality of cross-lingual parsers with predicted PoS labels. For this, we use the same strategy as in section 3.7 and concatenate the various data files to train parser models that combine all models and language pairs. In other words, we use the same models trained in section 3.7 but evaluate them on test sets that are automatically tagged with PoS labels. Again, we use two settings: 1) We apply PoS taggers trained on manually verified data sets—the monolingual target language treebanks, and 2) we use PoS taggers trained on projected and translated treebanks. For the latter we have now all data sets at our disposal and, therefore, expect a better PoS model as well. Table 13 lists the final results in comparison to the ones obtained with gold standard annotation.

	DE	EN	ES	FR	SV
monolingual baseline with gold PoS	78.38	91.46	82.30	82.30	84.52
delexicalized monolingual with gold PoS	70.84	82.44	71.45	73.71	74.55
best delexicalized cross-lingual with gold PoS	52.53	48.24	62.66	62.39	59.42
best cross-lingual model with gold PoS	**67.60**	**61.56**	**69.36**	**72.78**	**73.40**
monolingual PoS tagger accuracy	95.24	97.56	95.37	95.08	95.86
combined projected PoS tagger accuracy	88.47	88.24	88.06	89.83	88.07
monolingual baseline with predicted PoS	73.03	88.38	76.59	76.79	77.83
delexicalized monolingual with predicted PoS	64.25	72.81	60.49	64.06	65.77
best delexicalized cross-lingual with predicted PoS	48.36	43.87	52.94	52.47	49.84
combined cross-lingual with predicted PoS	**63.14**	**55.16**	**64.99**	**67.91**	**67.93**
combined cross-lingual with projected PoS model	**57.84**	**51.66**	**61.40**	**63.86**	**61.58**

Table 13: A comparison between models evaluated with gold standard PoS annotation (four top-level systems) and models tested against automatically tagged data.

First of all, we can see that our best cross-lingual models outperform delexicalized cross-lingual models by a large margin. They come very close to delexicalized models trained

on target language data with the exception of English which works much better with the original data set. In the lower part of the table, we observe that the scores drop significantly when gold standard PoS labels are replaced with predicted tags. Note that the four systems using predicted PoS labels apply the tagger trained on monolingual verified target language data which gives quite high accuracy. The final system in the table is the only one that applies the PoS model trained on projected and translated data. These tagger models are much less accurate, as shown in the middle of Table 13, and the influence of this degradation is visible in the attachment scores obtained by the systems. However, these models reflect a real-world scenario where no annotated is available for the target language, not even for training PoS taggers. The advantage of the projection and translation approaches is that such model is possible at all, whereas delexicalized and other transfer models always require existing tools that can produce the shared features used by the prediction system. Note also that the cross-lingual models now outperform some of the delexicalized models trained on verified target language data—with English as a striking exception—which is remarkable given the noisy data they are trained on.

3.9 Impact of Dataset Sizes

By and large, the data-driven dependency parsers benefit from introducing additional training data. In this subsection, we control for the amount of training data provided to our method, and observe the impact it has on LAS cross-lingually. We experiment with improved annotation projection (see Section 3.2), and we introduce up to 60 thousand sentences with projected dependency trees. For each of the five target UDT languages in the experiment, we provide four learning curves representing the four source languages. We plot the results in Figure 13.

We observe that virtually all transferred parsers benefit from the introduction of additional training data, albeit some of the improvements are only slight as some models level out at around 20 thousand sentences. All the source languages follow the same LAS learning curve patterns for all the targets, as we do not observe any trend violations for specific source-target pairs. Other than that, we observe clear source-target preferences, as the source orderings by LAS mostly remain the same for all training set sizes. Some of the lower-ranked sources do not benefit or even degrade by introducing more training data, for example, the Spanish parser induced from German data, or the English parser created by projecting Swedish trees. That said, it is worth noting that in the best source-target pairs, the targets always benefit from introducing more source data: German from English and Swedish, English from German and French, Spanish from French and vice versa, and Swedish from German and English. This is a very clear indicator for future improvements, as the method apparently benefits from adding more data. At the same time, our learning curves show benefits for truly under-resourced languages, as the largest relative gains are already reached at relatively modest quantities of 20 thousand sentence pairs. Moreover, the typological groupings in the former list of top-performing source-target pairs are quite apparent, as is the case throughout our experiments.

