

Cross-Lingual Constituency Parsing for Middle High German: A Delexicalized Approach

Ercong Nie^{1,2} Helmut Schmid¹ Hinrich Schütze^{1,2}

¹Center for Information and Language Processing (CIS), LMU Munich, Germany

²Munich Center for Machine Learning (MCML), Germany

nie@cis.lmu.de

Abstract

Constituency parsing plays a fundamental role in advancing natural language processing (NLP) tasks. However, training an automatic syntactic analysis system for ancient languages solely relying on annotated parse data is a formidable task due to the inherent challenges in building treebanks for such languages. It demands extensive linguistic expertise, leading to a scarcity of available resources. To overcome this hurdle, cross-lingual transfer techniques which require minimal or even no annotated data for low-resource target languages offer a promising solution. In this study, we focus on building a constituency parser for **Middle High German (MHG)** under realistic conditions, where no annotated MHG treebank is available for training. In our approach, we leverage the linguistic continuity and structural similarity between MHG and **Modern German (MG)**, along with the abundance of MG treebank resources. Specifically, by employing the *delexicalization* method, we train a constituency parser on MG parse datasets and perform cross-lingual transfer to MHG parsing. Our delexicalized constituency parser demonstrates remarkable performance on the MHG test set, achieving an F1-score of 67.3%. It outperforms the best zero-shot cross-lingual¹ baseline by a margin of 28.6% points. These encouraging results underscore the practicality and potential for automatic syntactic analysis in other ancient languages that face similar challenges as MHG.

1 Introduction

Constituency parsing, which involves analyzing the grammatical structure of sentences and identifying the hierarchical relationships between words, plays a crucial role in linguistic research, especially for

the analysis of ancient languages that are no longer spoken. Its significance extends beyond linguistic analysis, serving as a building block for various natural language processing (NLP) applications, such as information extraction (Jiang, 2012; Jiang and Diesner, 2019), sentiment analysis (Li et al., 2020), question answering (Hermjakob, 2001), etc. However, ancient languages lack large labeled and unlabeled corpora (Assael et al., 2022) and treebanks suitable for parser training are seldom available. This scarcity of resources can be attributed to two reasons. Firstly, ancient languages usually have a dearth of digital text resources. Secondly, the construction of a treebank for an ancient language requires substantial linguistic expertise and manual effort. Nonetheless, the continuity in the process of language evolution gives rise to linguistic similarities between ancient languages and their corresponding modern counterparts (Parravicini and Pievani, 2018). Cross-lingual transfer techniques (Ruder, 2019; Lauscher et al., 2020) are trained on high-resource languages and require little or no annotated data from low-resource target languages. They can effectively be applied to languages with similar sentence structure and word order. Hence, they can be a viable solution to this challenge.

In this work, we focus on building a constituency parser for Middle High German (MHG). MHG is a historical stage of the German language that was spoken between 1050 and 1350. It is the linguistic predecessor of Modern German (MG). Both languages have many similarities in word formation and grammatical features, e.g., similar word order patterns and inflectional systems (Salmons, 2018). The availability of MHG parse trees is extremely limited. The *Deutsche Diachrone Baumbank (German Diachronical Treebank, DDB)* (Hirschmann and Linde, 2023) comprises merely around 100 manually annotated parse trees, encompassing less than 3000 tokens. These resources are far from

¹As is prevalent in the realm of multilingual NLP, the term “zero-shot cross-lingual” in this context pertains to a transfer learning method where we finetune the model with task-specific data in a source language and test on the target language directly (Sitaram et al., 2023).

