Natural Language Processing (2025), **31**, pp. 277–305
doi:10.1017/nlp.2024.21

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Cross-lingual dependency parsing for a language with a unique script

He Zhou, Daniel Dakota and Sandra Kübler

Department of Linguistics, Indiana University, Bloomington, IN, USA
Corresponding author: He Zhou; Email: hzh1@iu.edu

(Received 23 December 2022; revised 27 July 2023; accepted 16 September 2023; first published online 9 September 2024)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

Syntactic parsing is one of the areas in Natural Language Processing. The development of large-scale multilingual language models has enabled cross-lingual parsing approaches, which allows us to develop parsers for languages that do not have treebanks available. However, these approaches rely on the assumption that languages share orthographic representations and lexical entries. In this article, we investigate methods for developing a dependency parser for Xibe, a low-resource language that is written in a unique script. We first investigate lexicalized monolingual dependency parsing experiments to examine the effectiveness of word, part-of-speech, and character embeddings as well as pre-trained language models. Results show that character embeddings can significantly improve performance, while pre-trained language models decrease performance since they do not recognize the Xibe script. We also train delexicalized monolingual models, which yield competitive results to the best lexicalized model. Since the monolingual models are trained on a very small training set, we also investigate lexicalized and delexicalized cross-lingual models. We use six closely related languages as source language, which cover a wide range of scripts. In this setting, the delexicalized models achieve higher performance than lexicalized models. A final experiment shows that we can increase performance of the cross-lingual model by combining source languages and selecting the most similar sentences to Xibe as training set. However, all cross-lingual parsing results are still considerably lower than the monolingual model. We attribute the low performance of cross-lingual methods to syntactic and annotation differences as well as to the impoverished input of Universal Dependency Part-of-Speech tags that the delexicalized model has access to.

Keywords: parsing; Xibe; multilinguality

1. Introduction

Syntactic parsing is one of the areas in Natural Language Processing that has immensely profited from deep learning, especially by the development in large scale pre-trained language models. One of the most difficult problems in the pre-neural era was data sparsity. Manually annotated treebanks are expensive and tend to be small, so that unknown words and unknown constructions were equally likely. Neural models have proven to be better suited for making correct parsing decisions, and embeddings working on subword levels reduce the problems of handling unknown words (Vania, Grivas, and Lopez 2018).

Another achievement resulting from the use of neural models is cross-lingual parsing, where we train on a source language and then parse a different target language, with a multilingual language model, such as multilingual Bidirectional Encoder Representations from Transformers (mBERT)

(Devlin *et al.* 2019), serving as a bridge (Das and Sarkar 2020). This is a solution for many low-resource languages for which no treebanks exist. However, while cross-lingual transfer learning has shown benefits in cross-lingual parsing (Das and Sarkar 2020), there are certain underlying assumptions that are often made in the process. One such assumption is that we can successfully leverage the orthographic representations of languages (henceforth: scripts) to help facilitate information sharing. This is particularly pertinent when sharing lexical information, such as word or character embeddings, or embeddings derived from pre-trained multilingual language models, as although transfer has improved performance, the success is still influenced by the data used to create the embedding (Joshi, Peters, and Hopkins 2018; Dakota 2021). Additionally, how the scripts of languages are handled by the tokenizer of the language model becomes central to their ability to deal with a large number of word representations to which they are exposed during training and testing. A central question then arises in how effective such lexicalized approaches are when the language is not only low-resource but it is written in a unique script that will not benefit from lexical sharing. And what additional strategies can be employed in such cross-lingual settings?

To examine these questions, we investigate how to effectively develop dependency parsing strategies for Xibe, which is not only a low-resource language, but written in a unique script that is not present in pre-trained language models, nor fully covered in any related languages. We investigate lexicalized (Kübler *et al.* 2009) and delexicalized parsing (Zeman and Resnik 2008). The latter approach makes cross-lingual parsing feasible in settings where large multilingual language models are not available or insufficient.

In a first step, we investigate monolingual parsing before performing cross-lingual parsing experiments, using both lexicalized and delexicalized approaches to determine the difficulties when applying standard cross-lingual approaches. We find that word plus character embeddings yield a high performance in monolingual parsing, which often deteriorate when language models are included, due to their inability to handle the Xibe script. Our best results are obtained with a lexicalized model using word and character embeddings plus the Pre-trained Language Model for Chinese Minority Languages (CINO; Yang *et al.* 2022). But we reach competitive results using delexicalized monolingual approaches, as cross-lingual approaches either suffer from an inability to transfer knowledge via script incompatibilities or from conflicting syntactic constructions between various source languages and Xibe.

In summary, the major contributions of our work presented here are the following: (1) We investigate parsing Xibe, a language with a unique script. To our knowledge, the only relevant work on that language is our work on delexicalized parsing of Xibe (Zhou and Kübler, 2021). Our results on suitable source languages builds the basis for our multilingual experiments here. (2) We show that monolingual lexicalized parsing is successful when trained on a small set of sentences, in combination with word and character embeddings. (3) We show that pre-trained multilingual language models do not improve performance in such a setting. (4) We show that monolingual delexicalized parsing is a competitive alternative. (5) We show that multilingual parsing results in a lower performance.

We first cover relevant cross-lingual parsing literature in Section 2 before introducing the Xibe language and treebank in Section 3. We present our methodology in Section 4 and initial monolingual parsing results in Section 5 in which we also provide discussions on how Xibe is handled by various embedding representations and on the delexicalized performance. Cross-lingual approaches are presented in Section 6, covering both single source and multi-source approaches, with a particular focus on how differences between languages and annotations in treebanks can impact performance. We conclude in Section 7.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

2. Related work

Since there is a wide range of literature on parsing in general, and on cross-lingual parsing more specifically, we restrict our overview to cross-lingual dependency parsing, with a specific focus on the different lexicalized and delexicalied approaches.

Cross-lingual transfer learning has been useful in improving the accuracy of low-resource target languages and has been applied in a multitude of tasks (Lin *et al.* 2019). This process refers to the application (and potentially adaptation) of resources and models from high-resource source languages to low-resource target languages on different levels with the assumption that existing similarities between languages can be exploited.

The main challenge for cross-lingual parsing is to reduce the language discrepancies on different levels (such as in the annotation, or on the lexical level, wrt. scripts, etc.), between the source language and the target language. For dependency parsing, there are four main cross-lingual parsing approaches: annotation projection (Yarowsky, Ngai, and Wicentowski 2001; Hwa *et al.* 2005), treebank translation (Tiedemann, Agić, and Nivre 2014; Tiedemann and Agić 2016), multilingual parsing models (Duong *et al.* 2015b; Ammar *et al.* 2016; Kondratyuk and Straka 2019), and model transfer (Zeman and Resnik 2008; McDonald, Petrov, and Hall 2011).

Annotation projection uses word alignment to project annotations from the source to the target language; treebank translation refers to the automatic translation of treebank data, while in model transfer approaches, models trained on source language treebanks are directly applied to parse target languages. For Xibe, there are currently no parallel corpora or machine translation systems available, which makes model transfer the most feasible approach.

2.1 Delexicalized approaches

A common approach to cross-lingual parsing is delexicalization. The strategy was first employed by Zeman and Resnik (2008) in constituency parsing, who replaced words with their morphological (part-of-speech (POS)) representations when adapting a Danish treebank to Swedish. Delexicalization results in similar input to the parser, independent of the lexical differences between the languages.[a] This approach produced a model that matched the performance of a model trained on approximately 1,500 sentences from the target language. The approach sparked several concurrent research directions. McDonald *et al.* (2011) concatenated multiple delexicalized source treebanks when training a dependency parser for a target language. Results were noticeably better than unsupervised parsing approaches. Simultaneous work by Cohen *et al.* (2011) used delexicalized models to support initial parameters for unsupervised parsing when using treebanks with non-parallel data. Rosa and Žabokrtský (2015b) trained an MSTParser model interpolation as an alternative for multi-source cross-lingual delexicalized dependency parser transfer, finding that performance was comparable to parse tree combinations, but computationally less expensive. 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 and yielded better trees than those generated from a single model, with the largest increases shown in lower resource settings.

One persistent issue in cross-lingual modeling is optimal data source selection for specific target data. Using perplexity for data point selection in a cross-lingual setup was shown by Søgaard (2011) as more suitable for selecting only similar sentences from the source treebank to parse the target language, in comparison to utilizing the entire source data. Rosa and Žabokrtský (2015a) used KL divergence (Kullback and Leibler, 1951) of POS trigrams to detect language similarity

[a]We note that a great deal of additional research exists for delexicalized constituency parsing, but we focus here primarily on dependency parsing.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

for source and target selection in delexicalized dependency parsing. Results showed good performance for single-source transfer and were further strengthened in a multi-source setup, but results seemingly degrade the further apart two languages are linguistically, suggesting that it may require additional language characteristics. In a similar vein, approaches by Dozat *et al.* (2017) used delexicalized language family models to parse additional surprise languages within the target family, assuming that language families share closer POS distributions and syntactic characteristics. Parallel work by Das *et al.* (2017) used transformations of syntactic features of the source languages to create delexicalized models to improve source-target selection criteria when parsing an unknown target language.

However, linguistic relatedness is not a guarantee of optimal source selection. Lynn *et al.* (2014) found that Indonesian was a better source language than the more closely related Indo-European languages when parsing Celtic in a cross-lingual setting. They noted that specific linguistic/annotation phenomena in treebanks, such as nominal modifications and long range dependencies, can be more closely aligned in a specific language/treebank pair, the impact of which will correspond to the prevalence of such phenomena in the target language.

Little delexicalized work exists for parsing Xibe. Zhou and Kübler (2021) performed cross-lingual experiments using three selection criteria: typology, perplexity, and LangRank as source selection, noting that syntactic similarity was the most important factor in source selection, with Japanese being the optimal source language.

2.2 Lexicalized approaches

While delexicalized parsing seems to be a natural choice in a cross-lingual experiments due to lexical differences in source and target languages, this also means that the parser has to rely on a very coarse-grained POS tagset, the Universal Dependencies (UD) POS tagset, to model syntactic phenomena. Consequently, lexicalized cross-lingual parsing has in some cases yielded superior results (Falenska and Çetinoğlu, 2017). Early work by Täckström *et al.* (2012) used parallel data to induce cross-lingual word clusters, and added them as features for their delexicalized parser. Results showed a reduced relative error across treebanks by up to 13 percent and importantly, such features never had a negative impact on model performance. 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. Results showed superior cross-lingual generalizability across nine languages.

