Delexicalized Cross-lingual Dependency Parsing for Xibe

He Zhou and **Sandra Kübler**
Indiana University
{hzh1, skuebler}@iu.edu

Abstract

Manually annotating a treebank is time-consuming and labor-intensive. We conduct delexicalized cross-lingual dependency parsing experiments, where we train the parser on one language and test on our target language. As our test case, we use Xibe, a severely under-resourced Tungusic language. We assume that choosing a closely related language as the source language will provide better results than more distant relatives. However, it is not clear how to determine those closely related languages. We investigate three different methods: choosing the typologically closest language, using `LangRank`, and choosing the most similar language based on perplexity.

We train parsing models on the selected languages using `UDify` and test on different genres of Xibe data. The results show that languages selected based on typology and perplexity scores outperform those predicted by `LangRank`; Japanese is the optimal source language. In determining the source language, proximity to the target language is more important than large training sizes. Parsing is also influenced by genre differences, but they have little influence as long as the training data is at least as complex as the target.

1 Introduction

For a severely low-resource language, constructing a dependency treebank is labor-intensive and time-consuming, and annotators are difficult to find. Expanding a small treebank via monolingual dependency parsing leads to suboptimal results since we lack enough training data to train a reliable parser. This situation has led to an increasing interest in techniques for supporting low-resource languages by taking advantage of high-resource languages together with methods for cross-lingual transfer (Meechan-Maddon and Nivre, 2019). This is facilitated by the Universal Dependencies (UD) project, which has resulted in a treebank collection covering a wide range of language, with the goal of facilitating multilingual parser development (Nivre et al., 2020). The latest release (v2.7) covers 183 treebanks in 104 languages (Zeman, 2020). In our current work, we carry out preliminary single-source cross-lingual delexicalized dependency parsing experiments for the Xibe language. With this method, we train a parser on the treebank of one *source* language and parse the *target* language, with both treebanks delexicalized to abstract away from lexical differences between the two languages.

Choosing the source language is crucial for single-source cross-lingual parsing. The optimal source language needs to be syntactically close to the target language as well as high-resourced. However, it is not obvious how to select this language. We investigate three methods for selecting the source language: We compare `LangRank` (Lin et al., 2019) and typology, and we investigate whether using perplexity as a similarity metric can approximate typological knowledge. Then we investigate whether the size of the source treebank or a genre mismatch affect the quality of the parser.

The remainder of this paper is organized as follows: Section 2 provides a short overview of Xibe syntax. In Section 3, we describe our research questions in more detail. In Section 4, we briefly summarize methods of cross-lingual transfer. The experimental settings are introduced in Section 5. We then explain the methods for selecting source languages in Section 6, and in Section 7, we discuss our results. We conclude in Section 8.

2 The Xibe Language and Treebank

Xibe is a Tungusic language. There are twelve languages in the Tungusic language family spoken in

Proceedings of Recent Advances in Natural Language Processing, pages 1626–1635
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_182

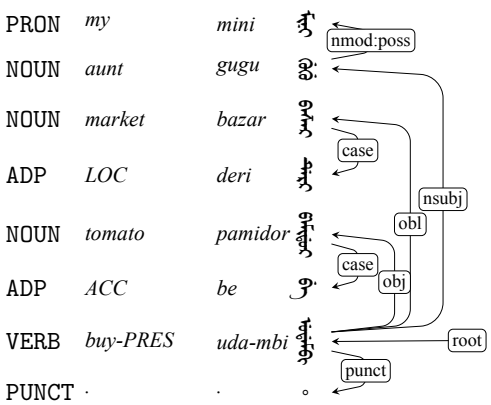

Figure 1: Dependency tree for 'My aunt buys tomatoes at the market'.

Central and Eastern Asia, but the numbers of Tungusic speakers have never been large (Robbeets and Savelyev, 2020). The language of Xibe is the one with a comparatively larger amount of active speakers in the whole language family.