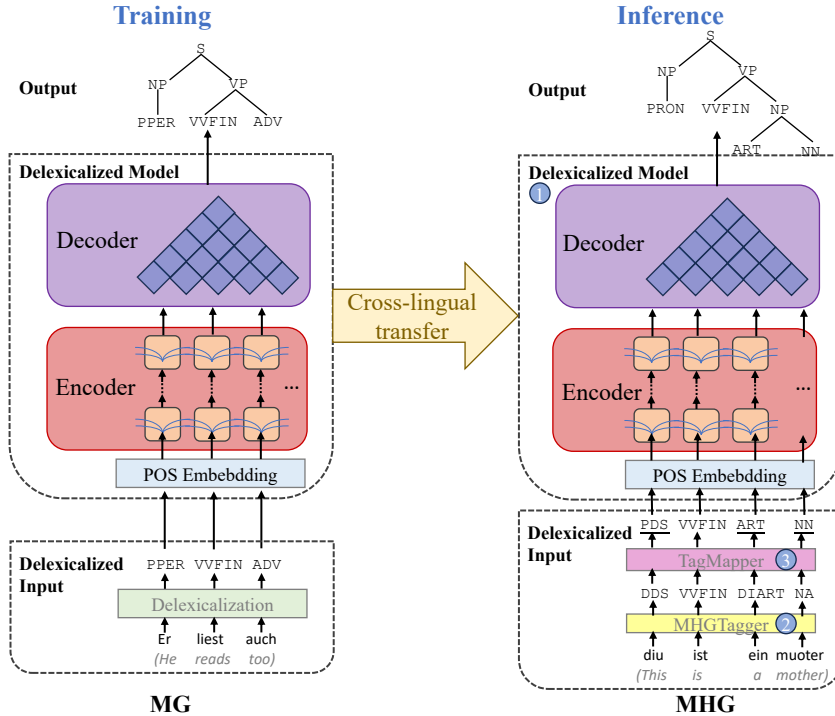


Figure 1: Overview of the cross-lingual delexicalized parsing system for MHG. In the training, the delexicalized parsing model is trained on the delexicalized MG trees. The trained parser is subsequently applied to MHG sentences. The delexicalized parsing system for MHG consists of three key modules: (1) *Delexicalized parsing model* trained on delexicalized MG trees, (2) *MHG POS tagger*, and (3) *Tag mapper*.

what is required to train an automatic syntactic analysis system, and are only suitable for use as test sets. On the other hand, there is an abundance of treebank resources available for MG, in particular the Tiger Treebank (Smith, 2003). Hence, we capitalize on the structural similarity between MHG and MG, as well as the rich MG treebank resources in order to develop a cross-lingual *delexicalized* constituency parsing model that we can directly apply to MHG sentences.

In the delexicalized approach, the parsing model operates on part-of-speech (POS) sequences rather than token sequences. We accomplish this by training a cross-lingual parser using POS sequences from high-resource source languages as input. Subsequently, we utilize this trained parser to directly parse POS sequences of low-resource target languages (McDonald et al., 2011).

In our work, we first train a delexicalized constituency parsing model on a delexicalized MG treebank. In order to parse MHG sentences with this model, we need to annotate them first with the POS tags used in the MG treebank. To this end, we train a POS tagger on an MHG corpus which has been manually annotated using a POS tag set similar, but not identical to the MG tag set. We em-

ploy a POS mapper to replace the MHG tags by the corresponding MG tags, ensuring the uniformity of the model’s inputs across the two languages, which is a prerequisite of the delexicalization method. The experimental results show that our delexicalized constituency parser substantially outperforms all other zero-shot cross-lingual parsing baselines, achieving an F1-score of 67.3% on the MHG parse test set.

The delexicalization method is particularly well-suited for languages which (1) lack treebank resources, (2) possess sufficient annotated data for training POS taggers, and (3) exhibit syntactic similarities with a high-resource language. Our investigation of this realistic scenario shows the feasibility of automatic syntactic analysis for an ancient language.

The subsequent sections of this paper are organized as follows. In Sec. 2, we discuss related work. Sec. 3 gives an overview of our research languages and the available corpora. The delexicalization method employed in our approach is detailed in Sec. 4. Sec. 5 describes our experimental setup, and in Sec. 6, we analyze the results. We conclude in Sec. 7.

2 Related Work

Cross-Lingual Transfer Learning The fundamental principle underlying cross-lingual transfer is that the processing of source and target languages uses a shared input representation, which can be either discrete or continuous. The delexicalization method is based on a shared discrete input representation, i.e., POS tags. Other discrete representation types include glossed words (Zeman and Resnik, 2008) and grounding texts in multilingual knowledge bases (Lehmann et al., 2015). Continuous cross-lingual representation spaces emerged with advancements in neural networks. Typical examples are cross-lingual word embeddings (Ammar et al., 2016) and sentence embeddings (Artetxe and Schwenk, 2019).

The emergence of massively multilingual transformers (Devlin et al., 2019; Conneau et al., 2020), which are jointly pretrained on multilingual corpora, introduces a novel pattern of zero-shot cross-lingual transfer learning. In this paradigm, a pretrained multilingual model is finetuned on a downstream NLP task dataset of a source language. The finetuned multilingual model is then directly applied to target language data for the same task (K et al., 2019; Hu et al., 2020; Liu et al., 2022; Nie et al., 2023).

Neural Constituency Parsing Recent advances in constituency parsing have witnessed a growing emphasis on harnessing neural network representations, making a shift from the previously prominent role of grammars, whose relevance has gradually diminished. Cross and Huang (2016) propose a span-based constituency parsing system specifically designed to leverage the powerful representation capabilities of the bidirectional long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). In this method, an input sentence is represented as a set of spans, and each span is assigned a score. The best-scoring parse tree is computed using dynamic programming techniques. They combine smaller spans into larger spans until the entire sentence is covered. Subsequently, several variations of the span-based method have been proposed, e.g. approaches replacing the inference algorithm with chart-based methods (Stern et al., 2017), using character-level representations instead of word-level representations (Gaddy et al., 2018), and replacing LSTMs with self-attention modules (Kitaev and Klein, 2018). Kitaev et al. (2019) take advan-

tage of the newly developed pretrained language models (PLMs) and use BERT (Devlin et al., 2019) to compute the span representations, resulting in enhanced performance. Kitaev and Klein (2020) improve the runtime complexity of constituency parsing to linear time by reducing parsing to tagging.