A hybrid representation was proposed by Duong *et al.* (2015a) where cross-lingual embeddings were generated by using lexical and POS representations, generating more syntactically aware embeddings. Results were higher than for corresponding delexicalized baselines. Ahmad *et al.* (2019) explored encoder representations pairing English as source language against a typologically diverse set of thirty target languages. While training was performed only on English, at testing times embeddings derived from pre-trained multilingual models were used as input and projected into a shared embedding space and aligned with the English embeddings. They found that RNN encoders were better for target languages that were syntactically closer to English while transformers are more flexible with word order modeling. He *et al.* (2019) focused on cross-lingual approaches to more distant languages by using URIEL (Littell *et al.* 2017), which encodes information such as typological, geographical, and phylogenetic features of languages, into vector representations to determine language distances. They utilized a structured flow model to induce interlingual embeddings to enhance the ability to learn and share syntactic information on a target language. Results using English as the source yielded improvements on a set of ten distant languages.

While pre-trained language models such as BERT (Devlin *et al.* 2019) have resulted in substantial performance gains in cross-lingual experiments with the use of pre-trained multilingual embeddings (e.g., mBERT), they still possess some limitations. Since those models split words into subwords, subword overlap between source and target language can be a potential indicator of source language selection (Wu and Dredze 2019). However, such models do not necessarily yield

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

optimal results for very low-resource languages. Wu and Dredze (2020) found that subword pieces of low-resource languages may be present in the base vocabulary of mBERT, but this does not mean these represent the most relevant subword types in the target language, which may be missing all together. They note that one of the main contributing factors to the limitation of mBERT is simple limitations in data availability and quality for low-resource languages and caution against relying solely on such paradigms in low-resource settings. Work by Rust *et al.* (2021) examining the tokenizers in monolingual and multilingual models shows the impact that monolingual tokenizers have on performance. When using a multilingual model, simply replacing the multilingual tokenizer with a monolingual tokenizer for the target language improves performance on multiple tasks and languages. Recent work by Blaschke *et al.* (2023) examined how subword overlap aids in source and target selection for cross-dialect POS tagging. They use metrics including the split word ratio difference (the ratio of words split into subwords between the source and target), the overlap of seen subwords in the source and target dialects, and the subword level type-token ratio. Experiments on a range of dialects across multiple languages indicate that similarity in the split word ratio is the stronger indicator of source and target performance.

3. The Xibe language and treebank

3.1 Xibe: A severely endangered Tungusic language in China

The Xibe language (ISO 639-3: sjo) is a Tungusic language used by the Xibe ethnic group in China. According to the Seventh National Population Census conducted in 2020 (Office for the National Population Census, 2022), Xibe has a population of around 190,000 members, which are mainly distributed in the northeastern provinces (Liaoning, Jilin, and Heilongjiang) and Xinjiang Uyghur Autonomous Region. For historical reasons, the Northeast Xibe has almost lost their language, whereas the Xinjiang Xibe still actively uses their native language.

The modern Xibe language is mainly spoken in the Cabcal Xibe Autonomous County and its adjacent areas including Huocheng, Tacheng, Gulja, and Urumqi (Gorelova, 2002). However, the actual number of speakers of the language ranges between 10,000 and 50,000 (Chog, 2016a); it is in decline as the social function of the language is weakening. UNESCO recognizes Xibe as a severely endangered language, that is, the language is spoken by grandparents and older generations, the parent generation may understand it but they do not speak it to their children or among themselves. Therefore, documenting this language is necessary from perspectives of linguistic research and preservation of culture.

Xibe shows significant differences between its spoken and written forms. Spoken Xibe is a collection of dialects, and there is no standard. Previous linguistic research on the Xibe language mostly focused on documenting Xibe dialectal differences or studying phonology and morphology of spoken Xibe (Norman 1974; Li 1979, 1982, 1984, 1985, 1988; Jang 2008; Zikmundová 2013). Written Xibe is written in Xibe script, and it is used less often, compared to spoken Xibe as the younger generation mostly cannot read or write the script (Chog, 2016b). Previous studies related to written Xibe are mainly concerned with comparisons to spoken Xibe or literary Manchu, a closely related language (Gu, 2016). In this study, we focus on parsing written Xibe.

3.2 Writing system

The modern Xibe writing system is slightly modified from the Manchu script *tongki fuka sindaha hergen* (Eng.: letters with circle and dot), which is derived from the traditional Mongolian writing system. The Xibe script is written vertically from top to bottom.[b] Xibe has five vowels and

[b]For Xibe script to be properly displayed here, we write in-text examples horizontally.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 1. Xibe alphabet of vowels and consonants, with Latin transliterations

Vowels (5)	a	e	i	o	u						
Consonants (19)	n	k	g	h	b	p	s	š	t	d	
	l	m	c	j	y	r	f	w	ng		
Foreign letters (10)	ck	cg	ch	z	dz	tsy	sy	c'y	j'y		

Table 2. Unified Mongolian graphemes and graphemes specific to the Xibe writing system

Mongolian	a	o	n	b	m	l	
	s	c	y	r	w	ck	
Xibe	e	i	u	k	g	h	p
	š	t	d	j	f	ng	cg
	ch	z	ts	dz	iy	c'	j'

nine consonants, plus ten "foreign" letters, which are specifically used for loanword transliteration (Šetuken, 2009), shown in Table 1. In addition, each Xibe letter has three shapes depending on its position in a word: at initial, middle or final positions. Vowels additionally have an isolated shape as they can be used alone.

According to the Unicode Standard Version 15.0 (The Unicode Consortium 2022), Xibe graphemes are encoded partly using traditional Mongolian graphemes and partly using graphemes specific to Xibe. Xibe shares twelve graphemes with traditional Mongolian and uses twenty-one Xibe specific graphemes; see Table 2.

3.3 The Xibe Universal Dependencies Treebank

Zhou *et al.* (2020a) annotated a small dependency treebank based on the Universal Dependencies framework (de Marneffe *et al.* 2021). The treebank currently contains a total of 1,200 sentences, including 544 grammar examples collected from a written Xibe grammar book (Šetuken 2009), 266 sentences from *Cabcal Newspaper* and 390 sentences from the introductory Xibe textbooks *Nimangga Gisun* (Eng.: Mother Tongue).[c]

In the treebank, annotation for each tree includes sentence-level and token-level information. Sentence information includes the Xibe sentence, Latinized transliteration and English translation. Each token is annotated for lexical, morphological and syntactic information. Figure 1 shows an example annotation for a simple sentence. *meiherehebi* (Eng.: have carried) is the ROOT, in sentence final position. The sentence has a core argument and an adjunct: *suduri i ujen tašan* (Eng.: serious duty of history) is the object marked by the accusative case marker *be*, and *musei meiren* is the adjunct marked by the locative case marker *de*. Xibe is a pro-drop language; in this sentence, the subject is dropped and consequently does not occur in the annotation. Within the object phrase, there is a nominal modifier marked by the genitive case *i*.

[c]The current release of the treebank in version 2.10 of the Universal Dependencies project (https://universaldependencies.org/) contains 810 sentences; the 390 textbook sentences will be released in the next UD release.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

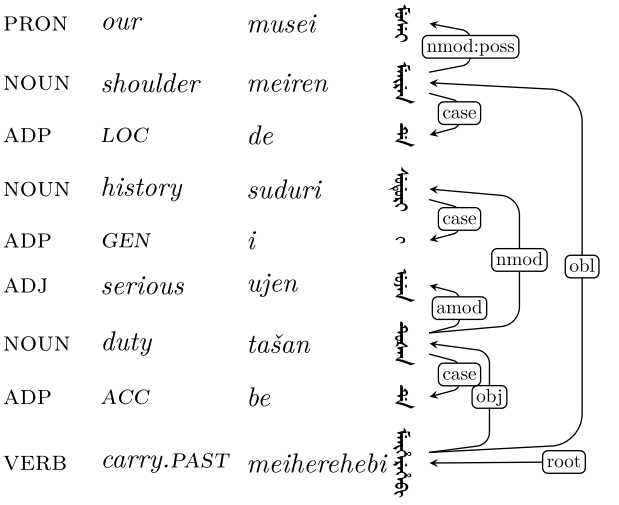


Figure 1. Example of a dependency tree; Eng.: "(We) have carried serious duty of history on our shoulder."

3.4 Xibe syntax

As a Tungusic language, Xibe shares a range of morphological and syntactic traits with other languages from the same language family: All have agglutinative morphology and use Subject-Object-Verb (SOV) word order. However, the languages differ in the degree of agglutination. In comparison to other Tungusic languages such as Evenki and Nanai, Xibe (as well as Manchu and Jurchen) has less inflectional morphology, which is assumed to be the result of language contact with Sinitic languages (Whaley and Oskolskaya 2020).

Xibe sentences follow a rigid SOV word order. Arguments are marked for case, but case marking is optional (see Section 3.4.1 for more details). Phrases are consistently head final, and subordinate clauses are located before the head they modify. The verbs of subordinate clauses are in non-finite form. Adverbial clauses are headed by converbs (see Section 3.4.2), and adnominal clauses are headed by participles. In written Xibe, inflectional morphology is mainly present on nouns and verbs. Nouns inflect for number and case. Verbal morphology of Xibe is the most complex. A verb in Xibe consists of a verb stem and an inflectional suffix. Inflectional suffixes include finite and infinite verb suffixes. Additional verbal suffixes express tense, aspect, voice, and mood.

For the remainder of this section, we focus on the two syntactic phenomena that are relevant for understanding our parsing results, namely the case marking system and converbs.

3.4.1 Case marking

Written Xibe has eight cases, as shown in Table 3. Subjects are unmarked. There are two cases of syncretism: ᠢ *i* marks genitive or instrumental case and ᠴᡳ *ci* marks ablative or lative case. The case markers follow their nouns, they can be written either as suffixes attaching to the nouns or as independent tokens. Following the UD guidelines (de Marneffe *et al*. 2021), separate case markers depend on their preceding nouns. If the case marker is suffixed on the noun, it is not annotated separately.

The example tree in Figure 1 has three separately written case markers. The sentence in Figure 2 has an accusative and a lative case marker, but the lative case is suffixed to ᠪᠣᠣ (*boo*, Eng.: home), which depends on the verb as an oblique.