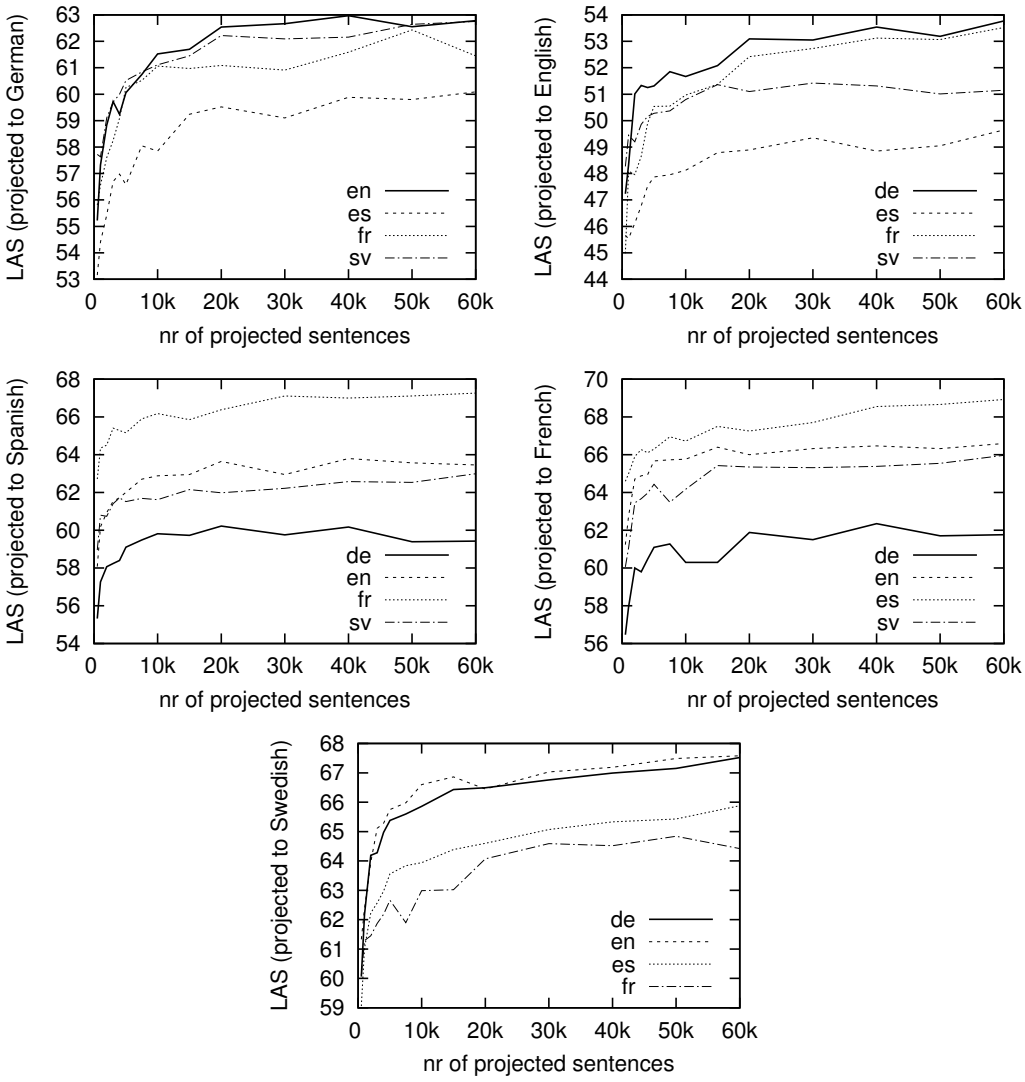

Figure 13: The impact of training data: Different sizes of projected data for training cross-lingual parsing models.

4. Comparison to Related Work

In this section—having thoroughly analyzed synthetic treebanking—we revert to a top-level discussion of cross-lingual parsing. In it, we contrast our approach to several selected alternatives from related work, and we sketch their properties from the viewpoint of enabling dependency parsing for truly under-resourced languages. We proceed by outlining the comparison.

We have already compared the various synthetic treebanking approaches to one another and to the delexicalized transfer baseline of McDonald et al. (2013) in section 3. Here, we aim at introducing a number of top-performing representatives of the methods discussed in

the overview section: a more competitive model transfer approach, an approach dealing with distributed word representations, and an annotation projection-motivated approach. As replicating all the approaches would be very time-consuming, we constrain our search to the approaches that also report their scores on UDT version 1 in their respective publication, as we can then compare by referencing. We select the following approaches for our discussion.