Cross-Lingual Constituency Parsing There has been relatively limited scholarly attention dedicated to cross-lingual constituency parsing in recent studies, especially for target languages situated in low-resource settings, such as MHG. Kitaev et al. (2019) have employed the multilingual BERT model to train a single parser with parameters shared across languages. They jointly finetune the multilingual BERT on 10 languages utilizing a common BERT backbone, but the model contains distinct MLP span classifiers for each language to accommodate the different tree labels. However, their approach necessitates the availability of treebanks of all the encompassed languages as training datasets. Kaing et al. (2021) undertake a comprehensive series of experiments to validate the efficacy of delexicalization techniques for zero-shot cross-lingual constituency parsing. Additionally, their study underscores significance of typological affinity in the source language selection. We build upon these investigations and apply their findings to the zero-shot parsing of MHG within a practical contextual framework.

Constituency Parsing on Historical German There is a notable scarcity of syntactically annotated corpora for historical German. In instances where annotated treebanks are absent, approaches such as rule-based, unsupervised, or zero-shot cross-lingual methods can be employed for constituency parsing. For instance, Chiarcos et al. (2018) have created a rule-based shallow parser for MHG. Recent advancements in the construction of such corpora encompass:

- *German Diachronical Treebank (DDB)*: a small yet syntactically deeply annotated corpus, comprising three subcorpora of different stages of German, i.e., Old High German, Middle High German and Early New High German (Hirschmann and Linde, 2023). The construction of the DDB corpus is oriented towards the Tiger Corpus (Smith, 2003), one of the largest German treebanks.
- *UP Treebank of Early New High German*

(*ENHG*): a syntactically annotated corpus of ENHG containing 21,432 sentences consisting of 600,569 word tokens based on the Reference Corpus of ENHG (Demske, 2019).

- *Corpus of Historical Low German (CHLG)*: a Penn-style treebank of Middle Low German (Booth et al., 2020)

Contemporary work on historical German parsing based on previously mentioned corpora includes endeavors such as cross-dialectal parsing for ENHG based on CHLG (Sapp et al., 2023).

3 Languages and Corpora

The ancient language which we study in this paper is Middle High German (MHG). MHG and Modern German (MG) are stages of the same Germanic language family, representing different historical periods. MHG emerged during the Middle Ages in the German-speaking regions of Central Europe. It was primarily used in literary and administrative contexts and played an important role in medieval literature, including epic poems such as the *Nibelungenlied* and *Minnesang* (courtly love poetry) (Salmons, 2018).

Linguistic Considerations of MHG MHG has a phonetic system that included a set of vowel and consonant sounds. The pronunciation and sound patterns differ from those of MG, but some MHG words are still recognizable in MG. MHG has a more complex grammatical system, such as a more extensive case system with different noun and adjective declensions. Besides, verb conjugation has more intricate forms and patterns (Jones and Jones, 2019). In terms of orthography, the spelling and writing conventions of MHG are different from MG. For example, *ü*, the umlaut of *u*, is usually written *iu* in MHG. The transition from MHG to MG was a gradual process, occurring over several centuries. MG can be considered the linguistic descendant of MHG, with linguistic changes and developments shaping the language over time.

MHG Corpora Resources During the MHG period, the amount of textual material that survives to the present increases markedly. The *Reference Corpus of Middle High German (ReM)* (Klein et al., 2016) encompasses a large collection of non-literary and non-religious texts. ReM is a corpus of diplomatically transcribed and annotated texts of MHG with a size of around 2 million word

forms. Texts in ReM have been digitized and richly annotated, e.g., with POS, morphological and lemma features. The morphological annotation uses the HiTS tag set (Dipper et al., 2013), a tag set for historical German, derived from the Stuttgart-Tübingen Tag Set (STTS) for modern German texts (Schiller et al., 1995). Although the ReM corpus provides rich morphologically annotated text data for MHG, the availability of syntactically annotated data for MHG is severely limited, with only approximately 100 MHG parse trees included in the DDB treebank. In contrast, the treebank resources for MG are abundant. The Tiger Treebank (Brants et al., 2002), for instance, consists of approximately 40,000 sentences of German newspaper text, taken from the *Frankfurter Rundschau*.