Note that the case markers are clear indicators of the grammatical function of the noun, and the separate case markers are accessible for the parser in lexicalized parsing settings. Whether the suffixes are also accessible may depend on the amount of training data. However, in delexicalized parsing, none of this information is available.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 3. Case markers in written Xibe

Case	Meaning	Form
NOM	Nominative	Ø
ACC	Accusative	be
GEN	Genitive	i
INS	Instrumental	i
DAT	Dative	de
ABL	Ablative	deri, ci
LAT	Lative	ci

Table 4. Converb suffixes in written Xibe

Meaning	Suffix
Conditional	ci
Concessive	cibe
Imperfect	me
Perfect	fi, pi, mpi
Durative	hai, hei, hoi
Terminative	tala, tele, tolo
Instrumental	tai, tei, toi
Preparative	nggala, nggele, nggolo

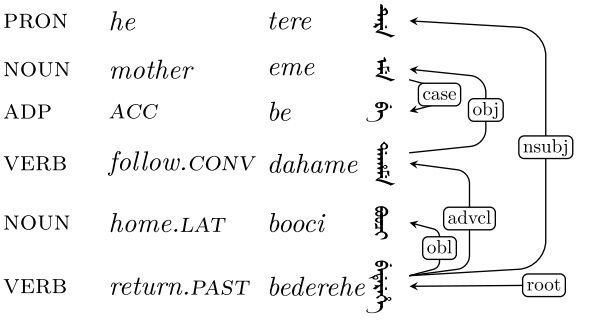


Figure 2. Example dependency tree; Eng.: "He returned home following (his) mother."

3.4.2 Converbs

Infinite verb forms in Xibe consist of converbs and participles. Converbs are nonfinite verbs that express adverbial subordination. This syntactic construction is typical for Tungusic, Turkic, and Mongolic languages. Xibe uses a range of converb suffixes, denoting different aspects or modal meanings, as shown in Table 4. In example (1), the suffix *-hai* in (*hadahai*) is the durative suffix, which adds the meaning that an action continues at the same time as another action.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

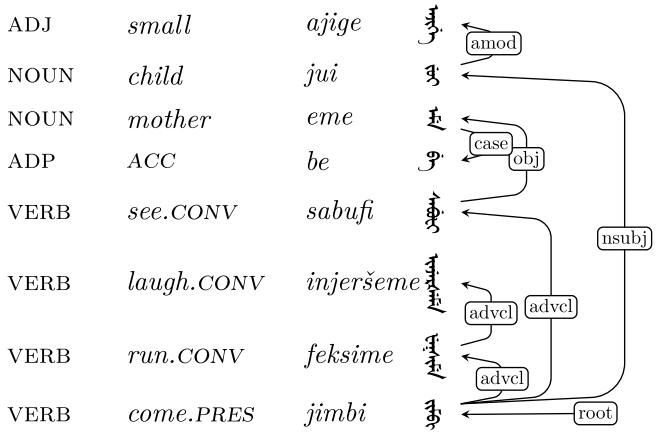


Figure 3. Dependency annotation of the converb construction from example 2.

(1)
yasa hadahai tuwambi
eye stare.CONV look.PRES

"The eye keeps staring."

In the Xibe UD treebank, converbs are dependents of their matrix verbs, via an adverbial clause (advcl) relation. This relation accounts for 9.04 percent of all dependencies in the Xibe treebank.

(2)
ajige jui eme be sabufi injeršeme feksime jimbi
small child mother ACC see.CONV laugh.CONV run.CONV come.PRES

"Having seen (his) mother, the small kid laughs and runs over."

Example (2) has a perfect converb (*sabufi*) and two imperfect converbs (*injeršeme*) and (*feksime*). In the annotation shown in Figure 3, the perfect converb (*sabufi*) depends on the main verb of the sentence converb (*jimbi*), to describe that the action of "seeing" is completed before the action "to come" is undertaken. The two imperfect converbs have the suffix converb *-me*, which denotes simultaneity of actions. In the example sentence, converb (*injeršeme*) depends on the other imperfect converb (*feksime*) as an accompanying action. converb (*feksime*) depends on the main verb as an accompanying action.

It is clear that in a cross-lingual setting, this construction can only be successfully parsed when the source language also has similar converb constructions.

4. Experimental setup

4.1 Treebank data

We use Xibe as our target language. The current Xibe treebank is relatively small, with approximately 1,200 sentences, and only provides a test set (see Table 5) according to UD guidelines.

For our source languages, we select languages based on language typology. Xibe is a Tungusic language. The Tungusic languages together with Mongolic and Turkic languages belong to the Altaic language family (Robbeets 2020). When we consider the broader category of the

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 5. Number of sentences and tokens of the selected UD treebanks

				Train		Dev		Test	
Family	Language	Treebank	Script	Sent	Token	Sent	Token	Sent	Token
Mongolic	Buryat	bxr_bdt	Cyrillic	19	153			908	10,032
	Kazakh	kk_ktb	Cyrillic	31	547			1047	10,142
	Uyghur	ug_udt	Arabic	1,656	19,262	900	10,644	900	10,330
Turkic	Turkish	tr_boun	Latin	7,803	99,165	979	12,174	979	12,207
	Turkish	tr_kenet	Latin	15,398	143,287	1,646	17,554	1643	17,817
	Turkish	tr_penn	Latin	14,850	165,513	622	6,994	924	10,047
Koreanic	Korean	ko_gsd	Hangul	4,400	56,687	950	11,958	989	11,677
	Korean	ko_kaist	Hangul	23,010	296,446	2,066	25,278	2,287	28,366
Japonic	Japanese	ja_gsd	Kanji/Kana	7,050	168,333	507	12,287	543	13,034
	Japanese	ja_gsdluw	Kanji/Kana	7,050	130,298	507	9,531	543	10,429
Tungusic	Xibe	sjo_xdt	Xibe					1,200	23,463

Transeurasian language hypothesis (Robbeets 2020), Korean and Japanese are also considered to be close relatives of Xibe.[d] We have decided to use both the Altaic and Transeurasian language neighbors since these language families exhibit similar linguistic characteristics, especially on the morpho-syntactic and syntactic level. Such shared characteristics are often important criteria for determining optimal source languages for cross-lingual parsing. Additionally, these characteristics have already been shown to also be of high relevance for Xibe in cross-lingual settings (Zhou and Kübler 2021).

The similarities in morphology and syntax between Xibe and the other Transeurasian languages (Robbeets and Savelyev, 2020) mainly include the following:

1. They have a predominant SOV word order, and the predicate verb is strictly located in a sentence final position.
2. Phrases follow a head-final word order.
3. Transeurasian languages all have a high degree of morphological agglutination, where bound morphemes are predominantly suffixing.

In UD version 2.10, there are twenty-five treebanks representing nine Transeurasian languages. Following Zhou and Kübler (2021), we select ten treebanks in six languages as shown in Table 5 and exclude three languages, Old Turkic, Tatar, and Yakut, due to their limited number of available trees (less than one hundred).

4.2 Data split

For most treebanks, we use the training, development, and test splits provided by the UD treebanks. For Xibe, Buryat, and Kazakh, however, only small treebanks are available and few non-test sentences are available. For these languages, we choose to split the data into three folds per treebank.

[d]While the Altaic hypothesis and the Transeurasian hypothesis are still under debate (Rybatzki 2020), the debate focuses primarily on whether these languages are descended from a common ancestral language, Proto-Altaic. Both sides of the debate agree that these languages have many linguistic similarities, despite the fact that the languages are geographically scattered throughout the vast Eurasian continent.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

For monolingual Xibe experiments (Section 5.1), this means we run threefold experiments where one fold serves as training, development, and test set, respectively. For all experiments where Xibe is used for testing, we use the full treebank as test set.

For the single-source cross-lingual experiments (Section 6.1), since Buryat and Kazakh are only used for training, we combine two folds into a training set while using the third as development set.

For multi-source cross-lingual experiments (Section 6.3), we concatenate training, development and test data of all selected treebanks. Then we use 80 percent for training and the remaining 20 percent as development set. For the training/development split, we use perplexity-based stratified sampling to ensure that both sets have the same distribution in terms of perplexity.

4.3 Multilingual word embeddings

For pre-trained multilingual language models, we use mBERT (multilingual Bidirectional Encoder Representations from Transformers; Devlin *et al.* 2019), cross-lingual RoBERTa (Robustly Optimized BERT Approach; XLM-R; Conneau *et al.* 2020), and the Pre-trained Language Model for CINO (Yang *et al.*, 2022). The three pre-trained language models are trained on corpora of different sets of languages and use different tokenizers.

mBERT is a large-scale pre-trained language model trained on 104 languages, which include typological relatives of Xibe: Kazakh, Turkish, Tajik, Kyrgyz, Tatar, Uzbek, Korean, and Japanese. It uses the WordPiece tokenizer (Wu *et al.* 2016), which breaks words into subword pieces.

XLM-R is a multilingual model that is trained on CommonCrawl data containing one hundred languages, including Kazakh, Kyrgyz, Turkish, Uyghur, Uzbek, Mongolian (Cyrillic), Korean, and Japanese. XLM-R uses the SentencePiece tokenizer (Kudo and Richardson 2018), which encodes a sentence as a Unicode sequence and decodes the tokenized Unicode sequence back to text.

CINO is built on XLM-R but has been adapted for minority languages by resizing its vocabulary and adopting a fast masked language modeling objective for pre-training. The extended languages include Cantonese, Tibetan, traditional Mongolian, Uyghur, Kazakh in Arabic script, Latinized Zhuang, and Korean. The CINO tokenizer also extends the XLM-R tokenizer via merging a Tibetan tokenizer and a traditional Mongolian tokenizer so that it can recognize these two languages (Yang *et al.* 2022).

All three language models are trained on more than one hundred languages and include syntactically similar languages, such as Turkic languages, Korean and Japanese, but to the best of our knowledge, Xibe is not present in any of the language models.

A complication arises from the fact that Xibe is written in a script derived from the Manchu alphabet, which is based on the traditional Mongolian script. Although the Xibe alphabet is included in the Unicode block for traditional Mongolian,[e] only twelve out of thirty-five Xibe letters share the same Unicode encodings with traditional Mongolian, while the remaining twenty-three letters use unique encodings. Therefore, we assume that mBERT and XLM-R will be at a serious disadvantage, while CINO can be assumed to have better but still limited capability in recognizing Xibe (see Section 5.2 for more details).

4.4 Parser

We use the implementation of the Deep Biaffine parser (Dozat and Manning 2017; Dozat *et al.* 2017) by Sayyed and Dakota (2021).[f] The parser is a neural graph-based dependency parser. The parser uses a biaffine classifier, and replaces a traditional bilinear or traditional MLP-based addition with biaffine attention. The deep bilinear attention allows for more relevant information to be retained before being used in the biaffine classifier.