- Delex: This is the delexicalized model transfer baseline of McDonald et al. (2013). We report the scores by Søgaard et al. (2015) who used the arc-factored adaptation of the `mate-tools` parser, and not our replication or the original, as they conveniently report multiple metrics. We discuss the metrics below, and we note that they used gold PoS.

- Multi: A reimplementation of McDonald et al. (2011) multi-source projected system (multi-proj. in the original paper) by Ma and Xia (2014). We provide it as a more competitive baseline system. The original work predates UDT and only evaluates on the heterogenous CoNLL treebanks, but Ma and Xia (2014) evaluate it on the UDT treebanks so we report their scores. Note that the parsing model and preprocessing is then inherent to their setup, differing from the original setup of McDonald et al. (2011). The setup details are described further in the text, under Xia.

- Proj: The improved annotation projection approach we described in section 3.2. It is the final approach of the subsection, in which the dependency relations over unary dummy nodes are collapsed, dummy leaves removed, and all Europarl trees with remaining dummy nodes discarded (see Table 4). These scores are given with gold PoS tags.

- Trans G & P: We report on our best syntax-based cross-lingual treebank translation scores with gold and predicted PoS, respectively. Our PoS predictions come from an HMM tagger (Halácsy et al., 2007). The taggers are trained on target language treebanks, and they score at 95% on average (see Table 10).

- Comb G & P: These are our multi-source syntax-based cross-lingual parsers. They build on the Trans G & P approaches: instead of just single sources, multiple treebanks are translated into the target languages, providing combined synthetic treebanks to train parsers on. As before, we also report scores with gold and HunPos-predicted PoS.

- Rosa: This is the multi-source delexicalized transfer approach of Rosa and Zabokrtsky (2015), in its weighted variant. In their method, each target is parsed by multiple sources, and each parse is assigned a weight based on an empirically established language similarity metric. For each target sentence, the multiple parses constitute a digraph, on top of which a (Sagae & Lavie, 2006)-style maximum spanning tree voting scheme is implemented. They use gold PoS tags.

- Søgaard: In this model, delexicalized model transfer is augmented by inter-lingual word representations based on inverted indexing via Wikipedia concept links (Søgaard et al., 2015). We choose it as a very recent and illustrative example of leveraging word

Target language	Baselines		Synthetic treebanking					Recent approaches		
	Delex	Multi	Proj	TransG	TransP	CombG	CombP	Rosa	Søgaard	Xia
de	56.80	69.21	72.65	70.62	67.59	75.27	71.79	56.80	56.56	74.01
en	–	–	62.79	65.10	63.62	64.54	63.15	42.60	–	–
es	63.21	72.57	74.92	75.71	72.16	76.85	73.20	72.70	64.03	75.60
fr	66.00	74.60	76.13	76.33	72.95	79.21	76.06	–	66.22	76.93
sv	67.49	75.87	76.96	76.98	73.61	81.28	76.83	50.80	67.32	79.27

Table 14: Comparison of cross-lingual parsing methods. In contrast to the rest of our paper, here we report UAS scores to attain maximum coverage of results reported in related work.

embeddings for improving cross-lingual dependency parsing. They use an embeddings-enabled version of Bohnet's parser (Bohnet, 2010) and gold PoS tags. We report their multi-source results.

- Xia: The approach by Ma and Xia (2014) is a novel method that leverages Europarl to train probabilistic parsing models for resource-poor languages by maximizing a combination of likelihood on parallel data and confidence on unlabeled data. We report on their best approach (marked as +U in their paper), which makes use of both parallel and unlabeled data. They use top-performing PoS taggers trained on the target languages, each of them reaching at least a 95% accuracy.

Before discussing the results, we make a number of remarks on the comparison. First, for each target language, we report the best obtained score for each method, rather than possibly misleading averages or more complex source-target matrices. In most related work, English is not used as a target language. Second, in contrast to the remainder of the paper—and contrary to the guidelines for evaluating cross-lingual parsers following McDonald et al. (2013)—we report on UAS only. This is targeted exclusively at facilitating the comparison to related work, as these contributions for the most part still report UAS scores, even when working with UDT. While we do see this as unfortunate, we also note that a LAS-enabled replication study exceeds the scope and does not match the focus of our contribution. Third, and also related to not being able to control for all the experiment parameters, we note the issue of reporting scores on gold and predicted PoS, and the different ways of obtaining the predicted annotations. We record the differences in the list above. Finally, we note that some of the referenced contributions do not explicitly state whether their scoring included punctuation or not, whereas we do include it in our experiments.