Xibe shares morphological and syntactic features with other transeurasian languages. The transeurasian languages have a very rich system of case marking through the use of affixes (or particles in Japanese and Korean). All transeurasian languages are head final, they use verb-final word order, and attributes, complements, and adjuncts precede their headwords. In Xibe, clausal constituents have a rigid Subject-Object-Verb (SOV) word order, and all phrasal categories are consistently head-final. Like other Tungusic languages, Xibe has agglutinative morphology, which mainly focuses on verbs in that verbs are marked for tense, aspect, mood and voice, as well as converbs and participles.

Zhou et al. (2020) describe a Xibe treebank annotated in the Universal Dependencies framework, containing 810 trees. Figure 1 shows an example Xibe dependency tree. The matrix predicate is at the sentence final position with the object and oblique constituent, marked by corresponding case markers, preceding the verb and the present tense suffix *-mbi* attached to the verb root.

3 Research Questions

Our research focuses on cross-lingual dependency parsing using a single target language and concentrates on determining the best method to select the optimal source language. More specifically, we investigate the following questions.

Question 1 What are the most important factors to consider when choosing the best source language(s)? We investigate three methods: one uses typological knowledge, the second uses `LangRank` (Lin et al., 2019), a machine learning approach to rank languages based on their relatedness. The third method uses POS n-gram perplexity to determine similarity between languages. Here, our goal is to determine whether perplexity can be used to model typological knowledge.

Question 2 Ideally, an optimal source language should be closely related to the target language as well as high-resourced, since we need a sizable treebank that can be used for training the parser. However, is a large treebank size more important than syntactic similarity with the target language in delexicalized dependency parsing?

Question 3 The Xibe treebank includes sentences from two different genres, grammar examples and news whereas most UD treebanks contain multiple written genres. Considering that the performance of the parsing models trained on one domain degrades on sentences drawn from a different domain (McClosky et al., 2010), we assume that this happens in our setting as well. Therefore, we investigate how a parser trained on multiple genres performs on the two target language genres. Is the mixture of genres in the source data robust enough to cover both of our target genres?

4 Related Work

Cross-lingual transfer learning has been useful in improving the accuracy of a low-resource target language and has been applied in a multitude of tasks (Lin et al., 2019). The process of cross-lingual transfer learning refers to resources and models from high-resource source languages to low-resource target languages on different levels.

There are four main cross-lingual parsing approaches for dependency parsing: annotation projection (Yarowsky et al., 2001; Hwa et al., 2005), model transfer (Zeman and Resnik, 2008; McDonald et al., 2011), treebank translation (Tiedemann et al., 2014; Tiedemann and Agić, 2016), and multilingual parsing models (Duong et al., 2015b; Ammar et al., 2016; Kondratyuk and Straka, 2019). The annotation projection approach requires parallel treebanks of both source language and target language, and the treebank translation approach requires a machine translation system, while in the

model transfer approach, models trained on source language treebanks are directly applied to parse target languages. A multilingual parsing model is either a model trained on one source language which is refined by taking advantage of similar structures shared with the target language, or a multilingual model using multilingual word clusters and embeddings or language-specific features.

The main challenge for cross-lingual parsing is to reduce the language discrepancies on different levels between the source language and the target language. To reduce the great differences in writing systems and vocabulary, Zeman and Resnik (2008) used delexicalization, based on the hypothesis that the interaction between morphology and syntax in two languages is similar; they applied this approach in parsing Swedish using Danish as the source language. Since this method does not require bilingual parallel data, it is extensively implemented combining with other features. McDonald et al. (2011) implemented the idea of delexicalizing the parsing models and adapting the parsers with a constraint driven learning algorithm that achieved accuracy gains. Søgaard (2011) improved the approach by Zeman and Resnik (2008) by selecting the source sentences that are most similar to the target language. Rosa and Žabokrtský (2015) trained an MSTParser model interpolation as an alternative for multi-source cross-lingual delexicalized dependency parser transfer. The work by Rosa (2015) involved the training of several independent parsers which were applied to the same input sentence. The resulting tree was obtained by finding the maximum spanning tree of a weighted directed graph of the potential parse tree edges from the different parsers.