[e] http://www.unicode.org/charts/PDF/U1800.pdf
[f] https://github.com/zeeshansayyed/multiparser

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 6. Hyper-parameter settings for the parser

Hyper-parameters	Value
Word embedding dimensions	100
POS tag embedding dimensions	100
Character embedding dimensions	50
PLM embedding dimension	100
Number of PLM layers Used	4
Number of LSTM layers	3
LSTM hidden layer dimensions	400
Learning rate	2e-3

The main reason for choosing this implementation is that the default parser implementation requires word embeddings, which are then concatenated with an additional embeddings model (e.g., word+POS, word+char). This behavior makes pure delexicalized experiments difficult: We would have to replace words by their POS tags (see Section 4.5), as the parser would concatenate these delexicalized embeddings with another embeddings model by default, for example, delexicalized+char, delexicalized+POS. The latter example would be a reduplication of POS embeddings. The implementation by Sayyed and Dakota (2021) has an option that enables single embeddings to be used without any concatenation, resulting in only a single delexicalized embeddings model when words are replaced with their POS tags, avoiding both embeddings concatenation and any reduplication.

We use the default hyper-parameters of the original base parser (Zhang, Li, and Min 2020), shown in Table 6. The word, POS tag, and character embeddings are randomly initialized. For the embeddings derived from the pre-trained language models, a scalar mixture of the last four layers is passed through a linear layer to produce the embeddings of the specified dimension (Peters *et al.* 2018; Tenney, Das, and Pavlick 2019a, 2019b).

4.5 Delexicalization

Since the Xibe alphabet is not represented in the training data of the pre-trained language models, we also experiment with delexicalized models of the parser. This will eliminate the script problem, but it reduces the parser's input to the coarse UD POS tags. We delexicalize all treebanks by replacing words by their UD POS tags. While this severely limits the amount of information available to the parser, it is a standard technique for pre-neural cross-lingual parsing (see Section 2.1).

4.6 Evaluation

We evaluate all experiments using the CoNLL 2018 Shared Task Scorer (Zeman *et al.* 2018).[g] Note that the scorer ignores dependency subtypes. Consequently, our evaluation is based on the main types, even though subtypes are present during training and are assigned in the parsing process. We report the unlabeled (UAS) and labeled attachment score (LAS), both macro-averaged, but mainly focus on the analysis of LAS. We additionally provide analyses by dependency labels and POS tags.

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

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

For significance testing, we use Dan Bikel's Randomized Parsing Evaluation Comparator (Yeh 2000).[h] The null hypothesis (H_0) is that for each test sentence, scores obtained from the two models are equally likely. For each setting, the weighted average UAS and LAS are calculated per model and per sentence. To determine significance, the script loops over the UAS and LAS scores of each sentence and randomly chooses to switch the sentence's scores across models. After a single pass, the scorer checks whether the new weighted UAS and LAS scores are more extreme than the originally calculated averages. This procedure is repeated for a predefined number of iterations after which a form of randomized paired-sample t-test for hypothesis testing is performed with respect to how often the new averages were more extreme. The results are the *p*-values for both the differences in UAS and LAS between the two models. We only report significance testing for LAS since in our experiments, both scores show the same trends.[i]

5. Monolingual parsing

In this section, we investigate parsing quality when training and testing on the small Xibe treebank. This gives us the (potentially competitive) baseline given a setting where we have a small amount of sentences available. It is an open question whether pre-trained multilingual word embeddings can provide useful information in a situation where the target language is not included in the embeddings, and neither is its script. We assume that the subword embeddings from these models can still provide useful information since the word embeddings are trained on a small training set of 400 sentences.

For the experiments, we train both lexicalized and delexicalized models. For the lexicalized models, we can use different combinations of representations as features: We use POS tag embeddings, character embeddings, and embeddings from large-scale pre-trained language models (see Section 4.3 for a description) in combination with word embeddings derived from the training data. The delexicalized model replaces words by their corresponding POS tags.

5.1 Results

The results for the monolingual experiments using different feature representations are shown in Table 7. When using only word embeddings trained on the Xibe training set, we reach an LAS of 56.57. Adding POS embeddings (WORD + POS) decreases the results, but adding character embeddings (WORD + CHAR) improves results, as expected, to an LAS of 62.34. That a concatenation of word and character embeddings can improve parsing performance has previously been shown across typologically diverse languages (Sayyed and Dakota 2021), pointing to their superior ability to handle out-of-vocabulary (OOV) words (Vania *et al*. 2018).

Results for using the pre-trained multilingual language models are mostly negative, with LAS results around or slightly below the results for word embeddings. The only exception is the setting when we use the combination of word embeddings, character embeddings, and CINO subword embeddings (WORD + CHAR + CINO). This results in the highest LAS of 62.47.

Table 7 also shows the difference in LAS between the best model and each of the remaining models, along with significance results for this comparison. Note that there is no significant difference between the best performing model and WORD + CHAR, WORD + CHAR + POS, and WORD + CHAR + POS + CINO, thus indicating that CINO's contribution is rather minimal. We assume that this may be due to the fact that the Xibe script has only minimal overlap with Mongolian in CINO, but no overlap with any of the other languages.

However, these results raise a number of questions: Is the missing overlap in scripts the reason for the poor performance of the pre-trained multilingual embeddings? Why do the POS

[h] https://github.com/tdozat/Parser-v1/blob/master/bin/compare.pl

[i] It is possible for either the UAS or LAS score to be significant while the other is not.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 7. Results for parsing Xibe in a monolingual setting using different feature combinations.

Feature	UAS	LAS	$diff_{LAS}$
WORD	68.19	56.57*	5.90
WORD + POS	67.19	55.52*	6.95
WORD + CHAR	71.65	62.34	0.13
WORD + CHAR + POS	71.69	62.17	0.30
POS ONLY (delexicalized)	70.06	59.93*	2.54
POS + VERBAL SUFFIX	71.08	61.67*	0.80
WORD + MBERT	67.82	56.09*	6.38
WORD + XLM-R	68.03	55.71*	6.76
WORD + CINO	69.25	57.79*	4.68
WORD + POS + CINO	69.07	57.44*	5.03
WORD + CHAR + CINO	**72.11**	**62.47**	-
WORD + CHAR + POS + CINO	71.94	62.26	0.21

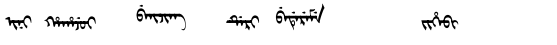
$diff_{LAS}$ shows the LAS difference to the best model (WORD + CHAR + CINO); * indicates significant differences ($p_{LAS} < 0.05$)

embeddings perform so poorly given that they provide gold POS information? How severe is the OOV problem when using Xibe word embeddings? We will investigate these issues below.

5.2 A closer look at pre-trained language models

Our results for monolingual parsing in Table 7 are somewhat surprising if we ignore the missing overlap in scripts: None of the pre-trained language models manage to outperform a model of Xibe word embeddings plus character embeddings (WORD + CHAR). In fact, two of the pre-trained models perform significantly worse. In this section, we will investigate whether the missing overlap in scripts is the reason for this performance.

As described in Section 4.3, the multilingual pre-trained language models that we use are mBERT, cross-lingual XLM-R, and CINO. Neither of the training corpora of the three models contains any Xibe data. However, the three language models use different tokenizers, which may influence parsing in different ways.

(3) 

ini hahajui beijing deri bedereme jihebi

his son Beijing ABL return.CONV come.PAST

"His son came back from Beijing."

We use a randomly chosen treebank sentence, see example (3), and tokenize it using the three pre-trained language models. The results are shown in Table 7. The table shows clearly that mBERT is not equipped to handle an unknown script: It does not recognize the Xibe characters. As a consequence, all words are treated as unknown tokens (UNK). A closer look at mBERT tokenization of the whole Xibe treebank shows that all words written in the Xibe script are represented as UNK by the mBERT tokenizer, only punctuation, numbers, and words spelled in Roman characters, such as "wifi," are handled by the tokenizer, but these are very few in number.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 8. Subwords for the sentence in example (3) tokenized by the pre-trained language models

mBERT:	["[UNK]," "[UNK]," "[UNK]," "[UNK]," "[UNK]," "[UNK]"]
XLM-R:	[_, ᡳᠨᡳ, _, ᡥ, ᠠ, ᡥ, ᠠ, ᠵᡠᡳ, _, ᠪᡝᡳᠵᡳᠩ, _, ᡩᡝᡵᡳ, _, ᠪᡝᡩᡝᡵᡝᠮᡝ, _, ᠵᡳᡥᡝᠪᡳ]
(transliterated)	[_, ini, _, h, a, h, a, jui, _, beijing, _, deri, _, bedereme, _, jihebi]
CINO:	[_, ᡳ, ᠨ, ᡳ, _, ᡥ, ᠠ, ᡥ, ᠠ, ᠵᡠᡳ, _ᠪ, ᡝᡳᠵᡳᠩ, _, ᡩᡝ, ᡵ, ᡳ, _ᠪ, ᡝᡩᡝ, ᡵ, ᡝ, ᠮ, ᡝ, _, ᠵᡳ, ᡥ, ᡝ, ᠪ, ᡳ, _]
(transliterated)	[_, i, n, i, _, h, a, h, a, jui, _b, eijing, _, de, r, i, _b, ede, r, e, m, e, _, ji, h, e, b, i, _]

XLM-R, in contrast, uses the SentencePiece tokenizer, which encodes the string as a Unicode sequence and decodes the normalized Unicode sequence to a string. This procedure seems to work better, and XLM-R is able to tokenize some words. However, there are considerable differences across the words, one word (*hahajui*; Eng: son) is split into characters, all the others are kept as complete words.

To determine the extent of the problem, we check how many words in the Xibe treebank are not subdivided into subwords (excluding all punctuation). Results show that 60.68 percent (12,461 out of 20,535) of the words are kept in their original form. This means that for Xibe, the XLM-R tokenization is closer to word-level tokenization, which explains why the XLM-R results are very similar to the results when using word embeddings in Table 7. The fact that the XLM-R results are slightly lower than those for word embeddings shows that the tokenization that does happen generally does not provide useful information to the parser.