4 Methods

In our work, we focus on developing a constituency parser for MHG. In the previous section, we reviewed annotated resources available for MHG and MG. Basically, we have ample treebank resources for MG and plenty of POS-tagged texts for MHG, whereas the treebank resources for MHG are extremely limited. Given the resource availability for MG and MHG along with the linguistic connection between the two languages, employing a cross-lingual constituency parsing approach utilizing delexicalization proves to be an effective solution. As Figure 1 shows, the delexicalized model is trained on the delexicalized inputs of MG. In the inference stage, the delexicalized parser is directly applied to MHG POS sequences. The delexicalization method requires that MHG and MG share the same set of POS tags. The final constituency parser for MHG (the right side of Figure 1 comprises three modules: (1) the delexicalized parser, (2) the MHG POS tagger, and (3) the POS mapper from MHG to MG. In the next section, we describe the delexicalized parsing system in more detail.

4.1 Delexicalized Parser

Our delexicalized MHG parser is based on the Berkeley neural parser (Benepar) (Kitaev and Klein, 2018), a span-based parser using self-attention. As illustrated in Figure 1, Benepar has an encoder-decoder architecture which combines a chart decoder with a sentence encoder based on self-attention. The sentence encoder computes contextualized representations for all word positions and combines them to form span representations.

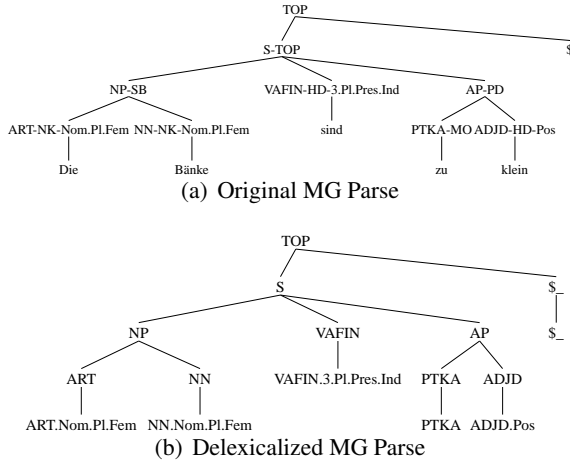


Figure 2: An example illustrating the delexicalization process of a MG tree.

From the span representations, the parser computes label scores, which are subsequently used to incrementally construct a tree using a chart parsing algorithm (Sakai, 1961).

According to Kaing et al. (2021), Benepar exhibits two key features which are advantageous for cross-lingual transfer. Firstly, it employs a self-attentive encoder that effectively captures global context information and exhibits less sensitivity to word order. Secondly, the parser independently scores each span without considering the label decisions of its children or parent. This means that a failure in label prediction for a certain span does not strongly impact the label prediction for other spans (Gaddy et al., 2018). Consequently, the prediction errors resulting from local syntax variations between two languages have a limited effect on the overall prediction.

While our delexicalized parser adopts the same architecture to Benepar, there exist distinctions in the inputs of the two. Specifically, Benepar is trained on parse trees with words, whereas our delexicalized parser operates on POS sequences as inputs, i.e. tree strings devoid of words. Therefore, the delexicalized version of the MG treebank is required to train the delexicalized parser. For the MHG parsing in the inference, we feed the delexicalized model with the POS sequences of MHG sentences.

4.2 Delexicalization for MG and MHG

Delexicalization for MG We use the Tiger Treebank to train the delexicalized parsing model on MG parse trees. The parse trees in the Tiger Tree-

bank contain additional semantic information, such as edge labels, and special structures, such as coreference indices and trace nodes. We remove all of them during delexicalization.

In the Tiger treebank, the label of each preterminal node contains not only the POS tag, but also morphological features, such as case, number, gender. During delexicalization, we overwrite the word at the leaf node with this extended POS tag, but only keep the POS information in the label of the preterminal node. This means that the input of our delexicalized parser contains information about morphological features. Figure 2 shows an example of the delexicalization for a MG sentence. As shown the edge labels, e.g., “NK” are removed and the tokens are replaced by the POS tag combined with morphological features, e.g., “ART.Nom.Pl.Fem”, where “ART” (determiner) is the POS tag, and “Nom.Pl.Fem” denotes the morphological information with case being nominative, number being plural, and gender being feminine.

MHG POS Tagger For the delexicalization of MHG sentences, we need a POS tagger for MHG. We use the RNNTagger of Schmid (2019) for this purpose, which annotates MHG sentences with POS tags as well as morphological features and has been trained on the ReM corpus. RNNTagger uses deep bidirectional LSTMs with character-based word representations.

4.3 Tag Set Mapping

The Tiger Treebank uses the STTS tag set, whereas the MHG version of the RNNTagger and the ReM corpus on which it was trained employ the HiTS tag set. Due to this discrepancy, we cannot directly use the POS labels from RNNTagger as input to the delexicalized parser. HiTS, for example, has separate tags for definite (DDART) and indefinite articles (DIART), whereas STTS uses the tag “ART” for both of them. Since the delexicalization method demands that the source and target languages share the same tag set, we have to map the MHG tags to the MG. The small MHG treebank that we use for evaluation purposes uses STTS and requires no mapping.