In addition to delexicalized methods, cross-lingual lexical representations can also be used in dependency parsing. Täckström et al. (2012) used parallel data to induce cross-lingual word clusters, and added them as features for their delexicalized parser. Xiao and Guo (2014) proposed that the source and target language words with the same meaning share a common embedding. The embeddings are jointly trained with a neural model and are used for dependency parsing. Duong et al. (2015a); Ahmad et al. (2019); He et al. (2019) proposed different methods to develop multilingual word representations and used them for dependency parsing. Also, these approaches utilize zero-shot parsing since the trained parsing models parse a target language without any training target instances (Romera-Paredes and Torr, 2015). It is a suitable method for parsing low-resource languages because knowledge between different languages is transferable and labeled low-resource language data is difficult to obtain.

For Xibe, there are currently no parallel corpora or machine translation systems available, which makes model transfer the most feasible approach. In order to achieve zero-shot single-source cross-lingual parsing, we first train a parsing model on one source language treebank, then parse the target language using this model. As Xibe is written in the traditional Mongolic alphabet, which differs greatly from all the candidate source languages, we must minimize these differences. Therefore, we use treebank delexicalization by replacing lexical items with only part-of-speech tags in both the source and target languages.

5 Experimental Settings

5.1 `LangRank` and Perplexity Calculation

5.1.1 LangRank

When predicting transfer languages, `LangRank` requires four types of input: a segmented target language dataset, an unsegmented target language dataset, target language code (in our case *sjo*) and task label (*DEP*). We use the 1 131 Xibe sentences (see Section 6.2) as the unsegmented dataset, and we create the segmented dataset with `SentencePiece` (Kudo and Richardson, 2018)[1]. `SentencePiece` is a language-independent subword tokenizer and detokenizer, which creates subword models directly from raw sentences, along with tokenization. Such a subword model is required by `LangRank`. We use the following `SentencePiece` parameters: We set the final vocabulary size to 8 000 since the Xibe dataset is small. We use the default value for the other two parameters, that is, the amount of characters covered by the model is set to *1.0*, and the model type is set to *unigram*.

5.1.2 Perplexity

We compute perplexity scores based on POS bigrams. We build the bigram language models using *NLTK* (Bird, 2006) and use Laplace Smoothing to avoid zero probability for unseen bigrams, then calculate perplexity of each Xibe sentence over the

[1] https://github.com/google/sentencepiece

source language model. The final score is averaged over all Xibe sentences per source language model.

5.2 Treebanks

The training data we use come from the Universal Dependencies (UD) project, version 2.7[2]. That is, we retrieve treebanks of the candidate source languages described in Section 6. Since Turkish, Korean, and Japanese have multiple treebanks in UD, we use three Turkish treebanks: `tr_gb`, `tr_imst` and `tr_boun`, and three Japanese treebanks, `ja_modern`, `ja_bccwj` and `ja_gsd`. The perplexity score of `ko_kaist` is 22.77 which is much higher than `ko_gsd`, we therefore only use `ko_gsd` (see Table 3). As for the remaining languages, if the language has more than one treebank, we only select the largest. We use the concatenation of train/dev/test splits per source language treebank as our training data. Moreover, the treebanks of candidate source languages differ from one another in size (see Table 4 and Table 5). `bxr_bdt`, `kk_kdt`, and `ja_modern` have only around 1 000 trees, the other treebanks range between around 3 000 and almost 90 000 trees. The size discrepancy is reduced by limiting each language to at most 3 000 trees, sampled randomly where necessary.

Since we use treebanks from the Universal Dependencies project, all treebanks share the same annotation scheme. However, we are aware that there may be differences in terms of annotation quality or the interpretation of language specific characteristics. Such issues are beyond the current project, but need to be addressed in future work.