CINO utilizes the same architecture as XLM-R. The main differences between CINO and XLM-R concern the SentencePiece tokenizer and the data used to train embeddings (see Section 4.3). Table 7 shows that the WORD + CINO model outperforms the WORD model by 1.22 percent points in LAS, suggesting that using the additional Tibetan and Mongolian information is helpful. The effectiveness of CINO can partially be attributed to its tokenizer which is more capable of splitting words into subwords, as shown in Table 8. When using the CINO tokenizer, only 7.43 percent (1,525 out of 20,535) of the words are not split into subwords. However, the example shows that this tokenizer has a tendency to tokenize words into individual characters, which explains the similarity of the parsing results to those of the WORD + CHAR model. One reason for this behavior may be found in the fact that Xibe and traditional Mongolian only share twelve out of thirty-five letters, that is, CINO is not capable of recognizing the remaining twenty-three letters. However, a closer look at the marked case, that is, subwords longer than one character, shows that these consist of only Mongolian characters, only Xibe characters, and a mix of both in almost equal parts. Consequently, familiarity with the Mongolian characters is not the only criterion for forming longer subwords.

Overall, it is evident that all pre-trained multilingual word embeddings have problems determining useful subword units given Xibe's unique script, which explains the low results for models using those embeddings.

5.3 A closer look at word plus part-of-speech embeddings

In general, integrating POS information with the Xibe word embeddings is expected to improve parsing results since the parser has access to gold POS tags and should thus be able to better disambiguate lexical ambiguities. However, from Table 7, we observe that parsing performance declines when part-of-speech embeddings are added to the word embeddings. In comparison to the model using word embeddings, the WORD + POS model reaches an LAS that is more than one point lower (55.52 vs. 56.57). Additionally, when adding POS embeddings to the best performing model (WORD + CHAR + CINO), the LAS deteriorates slightly from 62.47 to 62.26. The

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 9. Accuracy of correct head attachment per POS tag (sorted in descending order by the absolute frequency of the POS tag in the Xibe treebank)

POS	WORD + CHAR	WORD + CHAR + POS	diff	# POS
NOUN	60.88	60.86	−0.02	6,969
VERB	77.08	77.86	0.78	4,665
ADP	87.13	86.69	−0.44	2,697
ADJ	65.03	65.28	0.25	1,544
ADV	67.61	65.48	−2.13	985
PRON	69.87	68.9	−0.97	926
NUM	68.73	66.46	−2.27	793
PROPN	64.89	63.7	−1.19	675
X	42.95	44.26	1.3	305
PART	74.24	79.8	5.56	198
AUX	53.76	57.23	3.47	173
DET	74.39	75.0	0.61	164
CCONJ	55.28	52.17	−3.11	161
SCONJ	63.11	59.84	−3.27	122
INTJ	60.0	80.0	20.0	5

same trend also occurs when adding POS embeddings to WORD + CINO. While these differences are not significant, they still show that the gold POS information cannot be used successfully by the parser.

Table 9 provides a comparison of the WORD + CHAR and the WORD + CHAR + POS model in terms of the accuracy per POS tag of having been assigned the correct head, along with the absolute frequency of this POS tag in the Xibe treebank. The accuracies show that there is no clear trend, about half of the POS tags improve when POS tags are added, the other half deteriorates, and the split is distributed across frequencies and across open and closed class POS tags. This lack of trends is also evident in other forms of probing the results. We interpret this lack of trends as an indication of a complex interaction of the different types of embeddings, without any clear reasons for the lower accuracies when POS embeddings are used. However, we also acknowledge that this may be an artifact of the small training set size.

5.4 A closer look at word embeddings

Here, we investigate the reasons for the poor performance of the Xibe word embeddings model as compared to the delexicalized POS model. When representing a sentence via word embeddings, the LAS for this model is 3.36 points lower than the delexicalized POS model and 5.77 points lower than the WORD + CHAR model. Since the lexicalized model is trained on only 400 sentences and tested on a second fold of 400 sentences, we assume that the main problem consists of OOV words. We check the OOV ratio in the three folds of the test data, shown in Table 10. The table shows that the OOV ratio on average is 46.38 percent. Since nearly half of the words in the test data are unknown, it is not surprising that this model is struggling. We further check the number of unknown words on the sentence level, finding that the test sentences on average have 17.99 percent

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 10. Out-of-vocabulary ratio in three folds of the Xibe treebank

Train	Test	OOV	OOV/sent$_{test}$
fold$_1$	fold$_3$	46.99%	18.13%
fold$_2$	fold$_1$	46.30%	17.88%
fold$_3$	fold$_2$	45.85%	17.97%
Average		46.38%	17.99%

of unknown words relative to the training data. The missing information degrades prediction performance. The OOV problem is mitigated by the WORD + CHAR model, which reaches an LAS very close to the highest LAS for monolingual parsing.

5.5 A closer look at the delexicalized model

In contrast to the low performance of the WORD model, the delexicalized model performs significantly better. Using the coarse-grained universal part-of-speech tags instead of words alleviates the OOV problem of the word embeddings model. Part-of-speech sequences can encode local syntactic information, which the parser mainly relies on. However, the rather coarse-grained universal POS tags tend to cluster together many relevant syntactic distinctions. For example, all verbs are replaced by "VERB," and morphological information is largely lost. This may be detrimental if some of the encoded information is necessary for syntactic decisions.

We now investigate whether adding suffix information to the POS tag will provide the parser with better information. For this approach, we need to walk a fine line between adding syntactically relevant information and increasing the size of the POS tagset and consequently the OOV rate, given the small size of the treebank. We focus on verbs since these have the richest morphology in Xibe. For each verb, instead of representing the verb form by the POS tag "VERB," we attach the verbal suffix to the POS tag. For example, we use "VERB-ᡥᡳ" for ᡨᠠᠴᡳᠮᠪᡳ *taci-mbi*, "VERB-ᡥᡝ" for ᡤᡝᠨᡝᠮᡝ *gene-me*, and "VERB-ᡵᡝ" for ᠨᡳᡠᡵᡝ *niru-re*.

Since there does not exist a morphological analyzer for Xibe, we use a bootstrapping approach, extracting the relevant suffixes using a set of rules. Table 11 lists all the inflectional suffixes for Xibe verbs based on *Xiboyu Yufa Tonglun* (Eng.: General Introduction to Xibe Grammar; Šetuken 2009), including affirmative forms and negative forms. We use this list of suffixes to extend the verbal POS tag.

To ensure that we do not overgeneralize, we have manually compared the "VERB + suffix" tags and the original verbs. We found only three erroneous cases. These suffixes were removed before parsing.

After applying the rules, 95.6 percent of all verbs are represented by a more specific POS tag including a suffix. The remaining verbs are mainly uninflected verbs, participles marked for case, and irregular verbs.

In order to investigate whether adding verbal morphology to the POS tag provides relevant information, we compare results using this version of delexicalization to the standard method. We are aware that this model may not be completely delexicalized since we use suffix information, but POS tags often include partial morphological information, such as plural or past tense in the Penn Treebank POS tagset for English (Santorini 1990) and the finite/infinite verb distinction in the Penn Treebank POS tagset and the Stuttgart-Tübingen tagset for German (Schiller, Teufel, and Thielen 1995). However, we consider this method a knowledge-poor method for injecting morphological information into the parser.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 11. Inflectional suffixes of Xibe verbs based on the *General Introduction to Xibe Grammar* (Šetuken, 2009)

Suffix type	Meaning	Affirmative form	Negative form
Finite	Present/future	mbi	-
		ra, re, ro	rakv, rekv, rokv, rkv
	Progressive	mahabi	mahakv bi
	Simple past	ha, he, ho	hakv, hekv, hokv
		ka, ke, ko	kakv, kekv, kokv
	Past indefinite	habi, hebi, hobi	hakvbi, hekvbi, hokvbi
		kabi, kebi, kobi	-
	Question	mbio, rao, reo, roo, kvn	-
	Oblique moods	ki, cina, kini	-
Converb	Conditional	ci	-
	Concessive	cibe	-
	Imperfect	me	-
	Perfect	fi, pi, mpi	-
	Durative	hai, hei, hoi	-
	Terminative	tala, tele, tolo	-
	Instrumental	tai, tei, toi	-
	Preparative	nggala, nggele, nggolo	-
Participle	Imperfect	ra, re, ro	rakv, rekv, rokv, rkv
	Perfect	ha, he, ho	hakv, hekv, hokv
		ka, ke, ko	kakv, kekv, kokv

Table 12. Parsing results of the POS ONLY model and the POS + VERB SUFFIX model

Model	UAS	LAS	$diff_{LAS}$
POS ONLY (Delexicalized)	70.06	59.93	-
POS + VERB SUFFIX	71.08	61.67*	1.74

$diff_{LAS}$ shows LAS improved from the POS ONLY model to the POS + VERB SUFFIX model. * indicates significant difference ($p_{LAS} < 0.05$)

The results of this experiment are shown in Table 12. A comparison of the two models shows that the LAS significantly improves by 1.74 points to 61.67, thus indicating clearly that the morphological information present in the verbal suffixes is important for parsing.

Table 13 gives a selective overview of how adding suffixes changes the parser's performance on the verb POS tags as well as on clausal dependencies. The results show a considerable increase in accuracy for the POS tag VERB, which improves by 4.57 percent points. The dependency relations for clauses show increases between 3.78 and 37.18 percent points. The most significant change occurs for clausal complements, which increase from an F-score of 0.00 to 37.18.

A manual inspection of these cases between the two models shows that in the POS ONLY model, "ccomp" relations tend to be misclassified as "advcl." This over-generalization is partly corrected in the POS + VERB SUFFIX model. Figure 4 shows one example. Here, the main verb is the imperative *se*, *bi sain fonjimbi* is its clausal complement. Since "advcl" is more frequent than "ccomp,"

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 13. F1 score of VERB and dependency relations for clauses of the POS ONLY model and the POS + VERB SUFFIX model

POS/relation	POS ONLY	POS + VERB SUFFIX
VERB	75.26	79.83
acl (adnominal clause)	38.83	44.29
acl:relcl (relative clause)	26.80	30.58
advcl (adverbial clause)	64.65	76.67
ccomp (clausal complement)	0.00	37.18
parataxis (run-on sentences)	9.42	33.08
xcomp (open clausal complement)	25.41	30.63

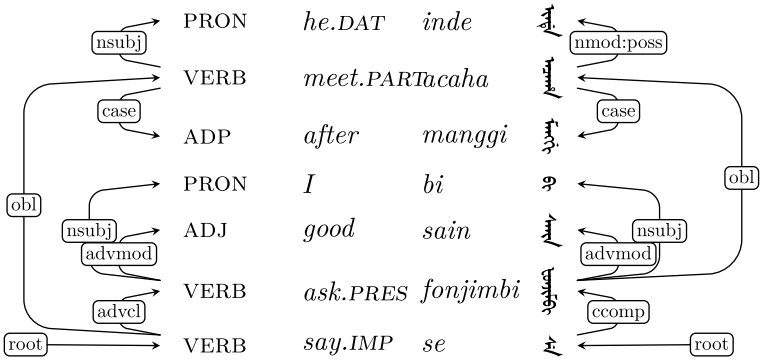


Figure 4. Parse by the POS ONLY model (left) and the POS + VERB SUFFIX model (right). Eng. "After (you) meeting with him, tell (him) that I ask after (his) health."

the parser prefers this dependency label over "ccomp" when it does not have access to verbal features. Adding verb suffixes to the POS tag provides the information that *fonjimbi* is finite and thus cannot be a converb.