The mapping process involves two dimensions. Firstly, we map the morphological features of MHG to those of MG. Secondly, we map the POS tags of MHG to those of MG primarily based on a mapping dictionary. Table 2 shows a selected part

	Type	Language	Size	Usage
Tiger	Treebank	MG	50,474 trees	Parser training
DDB	Treebank	MHG	96 trees	Parser evaluation
ReM	POS-tagged corpus	MHG	2,269,738 tokens	POS tagger training

Table 1: Overview of the datasets.

MHG Tag	MG Tag
CARDD	CARD
DDA	PDAT
DDART	ART
DIA	PIAT
DIART	ART
DID	PDAT
NA	NN
VAPS	ADJD.Pos

Table 2: Representative mapping pairs in the mapping dictionary.

of the POS tag mapping dictionary. It should be noted that our mapping is not flawless due to certain challenges. For instance, the composite word in MHG “*enerde (on earth)*” is separated into “*auf*” and “*Erde*” in MG and are tagged as “`APPR|NA`”. In the DDB treebank, such composite words are annotated with two separate tags combined with “`I`” in the DDB treebank. However, for simplification purposes, our mapping only retains the first part of the tag, leading to a loss of information.

5 Experiments

We begin by training Benepar on the delexicalized Tiger treebank for MG. Then we annotate the sentences of the small DDB treebank for MHG with RNNTagger and map the HiTS tags that it returns to STTS tags. Finally, we parse the POS tag sequences with the trained parser.

5.1 Datasets

In our experiments, we utilize the following three corpora (see also Table 1).

Tiger Treebank The delexicalized parser is trained on the Tiger Treebank (Smith, 2003), which comprises a total number of 50,474 parse trees for MG. We use a version of the Tiger Treebank which has been converted to the Penn Treebank format (Marcus et al., 1993). We delexicalize the Tiger corpus and divide it into a training set and a development set. The first 47,474 parse trees in the Tiger corpus comprise the training set and the last 3,000 parse trees comprise the development set.

DDB The German Diachronic Treebank (DDB) (Hirschmann and Linde, 2023) consists of a limited number of 100 parse trees for MHG. Due to the small data size, we utilize the DDB treebank solely for the cross-lingual evaluation of the delexicalized parser. To prepare the DDB treebank for evaluation, we perform preprocessing steps, including converting it to the format of the Penn Treebank and removing incomplete parse trees and parse trees with mostly Latin words. We also removed numbers and periods which formed the first token of a parse tree and corrected a few more minor problems. At the end, we had 96 sentences for evaluation purposes.

ReM The Reference Corpus for Middle High German (ReM) (Klein et al., 2016) is an extensive collection of texts written in MHG. This corpus encompasses approximately 2.3 million tokens and provides comprehensive linguistic annotations, including POS tags, morphological analysis, lemma features, and more. The ReM corpus has been used by Schmid (2019) to train the MHG version of his RNNTagger which annotates MHG texts with POS tags and morphological features.

5.2 Baselines

We evaluate the performance of our proposed delexicalized MHG parser which is based on the Benepar parser (Kitaev and Klein, 2018), and compare it with the cross-lingual transfer performance of the original Benepar without using the delexicalization method and other parsing approaches that incorporate pretrained language models, which have shown promising results in various NLP tasks.

Vanilla Benepar The vanilla Benepar model is trained directly on the original training set of the Tiger Treebank for MG without delexicalization. After training, the parser is directly used to parse the MHG sentences as token sequences. This allows us to compare the performance of the delexicalized MHG parser with the vanilla Benepar model, highlighting the impact of delexicalization on cross-lingual parsing performance.

	Recall		Precision		FScore		CM	
	MG	MHG	MG	MHG	MG	MHG	MG	MHG
<i>Baselines</i>								
Vanilla Benepar	84.18	34.41	87.57	44.40	85.84	38.77	45.80	0.00
Tetra-gBERT	86.31	23.20	88.19	29.53	87.24	25.98	51.70	3.12
Tetra-mBERT	60.68	19.69	65.61	23.25	63.15	21.32	21.35	0.00
<i>Our proposed method</i>								
Dexparser	81.39	64.72	84.89	70.19	83.10	67.34	39.03	12.50

Table 3: Main results of the cross-lingual parsing transfer performance of different parsers. **CM** refers to “complete match”. gBERT refers to the pretrained German BERT and mBERT refers to the multilingual version BERT. The best value of each column is indicated in **bold**.