The test data comes from the Xibe treebank, which generates three test datasets based on genre:

1. `grammar`: 544 grammar examples
2. `news`: 266 news sentences
3. `mixed`: the two genres combined, 810 trees

We delexicalize all the treebanks by replacing their word forms with their POS tags.

5.3 Parser

We use `UDify` (Kondratyuk and Straka, 2019) for the parsing experiments. `UDify` is a state-of-the-art multilingual multi-task model capable of accurately predicting universal parts of speech, morphology features, lemmas, and dependency trees simultaneously. It uses the pre-trained multilingual BERT model, which allows it to handle a large number of languages with reasonable performance, without requiring any language-specific components. On top of the BERT model, the parser uses an attention layer and a multi-task learning setup so that each of the linguistic tasks, predicting part-of-speech, morphological features, lemmas, and dependencies are single tasks that are learned jointly.

To determine whether `UDify` can parse Xibe straightforwardly without removing lexical items, we parse Xibe with the pre-trained UDify model, obtaining a UAS of 24.28% and an LAS of 6.79%. These results provide a strong indication that the vocabulary differences between Xibe and other languages cannot be bridged by the multilingual BERT model. Consequently, we decided to delexicalize our data and use (gold) POS sequences instead.

We train the individual models on the delexicalized treebank of each source language and parse the Xibe texts (also delexicalized) using those models. We use the default parameters, but set the *warmup_steps* and *start_step* to 256.

5.4 Evaluation

Evaluation is performed using the Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS), as computed by the official evaluation script provided for the CoNLL 2018 shared task[3].

6 Source Language Selection Methods

In this section, we describe the three methods for determining the best source languages, and present the languages chosen by these methods.

6.1 Typology Based Selection

The first approach uses linguistic knowledge: As described in Section 2, Xibe is a Tungusic language that shares morphological and syntactic features with other transeurasian languages. Therefore, transeurasian languages are assumed to be good candidates, including those belonging to Turkic, Mongolic, Tungusic, Koreanic and Japonic language families. To ensure that the candidate source languages have at least one dependency treebank, we limit our experiments to the following languages, which are included in the most re-

[2] https://universaldependencies.org/

[3] https://universaldependencies.org/conll18/evaluation.html

Genre	Top 3 predictions	
grammar examples	Czech	ces
	Norwegian	nor
	Spanish	spa
news	Finnish	fin
	Slovenian	slv
	Korean	kor
mixed	Finnish	fin
	Slovenian	slv
	Slovak	slk

Table 1: LangRank predictions on three genres.

Feature	Top 3 languages	
geographic	Russian	rus
	Hindi	hin
	Latvian	lav
genetic	Latvian	lav
	Czech	ces
	Norwegian	nor
word overlap	Chinese	zho
	Indonesian	ind
	English	eng

Table 2: Top 3 predictions using a single feature in LangRank.

cent release of the Universal Dependency treebank collection: Buryat (Mongolic), Japanese (Japonic), Korean (Koreanic), Kazakh (Turkic), Turkish (Turkic), and Uyghur (Turkic).

6.2 LangRank Based Selection

LangRank (Lin et al., 2019) is an approach for choosing source languages for cross-lingual NLP tasks including machine translation, entity linking, part-of-speech tagging, and dependency parsing. The task of selecting the optimal source languages for an NLP task is formulated as a ranking problem. Given a low-resource target language and a set of candidate source languages, a model is trained to rank the source languages according to the performance achieved when they are used in training to process the target language. Each candidate source language is represented with a set of dataset-dependent and dataset-independent features. The dataset-dependent features include dataset size, type-token ratio, word overlap and subword overlap, and the dataset-independent features include geographic distance, genetic distance, inventory distance, syntactic distance, phonological distance, and feature distance. Based on these features, the system implements gradient boosted decision trees (GBDT; Ke et al. (2017)) to select the best transfer languages for the four NLP tasks.

In our experiment, we use the Xibe treebank sentences (544 grammar examples and 266 news sentences) for prediction. We collect 321 more sentences from news to keep the two genres balanced, since LangRank does not define how much data is needed. Note that we use sentences as input for LangRank, we do not delexicalize the data for this step.