Our results show clearly that the verbal suffixes provide explicit cues for identifying syntactic patterns.

6. Cross-lingual dependency parsing

As we have seen in Section 5, monolingual parsing using a pre-trained multilingual language model is only moderately successful. In monolingual parsing, the parser only has a limited number of sentences from which to learn Xibe syntax, and the substantial proportion of unknown words creates a challenge. Another possibility of addressing the problem consists in cross-lingual parsing, by training a parser on a *source* language which is syntactically closely related and higher resourced,[j] and parse Xibe. Here, we rely on the commonality of the UD annotation scheme.

As shown in Section 2.2, lexicalized approaches can lead to good results in a cross-lingual setting. Since in the case of Xibe, there will be no or very little overlap in the scripts between source and target language, we assume that a lexicalized approach will not be successful. In Section 5,

[j]Note that two of the languages we use, Buryat and Kazakh, have fewer annotated sentences than Xibe. We include them to cover the Cyrillic script in the training data in addition to Arabic, Latin, Hangul, and Kanji/Kana.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 14. Cross-lingual dependency parsing results.

Feature	word		word+char		word+char+CINO		Delexicalized		
Source Treebank	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	$diff_{LAS}$
Buryat (split 1)	25.48	9.47	20.41	8.39	31.33	12.94	46.36	29.69*	30.24
Buryat (split 2)	16.88	3.23	15.50	2.74	32.34	12.03	47.34	30.58*	29.35
Buryat (split 3)	24.15	9.23	30.73	10.63	23.24	10.07	48.86	32.20*	27.73
Kazakh (split 1)	25.10	11.17	32.37	13.70	20.91	10.27	53.93	40.30*	19.93
Kazakh (split 2)	27.35	11.47	31.38	11.29	26.91	10.59	54.44	40.53*	19.40
Kazakh (split 3)	27.20	11.65	14.75	7.51	28.81	12.64	54.46	**41.42***	18.51
Uyghur	25.23	11.09	34.54	12.46	20.41	8.12	57.31	39.90*	20.03
Turkish_Kenet	17.92	4.54	24.05	11.57	36.40	13.95	54.15	39.21*	20.72
Turkish_Penn	17.62	3.53	33.19	13.02	40.90	15.63	51.92	34.58*	25.35
Turkish_BOUN	26.66	7.10	26.67	3.75	17.36	7.00	**58.58**	**41.00***	18.93
Korean GSD	26.98	6.66	12.79	4.24	28.91	13.12	47.00	33.37*	26.56
Korean-Kaist	30.80	9.26	10.95	7.24	41.21	15.67	35.70	22.03*	37.90
Japanese-GSD	12.35	5.36	24.78	8.78	37.43	15.38	51.22	35.04*	24.89
Japanese GSDLUW	18.96	6.86	26.47	9.18	40.37	15.92	53.21	36.75*	23.18
Xibe Baseline	68.19	56.57	71.65	62.34	72.11	62.47	70.06	59.93	-

$diff_{LAS}$ is the LAS difference between the delexicalized source model with the Xibe baseline. * indicates significant differences ($p_{LAS} < 0.05$)

we have shown that unlexicalized parsing is highly successful in a monolingual setting. Our assumption is that this will transfer to a cross-lingual setting as long as the languages are syntactically similar enough since in this case, the parser can profit from the (generally) higher number of training sentences. As source languages, we use the 6 languages from the Transeurasian language family, as described in Section 4.1.

Since in monolingual parsing, using word and character embeddings performs better than only using word embeddings and word plus POS embeddings, we also use the WORD + CHAR setting here. We also use character embeddings and CINO embeddings for the same reason.

6.1 Results

Results of the cross-lingual dependency parsing experiments are shown in Table 14. These results are considerably lower that the monolingual baseline (repeated in the last row of Table 14), across the board. The highest LAS is reached by training on one data split of the Kazakh treebank, in the unlexicalized setting. This reaches an LAS of 41.42, which is about 18.5 percent points lower than the result of the corresponding monolingual baseline.

As expected, all the lexicalized models perform poorly in the cross-lingual setting. The results in Table 14 show that the results of the word based models are in the range between 3.23 and 11.65 for LAS. Incorporating characters to the word model improves results in about two-thirds of the cases; however, it is unclear how to delineate the positive and negative cases, since even the different splits of Buryat and Kazakh show inconclusive results. Adding CINO embeddings generally improves performance to a limited extent, exceptions are Kazakh split 2 + 3 and Uyghur.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

However, even those improved results do not manage to improve beyond an LAS of 21.27 (based on the Japanese GSDLUW treebank).

Note that the six source languages cover a wide range of scripts, but do not have any overlap with the Xibe script. This leads us to the conclusion that lexicalized cross-lingual parsing is not a viable option in cases where the target script is not included in the training data.

In contrast to the lexicalized models, the delexicalized models perform better. The best setting reaches an LAS of 41.42 when training on the Kazakh treebank (split 3). However, there are significant differences across the three splits of the Kazakh treebank, indicating that this result may not be stable. The next highest result is based on training on the Turkish BOUN, which reaches an LAS of 41.00. However, these results are still substantially lower than those of the monolingual delexicalized model (LAS: 59.93). this means that the parser cannot harness the larger training data size to improve parsing results.

In the investigation of causes for the low performance of all cross-lingual models, the first assumption would be that the source and target language differ in terms of word order and types of syntactic constructions. However, we chose our source language mainly by language family, ensuring that all source languages show the same large scale syntactic characteristics as Xibe: All languages have a strict SOV word order, and phrases are all head final. This does not preclude other word order differences across languages, but these should have less effect on parsing accuracy. For this reason, we have a closer look at the dependency labels next, to determine possible reasons for the low performance.

6.2 Dependency relations

In this section, we investigate potential discrepancies in dependency labels between the source treebanks and Xibe as our target treebank. It is a well known fact that not all UD treebanks use the full set of dependency labels (Nivre *et al.* 2016, 2020; de Marneffe *et al.* 2021). For this reason, we first concentrate on the overlap between the labels in the Xibe treebank and each source treebank to decide whether there is a correlation between the (lack of) overlap and parsing performance.

The Xibe treebank uses all forty dependency relations, including the thirty basic dependency relations and ten relations with subtypes. Since the evaluation focuses on main types, we restrict our investigation to those as well.

Table 15 lists the dependency relations that do not occur in the training data of a given source treebank and the percentage of all the unknown labels in the Xibe test set. The table shows that there is a certain interdependence between the rate of unknown labels and parser performance: The source treebanks that reach an LAS higher than forty on Xibe, that is, Kazakh and the Turkish BOUN, all show minimal rates of unknown labels. However, the remaining picture is less clear. The Korean GSD, and the Japanese GSD and GSDLUW treebanks have the highest unknown label rates of around 7 percent. However, even though these rates are very similar, the parsing performance ranges between an LAS of 33.27 (Korean GSD) and 36.75 (Japanese GSDLUW). Additionally, the Korean Kaist treebank has an unknown label rate of only 1.49 percent, but results in a considerably lower LAS of 22.03. Consequently, we can conclude that while having a low unknown label rate is a prerequisite for high parsing accuracy, it alone is not sufficient to guarantee good performance.

We now look more closely at the parsing performance of specific dependency relations parsed by the highest performing model of Turkish BOUN along with Korean Kaist and the two Japanese models. Specifically, we choose the dependency relations with the highest frequency in the Xibe treebank: nominal subject (nsubj), object (obj), oblique (obl), case marking (case), and adverbial clause (advcl). These labels account for 40.85 percent of the dependency relations in the Xibe treebank.

The F1 scores for the five labels are shown in Table 16. The results show that all parsing models have significant difficulties locating subjects, objects, and obliques. All of the F-scores for these

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 15. Number of unseen labels between the training and test treebank. The average rate reports the percentage of these labels in the Xibe test set

Source Treebank	Unknown Label	Rate (%)	LAS
Buryat	clf	0.45	30.81
Kazakh	fixed	0.68	40.75
Uyghur UDT	clf, iobj	0.52	39.90
Turkish Kenet	cop	0.29	39.21
Turkish Penn	cop	0.29	34.58
Turkish BOUN	-	0	41.00
Korean GSD	compound, parataxis, xcomp clf, discourse, vocative	7.16	33.27
Korean Kaist	parataxis, clf	1.49	22.03
Japanese GSD	conj, parataxis, flat, xcomp appos, clf, vocative, iobj	6.98	35.04
Japanese GSDLUW	conj, parataxis, flat, xcomp appos, clf, vocative, iobj	6.98	36.75

Table 16. F1 scores for dependency relations obtained from four delexicalized models training, respectively, on the Turkish BOUN, Korean Kaist, and Japanese GSD and GSDLUW treebanks

Relation	Turkish BOUN	Korean Kaist	Korean GSD	Japanese GSD	Japanese GSDLUW
advcl	20.88	0.16	21.14	10.29	15.80
case	84.23	8.19	60.57	83.53	85.19
nsubj	17.99	6.57	21.59	0.77	0.50
obj	20.11	26.63	15.90	22.11	25.65
obl	22.14	0.00	0.25	19.22	19.13

dependency relations range between 0.00 and 26.63. Nominal subjects seem to be the most difficult to identify given Japanese source treebanks while Korean source treebanks lead to problems in identifying obliques.

As described in Section 3, Xibe has a strict SOV order, and the main verb is in sentence final position. This strict word order should provide the parser with cues about the order of arguments. However, in delexicalized parsing, it is more difficult to differentiate between the arguments. Figure 5 shows the dependency tree of example (1) plus an incorrect parse. Here, we have a pro-drop subject, plus an oblique and an object. In delexicalized parsing, the parser only has access to the POS level, which obscures all case information in the adpositions, thus leaving the parser guessing. The parse shows that all three adpositional phrases are attached to the verb as obliques. This is a clear indication that we need to provide more case information to the parser.