Tetra-Tagging with PLMs Tetra-tagging (Kitaev and Klein, 2020) is a technique for reducing constituency parsing to sequence labeling. In this approach, special parsing tags are predicted in parallel using a PLM, and then merged into a parse tree. In our experiment, we use the pretrained German BERT model (Chan et al., 2020) and the multilingual BERT model (Devlin et al., 2019) available on the HuggingFace website (Wolf et al., 2020). We start by finetuning these models on the Tiger Treebank using the Tetra-tagging technique. Subsequently, we evaluate their performance on the MHG parse test set.

5.3 Evaluation

Following Kitaev and Klein (2018), we use the the standard `evalb` measures (Sekine and Collins, 1997; Collins, 1997) for the parser quality evaluation. `evalb` is a software tool that provides metrics to assess the accuracy and similarity of parsed sentences against reference or gold standard parse trees, including precision, recall, F1 score, and complete match.

- **Precision** measures the proportion of predicted constituents in the generated parse tree which are also contained in the reference parse tree. It quantifies the accuracy of the parser in correctly identifying constituents.
- **Recall** measures the proportion of constituents in the reference parse tree which were predicted by the parser in the generated parse tree. It quantifies the parser’s ability to generate all the constituents present in the reference parse tree.
- **F1 Score** is the harmonic mean of precision and recall.
- **Complete Match** measures the proportion of

predicted parse trees which were exactly identical to the respective reference parse trees.

As is the standard practice, the evaluation disregards POS labels and punctuation.

5.4 Training Setup

For training the delexicalized parser, we adopt the same hyperparameter settings as described in (Kitaev and Klein, 2018). The encoder architecture consists of a character-level bidirectional LSTM neural network. We configure the encoder with a dimension of 1024, utilizing 8 layers, 8 attention heads, and a dimension of 64 for the key, query, and value. The size of the feedforward layer is set to 2048, and the character embedding dimension is 64. The batch size is set to 32, the learning rate is 5e-5, and the maximum sequence length of the encoder is 512. We use the random seed 10 for training. We conduct all our experiments using a server with 8 GPUs with 11GB RAM (NVIDIA GeForce GTX 1080 Ti).

6 Results and Analysis

6.1 Main Results

Table 3 shows the parsing performance of different cross-lingual parsers. Notably, our proposed parser attains the highest scores across all metrics for MHG, demonstrating that the delexicalized parser possesses superior cross-lingual parsing performance on MHG. Our delexicalized parser demonstrates substantial advantages in parsing MHG, achieving an impressive increase of almost 30% points in F1 score. Besides, it achieves comparable results on MG. In terms of the baselines, the Vanilla Benepar and the Tetra-gBERT parser both achieve relatively high recall and precision for MG but have noticeably lower values for MHG. The Tetra-mBERT parser exhibits lower values for both recall and precision for both MG and MHG. It is

	Recall	Precision	FScore	CM
Delexicalized parser using gold tags	66.18	71.17	68.59	14.58
- using predicted tags	64.72	70.19	67.34	12.50
- without mapping	59.16	68.82	63.63	7.29
- without morphological information	48.66	65.38	55.8	9.28

Table 4: The MHG parsing results with delexicalized parser in the ablation study.

worth noting that the parsing performance of the delexicalized model on the source language MG is surpassed by the two strong baselines, Vanilla Benepar and Tetra-gBERT. This outcome is expected as the delexicalization process diminishes the semantic information present in the input sequences. However, the trade-off of the performance loss in MG leads to a big leap in the cross-lingual parsing performance for MHG.

Our delexicalized constituency parser exhibits outstanding performance on the MHG test set, attaining an impressive F1-score of 67.3%. This substantial improvement outperforms the best zero-shot cross-lingual baseline by a considerable margin of 28.6%. Although there is a slight decline in the parsing performance for MG, the trade-off proves worthwhile considering the substantial gains achieved in parsing MHG. This emphasizes the effectiveness of the delexicalized approach in facilitating cross-lingual transfer and highlights its potential for parsing ancient and historical languages like MHG.

6.2 Ablation Study

We now examine how the parsing performance changes (i) as we replace predicted POS tags with goldstandard POS tags, (ii) as we use the original HiTS tags instead of mapping them to STTS tags, and (iii) as we remove the morphological features from the parser input. Table 4 presents the results of our ablation study.

Goldstandard POS Tags We observe that the f-score of the delexicalized parser increases by 1.3% points when it processes gold standard POS tag sequences instead of POS tag sequences predicted by RNNTagger. This finding underscores the quality of the POS tags predicted by RNNTagger. We lose very little performance due to POS tagging errors.

Tag Set Mapping Table 4 demonstrates a noticeable decline in parsing performance from 67.34% to 43.43% in terms of F1 score when the delexicalized MHG sequences are directly processed

by the cross-lingual parser without mapping them from HiTS to STTS. This finding highlights the indispensability of mapping from MHG to MG for maintaining satisfactory parsing performance. The results underscore the significance of aligning the tag sets between MHG and MG to ensure effective cross-lingual parsing and emphasize the necessity of this mapping process in our approach.