Table 1 lists the top three predicted source languages for each genre. Czech, Norwegian and Spanish rank among the top three when we feed in grammar examples. Finnish, Slovenian and Korean are the top three predictions when only news is used as input. Additionally, Finnish and Slovenian are also top languages when mixed data is used, followed by Slovak.

Lin et al. (2019) mentioned the possibility that LangRank cannot generalize well on certain languages since it is trained only on a few languages for the particular tasks. To obtain more educated guesses for choosing the transfer language, they analyzed the learned models and extracted the most important features for given tasks. In the dependency parsing task, *geographic distance*, *genetic distance* and *word overlap* are features that yield good scores on their own. Table 2 lists the top 3 predictions when only one relevant feature is used. In Table 1 and Table 2, only Czech and Norwegian appear in both results. But the results can be explained more easily. For example, Russia and India are geographically closer to the area where Xibe is spoken, and Xibe has larger word overlap with Chinese as a result of long-term language contact.

6.3 Perplexity Based Selection

Here, we attempt to automatically approximate the typological approach by determining similarity via POS bigrams. We use perplexity as a similarity metric. Basically, we determine the optimal source languages among the languages covered by Universal Dependencies by computing the perplexity between each of the treebanks (see Section 5.2) and Xibe. As vocabularies and orthographies among languages differ greatly, we use POS bigrams instead of words to calculate perplexity. The inherent assumption is that the POS bigrams

Language	ISO	Lang.family	Treebank	PP
Buryat	bxr	Mongolic	bxr_bdt	7.93
Kazakh	kaz	Turkic	kk_ktb	8.30
Turkish	tur	Turkic	tr_imst	8.59
Uyghur	uig	Turkic	ug_udt	8.69
Turkish	tur	Turkic	tr_boun	9.22
Turkish	tur	Turkic	tr_gb	9.25
Japanese	jpn	Japonic	ja_modern	12.48
Japanese	jpn	Japonic	ja_bccwj	13.94
Japanese	jpn	Japonic	ja_gsd	14.36
Korean	kor	Koreanic	ko_gsd	13.27
Korean	kor	Koreanic	ko_kaist	22.77

Table 3: Perplexity scores for source languages.

give us a local view of Xibe syntax that will allow us to determine syntactic similarities to other languages. A language that is close to Xibe should have a low perplexity score.

We find that the perplexity scores of the Mongolic and Turkic languages are closest to Xibe (lowest perplexity) among all languages (see Table 3). Comparing these languages with the ones chosen based on typology in Section 6.1, there are considerable overlaps, except for Japanese and Korean. We also examine the perplexity scores for Korean-Xibe and Japanese-Xibe: These scores are higher than those for Turkic and Mongolic languages but lower than most of the other languages.

7 Results and Analysis

In this section, we provide the results for our three research questions.

7.1 How to Choose the Source Language?

Table 4 shows the parsing results for source languages selected by typology. Among the transeurasian languages, Kazakh achieved the highest LAS of 58.69% on grammar examples while Japanese achieved the highest LAS of 38.59% when tested on news and the highest LAS of 44.91% when tested on mixed data. On all three test datasets, Korean had the lowest LAS, with 40.54% on grammar examples, 29.16% on news and 33.41% on mixed genres. Table 5 shows the results for source languages selected by `LangRank`. Korean scored the highest LAS whereas the lowest was achieved by Spanish with 15.11% on grammar examples, 8.45% on news and 10.94% on mixed genres.

Based on Table 4, we find the most suitable source language to be Japanese. Training on the `ja_gsd` treebank results in the highest LAS for news and mixed genres, but its LAS for grammar examples is 3.19% lower than when training on Kazakh. This proves that Kazakh is more accurate than Japanese at labeling dependency relations. In terms of news and mixed genres, the gap with Japanese is actually larger, which we will investigate in section 7.3. In addition, Uyghur also performs well, its LAS on mixed genre is only 1.21% lower.