Another problem that we observe from Table 16 is that none of the five source treebanks allow an accurate recognition of adverbial clauses (advcl), the F-scores range from 0.16 (Korean Kaist) to 20.88 (Turkish BOUN). Next, we compare the two Korean treebanks with the Xibe treebank. More specifically, we determine the types of heads and dependents that share the "advcl" dependency relation. We list the five most frequent head-dependent pairs per treebank in Table 17. In the Korean Kaist treebank, the most frequent types all have an adverbial (ADV) dependent. In the Korean GSD treebank, in contrast, the most frequent dependents are verbs. This is a clear

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 17. The five most frequent head-dependent pairs of adverbial clauses ("advcl") in the Korean Kaist, GSD, and Xibe treebank

Korean Kaist			Korean GSD			Xibe		
Dependent	Head	Rate (%)	Dependent	Head	Rate (%)	Dependent	Head	Rate (%)
ADV	VERB	60.26	VERB	VERB	81.52	VERB	VERB	93.96
ADV	SCONJ	11.69	VERB	ADJ	10.41	VERB	ADJ	1.41
ADV	AUX	6.26	VERB	NOUN	4.65	ADJ	VERB	1.27
ADV	ADJ	6.02	NOUN	VERB	1.65	VERB	NOUN	1.23
ADV	CCONJ	4.67	VERB	ADV	0.75	NOUN	VERB	0.61

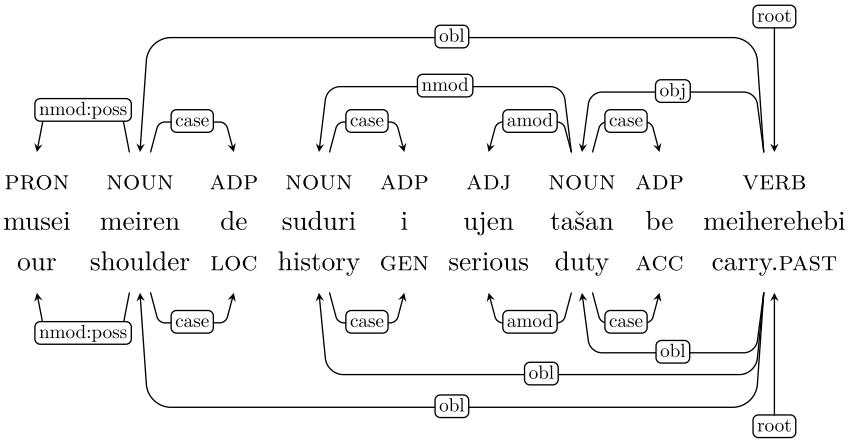


Figure 5. Dependency tree and an erroneous parse, for example (1). The edges above the POS tags show the treebank annotation, the edges below show the parse using the Turkish BOUN model for training.

indication of variability in annotation. For Xibe, 93.96 percent of the adverbial clause relations consist of a VERB depending on a VERB head. This is a consequence of the frequent use of converbs, as described in Section 3.4. Korean also uses converbs, but to a lesser degree. Whether the higher variability of dependents in the Xibe treebank is due to differences in language phenomena or due to annotation artifacts is difficult to determine, but it is clear that these differences lead to a deterioration in parsing performance in a cross-lingual setting.

In summary, our investigation has shown that the low results in cross-lingual parsing have a range of underlying causes, including mismatches in the set of dependency labels used, differences in annotation, and differences in syntactic preferences across different languages, even when we choose the languages that are the most closely related to the target language (and that possess a UD treebank). Additionally, since delexicalized parsing needs to rely on the seventeen POS tags in the UD POS tagset as representation of a sentence, these categories are too coarse-grained to allow the parser to make reliable syntactic decisions.

In the next section, we focus on one method to address differences in syntax and annotation automatically: we combine all source languages but then determine the subset of sentences that is the most similar to the target language.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 18. Multi-source parsing results using the full multi-source treebank, and subsets with sentence perplexity scores lower than $n = \{10, 15, 20\}$.

Train	Sent.num	UAS	LAS	$diff_{LAS}$
BOUN	9,768	58.58	41.00	–
full	17,905	65.63	49.58*	8.58
$PPL \leq 10$	11,837	65.13	49.21*	8.21
$PPL \leq 15$	15,907	**65.83**	**49.86***	8.86
$PPL \leq 20$	17,202	65.64	49.50*	8.50
Xibe	1,200	70.06	59.93*	18.93

$diff_{LAS}$ is the difference between the Turkish BOUN model, the multi-source models and the Xibe baseline.
* indicates significant differences ($p_{LAS} < 0.05$)

6.3 Multi-source parsing

As described above, in single source, delexicalized cross-lingual dependency parsing, the parser cannot profit from the larger training set size of a source language because of differences between languages, differences in annotation, and too coarse-grained information on the POS level. To address the first two issues, we investigate multi-source parsing, where we combine all reliable source languages, but then select only those sentences that are the most similar to the target language using perplexity.

We restrict ourselves to only those treebanks that have obtained an LAS greater than 36.00 in Table 14, that is, Kazakh, Uyghur, Turkish BOUN, Turkish Kenet, and Japanese GSDLUW. By having two Turkish treebanks, both of which are larger (Turkish Kenet with 18,687 sentences and Turkish BOUN with 9,768 sentences), this may give Turkish too much influence in the multi-source setup. Initial experimental results when including Turkish Kenet treebank yielded lower results, suggesting this may indeed be the case. For this reason, the final set of treebanks consists of only Kazakh, Uyghur, Turkish BOUN, and Japanese GSDLUW treebanks.

6.3.1 Perplexity

Perplexity has been shown to be good metric when selecting additional training data for cross-lingual parsing (Søgaard, 2011) as well as in other multi-source setups, such as domain adaptation (Hwa 2001; Khan, Dickinson, and Kübler 2013). To calculate perplexity we use NLTK (Loper, Bird, and Klein 2009) to train a POS bigram language model on the Xibe treebank. Laplace smoothing is performed with an α of 1 to handle the high likelihood of unseen POS sequences on such small datasets. Then for each sentence in the selected source treebanks, we compute the perplexity and select those below thresholds of 10, 15, and 20t, respectively, as additional training data.

6.3.2 Results

Table 18 shows the results of this experiment. For convenience, we repeat the results for using the Turkish BOUN treebank as source language and for monolingual training on Xibe from Table 14. The results show that we gain about 7 percent points in UAS and 8.6 percent points in LAS when we use all four reliable source treebanks for training. While this improves results, they are still about 4 percent points below the monolingual UAS and 10 percent points below the monolingual LAS. Restricting the sentences to those with a perplexity lower than 15 gives a slight boost to a UAS of 65.85 and an LAS of 49.86 by deleting about 2,000 sentences. Lowering the perplexity to 10 decreases the results below the scores of the full set of treebanks.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Table 19. Label performance in monolingual, single-source cross-lingual, and multi-source cross-lingual parsing

		Cross-lingual	
Label	Monolingual	Single source	Multi-source
advcl	64.65	20.88	45.92
case	86.64	84.23	87.53
nsubj	38.41	17.99	26.06
obj	45.17	20.11	20.30
obl	27.61	22.14	27.00

Table 19 provides an analysis of the dependency labels that show the most differences across the individual settings. We use the Turkish BOUN model for the single-source results and the multi-source ($PPL \leq 15$) for the multi-source results. The results show a sizable difference in adverbial clauses (advcl): This dependency relation is parsed most reliably in the monolingual setting, it fares poorly in the single-source cross-lingual setting, but is handled more successfully in the multi-source setting. This result makes sense when we consider what we know about converbs (see Section 3.4): They are a highly frequent construction in Xibe, which explains the good results in the monolingual setting. Since the other source languages also use converbs, the decrease in performance in the single-source model shows that the source languages use them in fewer or more restricted functions. By combining source languages into the multi-source training set, we provide the parser with a wider range of converb functionality, which helps the parser to handle this construction more successfully.

The results of the multi-source experiments fall in line with other work (Søgaard, 2011; Rosa and Žabokrtský, 2015a) showing that it is possible to improve the results of delexicalized cross-linguistic parsing by choosing a more relevant set of source data. We see that the pure size of a source treebank is less important than similarity to the target languages, since the Korean Kaist treebank is larger than our best multi-source treebank, and the Turkish Kenet treebank is about the same size, but both result in considerably lower performance. Combining source languages can provide the parser with a wider range of phenomena than are present in any single-source language. While this may lead to a decrease in performance if the target language does not use this full range, it can also improve performance, as in the case of the Xibe converbs.

However, we also see that the similarity of sentences is not a panacea, the cross-lingual results are still far below the monolingual ones. A comparison of POS bigrams, especially given the rather coarse-grained nature of UD POS tags, may not provide enough information to find the most relevant sentences.

7. Conclusion

In this work, we have investigated parsing for an under-resourced language, Xibe, which uses a unique script that is not present in any of the other languages for which we have resources. We first investigated a monolingual setting, determining how best to parse Xibe using the small treebank that is available. These results show that we reach the best results when we combine word and character embeddings with the CINO language model. Since all other language models lead to a deterioration of parsing performance, we come to the conclusion that this improvement results from having Mongolian included in the language model, since Mongolian is the only language that shares a subset of characters with Xibe. We also show that we reach competitive results when

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

parsing completely delexicalized, that is, by focusing on POS instead of word embeddings. We also show that the UD POS tagset is impoverished and that adding automatically extracted verbal suffixes to the POS tags improves results.

In a second set of experiment, we have focused on a cross-lingual setting where we train on related source languages. Our investigation has shown that this setting leads to significantly lower results than training on the small Xibe training set. These low results have a range of underlying causes, including mismatches in the set of dependency labels, differences in annotation, and differences in syntactic preferences across different languages, even though we chose closely related languages. None of these issues are new or different from settings where source and target share the same script, but they are exacerbated by the difference in script since a multilingual language model cannot provide the necessary bridge across the languages. Using delexicalization to bridge the languages is necessary but also comes with a high price: We abstract away from specific types of information, such as case information, which is necessary for many parsing decisions.

Our next steps need to focus on injecting more information into the POS tags. In a way, this is reminiscent of the parsing situation before the arrival of large scale neural language models, where much attention was paid to finding the optimal level of abstraction between using words or POS tags as input for the parser. While neural parsers have drastically improved performance, they can still be too dependent on lexical information making them not robust enough to lexical variation (Kasai and Frank 2019), while simultaneously questions persist about the granularity of POS tags for a given language within a standard annotation scheme (e.g., UD; Anderson and Gómez-Rodríguez 2020) or in some instances even their necessity (Zhou *et al.* 2020b). This shows very clearly that the problems in parsing remain the same over time, even though we are making progress in addressing them.