Morphological Information The inclusion of morphological markers provides the neural model with valuable additional information for parsing MHG sentences. In our experiments, we augment the delexicalized MHG sequences with morphological information, such as case, gender, number, and more. The outcomes of the ablation study clearly indicate that removing this morphological information from the delexicalized input sequences obviously impairs parsing performance. Specifically, this exclusion leads to a noticeable decline in the F1 score, amounting to a reduction of 11.5%.

6.3 Case Study

Figure 3 shows two MHG trees generated by our delexicalized parser and the corresponding gold standard trees for comparison. This case study reveals that the delexicalized parser demonstrates relatively accurate predictions of constituents when compared to the reference trees, especially for short MHG sentences. Some prediction errors in constituents stem from the intricacy and the ambiguity of the MHG grammar, as exemplified by the case of “her” in Example 2. From a linguistic perspective, determining whether “her” functions as an adverb (ADV) or a separated verb prefix (PTKVZ) poses challenges. However, in longer and more complex sentences, e.g., the sentence in Example 1, the parser typically maintains a high level of accuracy locally while occasionally struggling to accurately determine the overall structure of the entire sentence. Besides, the presence of noise in the ancient texts is another factor that can impact the effectiveness of the cross-lingual parsing for MHG. Overall, the qualitative analysis provides further evidence of the effectiveness of the delexicalized parser for

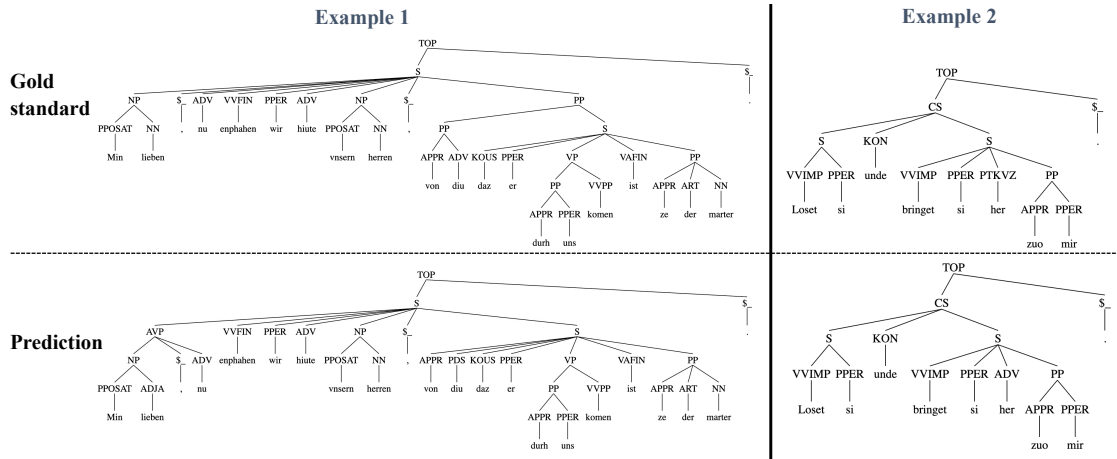


Figure 3: Two examples of the trees generated by our delexicalized parser compared to the reference parses.

MHG, emphasizing its ability to accurately predict constituents, especially in shorter sentences. While challenges may arise in handling longer and more complex sentences, the delexicalized parser showcases promising results, contributing to the advancement of MHG parsing.

7 Conclusion

In conclusion, our study presents an effective cross-lingual constituency parsing approach for ancient languages, specifically focusing on the parsing of Middle High German (MHG) sentences. Through the utilization of delexicalization and the similarities between MHG and Modern German (MG), we have developed a delexicalized parser based on the rich treebank resources of MG, which demonstrates remarkable performance in parsing MHG sentences. Our experimental results showcase the efficacy of the delexicalized approach, outperforming existing baselines and achieving substantial improvements in parsing accuracy. These findings highlight the practicality and promise of our approach for parsing historical and ancient languages, addressing the challenges posed by limited annotated data and linguistic variations.

Limitations

One limitation of our study is the need for further improvement in the robustness of the delexicalized parsing method, particularly when applied to ancient texts. By addressing this limitation, we can further enhance the applicability of our approach to a wider range of ancient languages and ensure more reliable parsing results. Besides, our proposed method is only applicable to the scenario

where a POS tagger for the target language and a related language with a treebank exist.

Acknowledgements

We extend our sincere gratitude to the anonymous reviewers for their invaluable contributions and constructive feedback that have greatly enriched the quality and scope of this paper. This work was supported by China Scholarship Council (CSC).

References

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. [Many languages, one parser](#). *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.
- Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. [A Penn-style treebank of Middle Low German](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 766–775, Marseille, France. European Language Resources Association.
- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168, pages 24–41.