When using perplexity on POS bigrams to choose the source language, we assume that a low perplexity corresponds to a good match. However, when we compare the complexity scores in Table 3 and the parsing results in Tables 4 and 5, the situation is more complex: The Japanese treebank `ja_gsd` performs best in parsing even though it has a high perplexity score. The Korean treebank `ko_gsd` has a slightly lower perplexity than the Japanese `ja_bccwj`, but the Japanase LAS is about 11 points higher than the Korean LAS (on mixed). Similarly, Kazakh, Uyghur, and the Turkish tr_imst have similar perplexities, but the Kazakh and Uyghur LAS are about 10 points higher than the Turkish LAS (on mixed), even though the Kazakh treebank is by far the smallest. This shows that bigram POS perplexity is not an ideal measure of syntactic similarity, even though it performs better than `LangRank`.

As described in Section 6.2, standard `LangRank` may not be able to provide the best predictions. Therefore, we also investigate single features that are important for dependency parsing (see Table 2). According to the *geographic feature*, Hindi has the highest LAS 39.93% (on mixed genre, see Table 6). Similar to Xibe, Hindi has a Subject-Object-Verb (SOV) word order. Hence, we assume that the good performance of Hindi is a result of its syntactic similarity to Xibe rather than its geographic proximity. The *genetic feature* alone is not a good indicator for source language selection as all three languages achieve LAS around or below 20% (see Table 6). Languages selected via the *word-overlap feature* have poor results as well. On the mixed genre, Chinese achieves an LAS of 21.03% while Indonesian and English achieve only 13.97% and 13.22% respectively (see Table 6). Since we only used POS tags, we ignore borrowed Chinese words in Xibe sentences, and the higher performance of Chinese shows that Xibe is syntactically closer to Chinese than to English and Indonesian. Nevertheless, the LAS of Chinese is much lower than that of

Language	Treebank name	Treebank size	Training size	grammar		news		mixed	
				UAS	LAS	UAS	LAS	UAS	LAS
Buryat	bxr_bdt	927	927	65.00	43.70	44.89	32.45	52.40	36.65
Kazakh	kk_kdt	1 078	1 078	72.17	**58.69**	45.40	34.30	55.41	43.42
Turkish	tr_gb	2 880	2 880	69.22	50.39	33.08	23.91	46.58	33.81
Turkish	tr_imst	5 635	3 000	65.03	43.74	48.04	32.77	54.39	36.87
Turkish	tr_boun	9 761	3 000	66.61	47.21	51.97	37.77	57.44	41.30
Uyghur	ug_udt	3 456	3 000	69.45	52.48	54.60	38.46	60.15	43.70
Korean	ko_gsd	6 339	3 000	54.35	40.54	41.26	29.16	46.15	33.41
Japanese	ja_modern	822	822	69.94	52.42	51.82	38.95	58.60	43.98
Japanese	ja_bccwj	57 028	3 000	73.34	55.41	53.68	37.26	61.03	44.04
Japanese	ja_gsd	8 071	3 000	**73.68**	55.50	**55.01**	**38.59**	**61.99**	**44.91**

Table 4: Parsing results with typologically related languages as source languages, based on perplexity.

Language	Treebank name	Treebank size	Training size	grammar		news		mixed	
				UAS	LAS	UAS	LAS	UAS	LAS
Czech	cs_pdt	87 913	3 000	31.74	19.68	18.13	10.54	23.00	13.95
Norwegian	no_bokmaal	20 044	3 000	33.35	21.87	21.02	14.48	25.63	17.25
Spanish	es_ancora	17 680	3 000	23.03	15.11	13.87	8.45	17.30	10.94
Finnish	fi_ftb	18 723	3 000	52.72	37.47	34.32	26.02	41.20	30.30
Slovenian	sl_ssj	8 000	3 000	32.64	19.66	18.68	10.15	23.89	13.71
Korean	ko_gsd	6 339	3 000	**54.35**	**40.54**	**41.26**	**29.16**	**46.15**	**33.41**
Slovak	sk_snk	10 604	3 000	30.19	19.94	14.48	8.74	20.35	12.89

Table 5: Parsing results for languages chosen by `LangRank`.

any transeurasian language in Table 4, even lower than Korean by 12.38 points.