The results are given in Table 14 and we now proceed to discuss them in more detail, reflecting on the methods' intricacies and requirements in the process.

In the table, we visually group the methods into the baselines (Delex, Multi), our proposed approaches (Proj, Trans, Comb), and selected recent contributions to cross-lingual dependency parsing (Rosa, Søgaard, Xia). By design, we do not highlight the best scores, as not all the results are directly comparable, especially with respect to the lack of control for sources of features facilitating the parsing, such as the PoS tags. We also note that Rosa is evaluated on the HamleDT treebanks (Rosa, Mašek, Mareček, Popel, Zeman,

& Žabokrtský, 2014) and not UDT, but we still provide it for reference, as it implements an interesting addition to Delex as a sort of an intermediate step towards Multi.

We first observe that Rosa and Søgaard rarely surpass our Delex baseline. This does not come as a surprise, as our baseline uses a more advanced graph-based dependency parser (Bohnet, 2010): in contrast, Rosa uses an arc-factored parser (McDonald, Pereira, Ribarov, & Hajič, 2005), while Søgaard implements a first-order version of the parser by Bohnet (2010) that leverages cross-lingual word representations. That said, the discrepancy between the first- and second-order graph-based parsers appears not to be the only factor in explaining the slight (if any) gains provided by these two approaches. Namely, Rosa is an approach to multi-source delexicalized parsing based on maximum spanning tree-style voting, and it uses empirically obtained dataset similarity metrics for weighting the arcs in the voting schemes. As such, even if it yields slight improvements over the respective fair baselines—as provided in the paper describing the approach (Rosa & Zabokrtsky, 2015)—it is still bound by the impoverished feature representation informing the parser, inherited from the Delex it builds on, preventing the method from reaching higher accuracies. Søgaard attempts to alleviate this by introducing cross-lingual word representations to the feature space. In their report on the approach, Søgaard et al. (2015) observe slight improvements over the baselines, but it is apparent that the word representations they utilize work much better for NLP tasks that don't involve syntactic representations, indicating they might not be appropriate for facilitating cross-lingual parsing more substantially.

Having considered Rosa and Søgaard—comparing the two approaches to the Delex baseline, and establishing their inferiority to the remaining approaches, including synthetic trebanking—we turn to the more interesting part of the discussion, in which our contributions are compared to one another, and to Xia. We also include the competitive Multi baseline of McDonald et al. (2011) to this discussion.

Our improved annotation projection Proj appears to be a very competitive method, as none of the other approaches surpass it by a large margin. It also consistently beats Multi, albeit their PoS annotations are not comparable. Syntax-based treebank translation (Trans) surpasses it by a very narrow margin on four out of five targets, with German as the exception, while the multi-source variant (Comb) adds approximately 3-5 LAS points to the difference, with English as the exception. Only the approaches using predicted PoS tags are contrasted to Xia, but noting that on these datasets, our tagging approach (HunPos) performs slightly under theirs (Stanford) on average. We observe that Xia exhibits a slight advantage over out top approach (CombP) across the targets, but we also note—on top of the differences in taggers—that their approach also utilizes unlabeled data for semi-supervised parser augmentation. That said, Ma and Xia (2014) document only minor decreases when removing the unlabeled sources, and they implement an arc-factored dependency parser in the pipeline. Thus, we note that i) our synthetic treebanking approaches and Xia currently represent the most competitive approaches to cross-lingual dependency parsing, with a slight empirical edge for the latter, and that ii) further research is needed—in the form of an extensive replicative survey of cross-lingual parsing—to empirically gauge the various intricacies of these two approaches, and other influential contributions to the field, such as the work of McDonald et al. (2011) or Xiao and Guo (2014). We also note a very recent contribution by Rasooli and Collins (2015), which also deals with parallel corpora and projections, showing very promising results.

At this point, from the viewpoint of enabling the processing of truly under-resourced languages, it is interesting to mark the following observation. In Table 14, there is an apparent disconnect in scores between the methods that exploit parallel data sources (Multi, Proj, Trans, Comb, Xia), and the methods that don't (Delex, Rosa, Søgaard): the methods that make use of the parallel resources all perform significantly better. This is a clear indicator that for reaching top-level cross-lingual parsing performance, at least with the current line-up of standard dependency parsers, we need the lexical features provided by parallel corpora. The observation appears to us as a clear guideline for future work in cross-lingual parsing, and in the enablement of NLP for under-resourced languages.