- Branden Chan, Stefan Schweter, and Timo Möller. 2020. **German’s next language model**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Christian Chiarcos, Benjamin Kosmehl, Christian Fäth, and Maria Sukhareva. 2018. **Analyzing Middle High German syntax with RDF and SPARQL**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Michael Collins. 1997. **Three generative, lexicalised models for statistical parsing**. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Madrid, Spain. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. **Unsupervised cross-lingual representation learning at scale**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- James Cross and Liang Huang. 2016. **Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1–11, Austin, Texas. Association for Computational Linguistics.
- Ulrike Demske. 2019. Referenzkorpus frühneuhochdeutsch: Baumbank. *UP. Universität Potsdam: Institut für Germanistik (<https://hdl.handle.net/11022/0000-0007-EAF7-B>)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. **Hits: ein tagset für historische sprachstufen des deutschen**. *Journal for Language Technology and Computational Linguistics*, 28(1):85–137.
- David Gaddy, Mitchell Stern, and Dan Klein. 2018. **What’s going on in neural constituency parsers? an analysis**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 999–1010, New Orleans, Louisiana. Association for Computational Linguistics.
- Ulf Hermjakob. 2001. **Parsing and question classification for question answering**. In *Proceedings of the ACL 2001 Workshop on Open-Domain Question Answering*.
- Hagen Hirschmann and Sonja Linde. 2023. **Deutsche diachrone baumbank (version 1.0)**. Humboldt-Universität zu Berlin.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. **XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation**. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Jing Jiang. 2012. Information extraction from text. *Mining text data*, pages 11–41.
- Ming Jiang and Jana Diesner. 2019. **A constituency parsing tree based method for relation extraction from abstracts of scholarly publications**. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 186–191, Hong Kong. Association for Computational Linguistics.
- Howard Jones and Martin H Jones. 2019. *The Oxford Guide to Middle High German*. Oxford University Press.
- Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. *ArXiv*, abs/1912.07840.
- Hour Kaing, Chenchen Ding, Masao Utiyama, Eiichiro Sumita, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Constituency parsing by cross-lingual delexicalization. *IEEE Access*, 9:141571–141578.
- Nikita Kitaev, Steven Cao, and Dan Klein. 2019. **Multilingual constituency parsing with self-attention and pre-training**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.
- Nikita Kitaev and Dan Klein. 2018. **Constituency parsing with a self-attentive encoder**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

- Nikita Kitaev and Dan Klein. 2020. [Tetra-tagging: Word-synchronous parsing with linear-time inference](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6255–6261, Online. Association for Computational Linguistics.
- Thomas Klein, Klaus-Peter Wegera, Stefanie Dipper, and Claudia Wich-Reif. 2016. [Reference Corpus of Middle High German \(1050–1350\) \(Version 1.0\)](#). Rheinische Friedrich-Wilhelms-Universität Bonn, Ruhr-Universität Bochum.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Yuncong Li, Cunxiang Yin, and Sheng-hua Zhong. 2020. Sentence constituent-aware aspect-category sentiment analysis with graph attention networks. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 815–827. Springer.
- Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2022. [MulZDG: Multilingual code-switching framework for zero-shot dialogue generation](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 648–659, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. [Multi-source transfer of delexicalized dependency parsers](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. [Cross-lingual retrieval augmented prompt for low-resource languages](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, Toronto, Canada. Association for Computational Linguistics.
- Andrea Parravicini and Telmo Pievani. 2018. Continuity and discontinuity in human language evolution: putting an old-fashioned debate in its historical perspective. *Topoi*, 37(2):279–287.
- Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.
- Itiroo Sakai. 1961. [Syntax in universal translation](#). In *Proceedings of the International Conference on Machine Translation and Applied Language Analysis*, National Physical Laboratory, Teddington, UK.
- Joseph Salmons. 2018. *A history of German: What the past reveals about today’s language*. Oxford University Press.
- Christopher Sapp, Daniel Dakota, and Elliott Evans. 2023. [Parsing early New High German: Benefits and limitations of cross-dialectal training](#). In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 54–66, Washington, D.C. Association for Computational Linguistics.
- Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das tagging deutscher textcorpora mit stts. *Universität Stuttgart, Universität Tübingen, Germany*.
- Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, pages 133–137.
- Satoshi Sekine and Michael Collins. 1997. Evalb bracket scoring program. URL: <http://www.cs.nyu.edu/cs/projects/proteus/evalb>.
- Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. [Everything you need to know about multilingual LLMs: Towards fair, performant and reliable models for languages of the world](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 21–26, Toronto, Canada. Association for Computational Linguistics.
- George Smith. 2003. A brief introduction to the tiger treebank, version 1. Technical report, Universität Potsdam.
- Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. [A minimal span-based neural constituency parser](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. [Cross-language parser adaptation between related languages](#). In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.