References

Ahmad W., **Zhang Z.**, **Ma X.**, **Hovy E.**, **Chang K.-W. and Peng N.** (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*, Minneapolis, MN, pp. 2440–2452.

Ammar W., **Mulcaire G.**, **Ballesteros M.**, **Dyer C. and Smith N. A.** (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics* **4**, 431–444.

Anderson M. and Gómez-Rodríguez C. (2020). On the frailty of universal POS tags for neural UD parsers. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Online, pp. 69–96.

Blaschke V., **Schütze H. and Plank B.** (2023). Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 40–54.

Chog (2016a). Area Strategy and Northeast Asia. Social Sciences Academic Press (China)*(In Chinese)*.

Chog (2016b). Protection and academic values of endangered minority languages in China. In Chog (2016a), pp. 17–35 (In Chinese).

Cohen S. B., **Das D. and Smith N. A.** (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, pp. 50–61.

Conneau A., **Khandelwal K.**, **Goyal N.**, **Chaudhary V.**, **Wenzek G.**, **Guzmán F.**, **Grave É.**, **Ott M.**, **Zettlemoyer L. and Stoyanov V.** (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8440–8451.

Dakota D. (2021). Genres, parsers, and BERT: The interaction between parsers and BERT models in cross-genre constituency parsing in English and Swedish. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, Online. Association for Computational Linguistics, pp. 59–71.

Das A. and Sarkar S. (2020). A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **19**(5), 1–60.

Das A., **Zaffar A. and Sarkar S.** (2017). Delexicalized transfer parsing for low-resource languages using transformed and combined treebanks. In *Proceedings of the CoNLL. 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 182–190.

de Marneffe M.-C., **Manning C. D.**, **Nivre J. and Zeman D.** (2021). Universal dependencies. *Computational Linguistics* **47**(2), 255–308.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN.

Dozat T. and Manning C. (2017). Deep biaffine attention for neural dependency parsing. In *5h International Conference on Learning Representations (ICLR)*, Toulon, France.

Dozat T., **Qi P. and Manning C. D.** (2017). Stanford's graph-based neural dependency parser at the CoNLL. 2017 shared task. In *Proceedings of the CoNLL. 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 20–30.

Duong L., **Cohn T.**, **Bird S. and Cook P.** (2015a). Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, Beijing, China, pp. 113–122.

Duong L., **Cohn T.**, **Bird S. and Cook P.** (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*, Beijing, China, pp. 845–850.

Falenska A. and Çetinoğlu Ö. (2017). Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proceedings of the 15th International Conference on Parsing Technologies*, Pisa, Italy, pp. 18–24.

Gorelova L. M. (2002). Manchu Grammar. Brill.

Gu S. (2016). A literature review on Sibe language. In *Manchu Studies*, vol. **2**, pp. 83–87 (In Chinese).

He J., **Zhang Z.**, **Berg-Kirkpatrick T. and Neubig G.** (2019). Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3211–3223.

Hwa R. (2001). On minimizing training corpus for parser acquisition. In *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL)*, Toulouse, France.

Hwa R., **Resnik P.**, **Weinberg A.**, **Cabezas C. and Kolak O.** (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering* **11**(3), 311–326.

Jang T. (2008). *Sibe Grammar*. Kunming, Yunnan, China: The Nationalities Publishing House of Yunnan (In Chinese).

Joshi V., **Peters M. and Hopkins M.** (2018). Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp. 1190–1199.

Kasai J. and Frank R. (2019). Jabberwocky parsing: Dependency parsing with lexical noise. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pp. 113–123.

Khan M., **Dickinson M. and Kübler S.** (2013). Towards domain adaptation for parsing web data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria, pp. 357–364.

Kondratyuk D. and Straka M. (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)*, Hong Kong, China, pp. 2779–2795.

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

Kudo T. and Richardson J. (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*, Brussels, Belgium, pp. 66–71.

Kullback S. and Leibler R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79–85.

Li S. (1979). A survey on the Sibe language. *Minority Languages of China* **6**(3), 221–232 (In Chinese).

Li S. (1982). Possession category in Sibe.*Minority Languages of China*, **6**, 50–57 (In Chinese).

Li S. (1984). Moods of indicative verbs in Sibe. *Minority Languages of China* **6**(6), 26–32 (In Chinese).

Li S. (1985). Adverbials in Sibe. *Minority Languages of China* **6**(5), 12–25 (In Chinese).

Li S. (1988). Auxiliaries in Sibe. *Minority Languages of China* **6**(6), 27–32 (In Chinese).

Lin Y.-H., **Chen C.-Y.**, **Lee J.**, **Li Z.**, **Zhang Y.**, **Xia M.**, **Rijhwani S.**, **He J.**, **Zhang Z.**, **Ma X.**, **Anastasopoulos A.**, **Littell P. and Neubig G.** (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3125–3135.

Littell P., **Mortensen D. R.**, **Lin K.**, **Kairis K.**, **Turner C. and Levin L.** (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 8–14.

Loper E., **Bird S. and Klein E.** (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Lynn T., **Foster J.**, **Dras M. and Tounsi L.** (2014). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, Dublin, Ireland, pp. 41–49.

McDonald R., **Petrov S. and Hall K.** (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, pp. 62–72.

Nivre J., **De Marneffe M.-C.**, **Ginter F.**, **Goldberg Y.**, **Hajic J.**, **Manning C. D.**, **McDonald R.**, **Petrov S.**, **Pyysalo S.**, **Silveira N. and et al.** (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, pp. 1659–1666.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C. D., Pyysalo S., Schuster S., Tyers F. and Zeman D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 4034–4043.

Norman J. (1974). A sketch of Sibe morphology. *Central Asiatic Journal* **18**(3), 159–174.

Office for the National Population Census (2022). *China Population Census Yearbook*. China Statistics Press (In Chinese).

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana: Association for Computational Linguistics, vol 1 (Long Papers), pp. 2227–2237.

Robbeets M. (2020). The classification of the Traneurasian languages. In Robbeets and Savelyev (2020), pp. 31–39.

Robbeets M. and Savelyev A. (eds), (2020). *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

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

Rosa R. and Žabokrtský Z. (2015a). KLcpos3 - A language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, pp. 243–249.

Rosa R. and Žabokrtský Z. (2015b). MSTParser model interpolation for multi-source delexicalized transfer. In *Proceedings of the 14th International Conference on Parsing Technologies*, Bilbao, Spain, pp. 71–75.

Rust P., Pfeiffer J., Vulić I., Ruder S. and Gurevych I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, pp. 3118–3135.

Rybatzki V. (2020). The Altaic languages: Tungusic, Mongolic, Turkic. In Robbeets and Savelyev (2020), pp. 22–28.

Santorini B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. 3rd Revision, 2nd Printing. Department of Computer and Information Science, University of Pennsylvania.

Sayyed Z. A. and Dakota D. (2021). Annotations matter: Leveraging multi-task learning to parse UD and SUD. In *Findings of the Association for Computational*. Linguistics, Online: ACL-IJCNLP, pp. 3467–3481.

Schiller A., Teufel S. and Thielen C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS, Universität Stuttgart and Universität Tübingen, Technical report.

Søgaard A. (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*, Portland, OR, pp. 682–686.

Täckström O., McDonald R. and Uszkoreit J. (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*, Montréal, Canada, pp. 477–487.

Tenney I., Das D. and Pavlick E. (2019a). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601.

Tenney I., Xia P., Chen B., Wang A., Poliak A., McCoy R. T., Kim N., Van Durme B., Bowman S. R., Das D., Pavlick E. (2019b). What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*

The Unicode Consortium (2022). The Unicode standard, version 15.0.0. *Available at:* https://www.unicode.org/versions/Unicode15.0.0/.

Tiedemann J. and Agić Z. (2016). Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* **55**, 209–248.

Tiedemann J., Agić Ž. and Nivre J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Ann Arbor, MI, pp. 130–140.

Vania C., Grivas A. and Lopez A. (2018). What do character-level models learn about morphology? The case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 2573–2583.

Šetuken (2009). *General Introduction to Xibe Grammar*. Urumqi, Xinjiang, China: Xinjiang People's Publishing House (In Chinese).

Whaley L. and Oskolskaya S. (2020). The classification of the Tungusic languages. In Robbeets and Savelyev (2020), pp. 81–91.

Wu S. and Dredze M. (2019). Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. 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)*, Hong Kong, China, pp. 833–844.

Wu S. and Dredze M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, Online, pp. 120–130.

https://doi.org/10.1017/nlp.2024.21 Published online by Cambridge University Press

Wu Y., **Schuster M.**, **Chen Z.**, **Le Q. V.**, **Norouzi M.**, **Macherey W.**, **Krikun M.**, **Cao Y.**, **Gao Q.**, **Macherey K.**, **et al.** (2016). Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144.

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

Yang Z., **Xu Z.**, **Cui Y.**, **Wang B.**, **Lin M.**, **Wu D. and Chen Z.** (2022). CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, pp. 3937–3949.

Yarowsky D., **Ngai G. and Wicentowski R.** (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, San Diego, CA.

Yeh A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics COLING*, Saarbrücken, Germany.

Zeman D., **Hajič J.**, **Popel M.**, **Potthast M.**, **Straka M.**, **Ginter F.**, **Nivre J. and Petrov S.** (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL. 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, pp. 1–21.

Zeman D. and Resnik P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Zhang Y., **Li Z. and Min Z.** (2020). Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of ACL*, pp. 3295–3305.

Zhou H., **Chung J.**, **Kübler S. and Tyers F.** (2020a). Universal Dependency treebank for Xibe. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, Barcelona, Spain (Online), pp. 205–215.

Zhou H. and Kübler S. (2021). Delexicalized cross-lingual dependency parsing for Xibe. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Online, pp. 1626–1635.

Zhou H., **Zhang Y.**, **Li Z. and Zhang M.** (2020b). Is POS tagging necessary or even helpful for neural dependency parsing?. In Zhu X., Zhang M., Hong Y. and He R., (eds), *Natural Language Processing and Chinese Computing*. Cham: Springer International Publishing, pp. 179–191.

Zikmundová V. (2013). *Spoken Sibe: Morphology of the Inflected Parts of Speech*. Charles University Prague: Karolinum Press.

Cite this article: Zhou H, Dakota D and Kübler S (2025). Cross-lingual dependency parsing for a language with a unique script. *Natural Language Processing* **31**, 277–305. https://doi.org/10.1017/nlp.2024.21