5. Conclusions and Future Work

In this paper we discussed the various approaches for cross-lingual dependency parsing, reviewing and comparing a number of commonly used methods. Furthermore, we included an extensive study of annotation projection and treebank translation, and presented very competitive results in cross-lingual dependency parsing for the task of parsing data with cross-lingually harmonized annotation as included in the Universal Dependency Treebank.

Our future work includes the incorporation of cross-lingual word embeddings in model transfer as another component of the system combinations we discuss in the paper. We will also look at a wider range of languages using the growing set of harmonized data sets in the Universal Dependencies project. Especially interesting is the use of our techniques for truly under-resourced languages. We will explore cross-lingual parsing as a means of bootstrapping tools for those languages. We also aim at implementing a large-scale replicative survey of cross-lingual dependency parsing, as we show in our contribution that such an empirical assessment would be very timely and beneficial to this fast-developing field.

Acknowledgements

We thank the four anonymous reviewers for their detailed comments, which significantly contributed to improving the quality of the publication. We also acknowledge Joakim Nivre for the discussions on synthetic treebanking, and Héctor Martínez Alonso for his suggestions on improving the readability of the paper.

References

Abeillé, A. (2003). *Treebanks: Building and Using Parsed Corpora*. Springer.

Agić, Ž., Merkler, D., & Berović, D. (2012). Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing. In *Proceedings of IS-LTC*, pp. 5–9.

Agić, Ž., Hovy, D., & Søgaard, A. (2015). If All You Have is a Bit of the Bible: Learning POS Taggers for Truly Low-resource Languages. In *Proceedings of ACL*, pp. 268–272.

Agić, Ž., Tiedemann, J., Merkler, D., Krek, S., Dobrovoljc, K., & Može, S. (2014). Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets. In *Proceedings of LT4CloseLang*, pp. 13–24.

Ballesteros, M., & Nivre, J. (2012). MaltOptimizer: An Optimization Tool for MaltParser. In *Proceedings of EACL*, pp. 58–62.

Bender, E. M. (2011). On Achieving and Evaluating Language-independence in NLP. *Linguistic Issues in Language Technology*, *6*(3), 1–26.

Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers.

Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B. (2003). The Prague Dependency Treebank. In *Treebanks*, pp. 103–127. Springer.

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of COLING*, pp. 89–97.

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL*, pp. 149–164.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Das, D., & Petrov, S. (2011). Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. In *Proceedings of ACL*, pp. 600–609.

De Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, pp. 449–454.

Durrett, G., Pauls, A., & Klein, D. (2012). Syntactic Transfer Using a Bilingual Lexicon. In *Proceedings of EMNLP-CoNLL*, pp. 1–11.

Elming, J., Johannsen, A., Klerke, S., Lapponi, E., Martinez Alonso, H., & Søgaard, A. (2013). Down-stream Effects of Tree-to-dependency Conversions. In *Proceedings of NAACL*, pp. 617–626.

Faruqui, M., & Dyer, C. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of EACL*, pp. 462–471.

Garrette, D., Mielens, J., & Baldridge, J. (2013). Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages.. In *Proceedings of ACL*, pp. 583–592.

Gouws, S., & Søgaard, A. (2015). Simple Task-specific Bilingual Word Embeddings. In *Proceedings of NAACL*.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos – An Open-source Trigram Tagger. In *Proceedings of ACL*, pp. 209–212.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, pp. 690–696.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, *11*(3), 311–325.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing Crosslingual Distributed Representations of Words. In *Proceedings of COLING*, pp. 1459–1474.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit*, pp. 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pp. 177–180.

Koo, T., Carreras, X., & Collins, M. (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of ACL*, pp. 595–603.

Kübler, S., McDonald, R., & Nivre, J. (2009). *Dependency Parsing*. Morgan & Claypool Publishers.

Li, S., Graça, J. V., & Taskar, B. (2012). Wiki-ly Supervised Part-of-speech Tagging. In *Proceedings of EMNLP-CoNLL*, pp. 1389–1398.

Ma, X., & Xia, F. (2014). Unsupervised Dependency Parsing with Transferring Distribution via Parallel Guidance and Entropy Regularization. In *Proceedings of ACL)*, pp. 1337–1348.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*, pp. 92–97.

McDonald, R., Pereira, F., Ribarov, K., & Hajič, J. (2005). Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of EMNLP*, pp. 523–530.