7.2 Syntactic Similarity vs. Data Size

In the previous section, we have found Japanese to be the optimal source language for Xibe, followed by Uyghur and Kazakh. However, the Kazakh treebank only contains 1 078 trees while the Japanese `ja_gsd` and Uyghur models are trained with 3 000 trees. We investigate whether the training set size is the main factor in reaching good parsing accuracy. Consequently, we sample 1 000 trees from the Japanese `ja_gsd` treebank, making the training set size comparable to Kazakh `kk_kdt`. Parsing results are displayed in Table 7. On all three test datasets, when training with 1 000 trees, the LAS slightly decreases compared to 3 000 trees. Despite this, both LAS and UAS are still higher for the 1 000 Japenese trees than for Kazakh, with the exception of the LAS on the grammar examples. This shows clearly that the training set size is contributing only minimally.

As the Japanese results increase slightly when increasing training data from 1 000 to 3 000 trees, an obvious question is whether we can improve results by increasing the training set size further. Thus, we train parsing models by sampling 6 000 trees from `ja_gsd` and using all 8 071 trees respectively (see Table 7). However, we only see a minimal increase in LAS (45.03% vs. 44.91%) and a small decrease in UAS (on mixed). Thus we can conclude that larger training data do not necessarily lead to an improvement in performance.

We also had a closer look at Japanese and Korean, which share many linguistic features, despite which Japanese performs better than Korean. On mixed data, Korean obtains an LAS of 33.41%. One possible reason for such a large gap can be found in the differences in annotations between the two languages. As described by Han et al. (2020), in the UD Korean treebanks, a sentence is segmented into *eojeols*. An *eojeol* can consist of lexical morphemes and functional morphemes, which means the functional morpheme is agglutinated to the lexical item preceding it. In contrast, the Japanese treebank adopts the Short Unit Word (SUW). This means that functional morphemes are annotated as separate units in the Japanese treebank, and their dependency relations are present. In Xibe, function words are written as

Language	Treebank name	Treebank size	Training size	grammar UAS	grammar LAS	news UAS	news LAS	mixed UAS	mixed LAS
geographic feature									
Russian	ru_syntagrus	61 889	3 000	28.14	17.84	18.79	9.42	22.28	12.56
Hindi	hi_hdtb	16 647	3 000	67.99	50.65	50.80	33.53	57.22	39.93
Latvian	lv_lvtb	13 643	3 000	37.78	23.68	26.67	18.36	30.82	20.35
genetic feature									
Latvian	lv_lvtb	13 643	3 000	37.78	23.68	26.67	18.36	30.82	20.35
Czech	cs_pdt	87 913	3 000	31.74	19.68	18.13	10.54	23.00	13.95
Norwegian	no_bokmaal	20 044	3 000	33.35	21.87	21.02	14.48	25.63	17.25
word-overlap feature									
Chinese	zh_gsdsimp	4 997	3 000	47.16	25.67	33.80	18.26	38.80	21.03
Indonesian	in_gsd	5 593	3 000	23.40	16.28	16.39	12.60	19.01	13.97
English	en_ewt	16 662	3 000	28.51	18.45	19.03	10.10	22.58	13.22

Table 6: Parsing results for source languages chosen using a single feature in `LangRank`.

Language	Treebank name	Treebank size	Training size	grammar UAS	grammar LAS	news UAS	news LAS	mixed UAS	mixed LAS
Japanese	ja_gsd	8071	1000	72.57	54.79	53.27	37.74	60.48	44.11
			3000	**73.68**	55.50	**55.01**	38.59	**61.99**	44.91
			6000	73.16	55.62	54.43	38.38	61.43	44.82
			8071	73.20	**55.74**	54.33	**38.64**	61.39	**45.03**

Table 7: Parsing results with sampling different amounts of data from `ja_gsd`

separate words in most cases and overtly annotated, which is more similar to the Japanese treebank.