McDonald, R., Petrov, S., & Hall, K. (2011). Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP*, pp. 62–72.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. http://arxiv.org/pdf/1309.4168.pdf.

Naseem, T., Barzilay, R., & Globerson, A. (2012). Selective Sharing for Multilingual Dependency Parsing. In *Proceedings of ACL*, pp. 629–637.

Nivre, J. (2006). *Inductive dependency parsing*. Springer.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., & Ginter, F. e. a. (2015). Universal dependencies 1.0..

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 915–932.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pp. 160–167.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19–52.

Petrov, S. (2014). Towards Universal Syntactic Processing of Natural Language. In *Proceedings of LT4CloseLang*, p. 66.

Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of LREC*, pp. 2089–2096.

Plank, B., Martínez Alonso, H., Agić, v., Merkler, D., & Søgaard, A. (2015). Do Dependency Parsing Metrics Correlate with Human Judgments?. In *Proceedings of CONLL*, pp. 315–320.

Rasooli, M. S., & Collins, M. (2015). Density-Driven Cross-Lingual Transfer of Dependency Parsers. In *Proceedings of EMNLP*.

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., & Žabokrtský, Z. (2014). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of LREC*, pp. 2334–2341.

Rosa, R., & Zabokrtsky, Z. (2015). KLcpos3 - a Language Similarity Measure for Delexicalized Parser Transfer. In *Proceedings of ACL*, pp. 243–249.

Sagae, K., & Lavie, A. (2006). Parser Combination by Reparsing. In *Proceedings of NAACL*, pp. 129–132.

Søgaard, A. (2011). Data Point Selection for Cross-language Adaptation of Dependency Parsers. In *Proceedings of ACL*, pp. 682–686.

Søgaard, A. (2012). Unsupervised Dependency Parsing Without Training. *Natural Language Engineering*, *18*(02), 187–203.

Søgaard, A. (2013). *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool Publishers.

Søgaard, A., Agić, Ž., Martínez Alonso, H., Plank, B., Bohnet, B., & Johannsen, A. (2015). Inverted Indexing for Cross-lingual NLP. In *Proceedings of ACL*, pp. 1713–1722.

Täckström, O., Das, D., Petrov, S., McDonald, R., & Nivre, J. (2013a). Token and Type Constraints for Cross-lingual Part-of-speech Tagging. *Transactions of the Association for Computational Linguistics*, *1*, 1–12.

Täckström, O., McDonald, R., & Nivre, J. (2013b). Target Language Adaptation of Discriminative Transfer Parsers. In *Proceedings of NAACL*, pp. 1061–1071.

Täckström, O., McDonald, R., & Uszkoreit, J. (2012). Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of NAACL*, pp. 477–487.

Tiedemann, J. (2014). Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proceedings of COLING*, pp. 1854–1864.

Tiedemann, J., Agić, Ž., & Nivre, J. (2014). Treebank Translation for Cross-Lingual Parser Induction. In *Proceedings of CoNLL*, pp. 130–140.

Tiedemann, J., & Nakov, P. (2013). Analyzing the use of character-level translation with sparse and noisy datasets. In *Proceedings of RANLP*, pp. 676–684.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of LREC*, pp. 2214–2218.

Tiedemann, J. (2015). Improving the Cross-Lingual Projection of Syntactic Dependencies. In *Proceedings of NoDaLiDa*.

Uszkoreit, H., & Rehm, G. (2012). *Language White Paper Series*. Springer.

Xiao, M., & Guo, Y. (2014). Distributed Word Representation Learning for Cross-Lingual Dependency Parsing. In *Proceedings of CoNLL*, pp. 119–129.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In *Proceedings of HLT*, pp. 1–8.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtskỳ, Z., & Hajič, J. (2014). HamleDT: Harmonized Multi-language Dependency Treebank. *Language Resources and Evaluation*, *48*(4), 601–637.

Zeman, D., & Resnik, P. (2008). Cross-Language Parser Adaptation between Related Languages. In *Proceedings of IJCNLP*, pp. 35–42.

Zhang, Y., Reichart, R., Barzilay, R., & Globerson, A. (2012). Learning to Map into a Universal POS Tagset. In *Proceedings of EMNLP-CoNLL*, pp. 1368–1378.

Zhao, H., Song, Y., Kit, C., & Zhou, G. (2009). Cross Language Dependency Parsing Using a Bilingual Lexicon. In *Proceedings of ACL-IJCNLP*, pp. 55–63.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of EMNLP*, pp. 1393–1398.