7.3 Does Genre Matter?

As shown in Section 5.2, the Xibe treebank consists of two different genres (grammar and news) while most of the source treebanks have multiple genres. This design allows us to see how different genres influence parsing results. One prominent difference between Xibe grammar examples and news is that news sentences are much longer and use more complex syntactic structure. Thus, we expect to reach higher accuracy on the grammar examples. This is born out by the results in Tables 4 and 5: Among transeurasian languages in Table 4, Turkish `tr_gb` has the largest LAS difference between grammar examples and news by 26.48% whereas Uyghur has the smallest by 9.44%. In Table 5, Finnish displays the largest LAS discrepancy between the two genres by 11.45% whereas Spanish has the smallest difference by 6.66%. We find a general tendency that results on `grammar` are considerably higher than those on `news`, with sizable differences.

We now have a closer look at the three Turkish treebanks since `tr_gb` mainly contains grammar examples while the other two contain news and non-fictional data. Comparing performance of the models trained on the three treebanks, when we test with grammar examples, `tr_gb` outperforms the other two even though it is smaller in size. When testing on news, `tr_boun` and `tr_imst` reach similar results: `tr_boun` reaches an LAS of 37.77% and `tr_imst` reaches an LAS of 32.77%. However, `tr_gb` declines by 13.86% in LAS compared to `tr_boun`. The results indicate that genre does influence parsing. When the training data contains mainly simpler syntactic structures than the test data, the parser cannot analyze the more complex test data adequately.

8 Conclusion

In this research, we have investigated cross-lingual dependency parsing for Xibe. As we do not have parallel data or a machine translation system for this language, we delexicalize treebanks to avoid orthographic and lexical differences. We propose three criteria to select source languages, that is, typology, perplexity, and automatic predictions by the `LangRank` tool. Then, we train parsing models with `UDify` and test them with the three sets of

Xibe data. Our results demonstrate that syntactic similarity is considered the most important factor in delexicalized cross-lingual parsing. Japanese is found to be the optimal source language for parsing Xibe. Differences in genre also influence parsing. Parsers trained on simpler sentence structures cannot analyze more complex test data.

In our current work, we use only one source language to parse Xibe. We will determine the best concatenation of source languages for multilingual parsing in the future. Additionally, we will resegment the current units of the Korean treebanks into smaller units and create dependency relations by rules in order to determine if a more similar segmentation will lead to an improvement. Alternatively, we can use lexical information in parsing, such as creating a bilingual dictionary or training Xibe word embeddings.

Acknowledgments

We would thank the three anonymous reviewers for their helpful comments. He Zhou is supported by China Scholarship Council.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2440–2452, Minneapolis, MN.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 214–217, Syndey, Australia.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China. Association for Computational Linguistics.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, pages 99–108, Barcelona, Spain (Online).

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223, Florence, Italy.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–326.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *31st Conference on Neural Information Processing Systems (NIPS)*, pages 3146–3154, Long Beach, CA.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing:*, pages 66–71, Brussels, Belgium.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, CA.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, UK.

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120, Paris, France.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France.

Martine Robbeets and Alexander Savelyev. 2020. *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161.

Rudolf Rosa. 2015. Multi-source cross-lingual delexicalized parser transfer: Prague or Stanford? In *Proceedings of the Third International Conference on Dependency Linguistics (Depling)*, pages 281–290, Uppsala, Sweden.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. MST-Parser model interpolation for multi-source delexicalized transfer. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 71–75, Bilbao, Spain.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada.

Jörg Tiedemann and Zeljko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, MI.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, MI.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*.

Daniel et al. Zeman. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

He Zhou, Juyeon Chung, Sandra Kübler, and Francis Tyers. 2020. Universal dependency treebank for xibe. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 205–215.