Parsing Early New High German: Benefits and limitations of cross-dialectal training

Christopher Sapp
Department of Germanic Studies
Indiana University
csapp@iu.edu

Daniel Dakota
Department of Linguistics
Indiana University
ddakota@iu.edu

Elliott Evans
Department of Germanic Studies
Indiana University
evansell@iu.edu

Abstract

Historical treebanking within the generative framework has gained in popularity. However, there are still many languages and historical periods yet to be represented. For German, a constituency treebank exists for historical Low German, but not Early New High German. We begin to fill this gap by presenting our initial work on the Parsed Corpus of Early New High German (PCENHG). We present the methodological considerations and workflow for the treebank's annotations and development. Given the limited amount of currently available PCENHG treebank data, we treat it as a low-resource language and leverage a larger, closely related variety—Middle Low German—to build a parser to help facilitate faster post-annotation correction. We present an analysis on annotation speeds and conclude with a small pilot use-case, highlighting potential for future linguistic analyses. In doing so we highlight the value of the treebank's development for historical linguistic analysis and demonstrate the benefits and challenges of developing a parser using two closely related historical Germanic varieties.

1 Introduction

The most common system for historical treebanks in the generative framework is the Penn family of parsed historical corpora. These are valuable resources for analyzing syntactic change and have resulted in an explosion of research in this area, including the annual Diachronic Generative Syntax Conference and *Journal of Historical Syntax*. The Germanic family is well-represented (see section 3.1), with the exception of High German (HG). Our broader research agenda seeks to fill this gap by creating a parsed corpus of Early New High German (ENHG; 1350-1650).

Although there is no Penn-style treebank for any stage of HG on which to train a parser[1], there does exist the Corpus of Historical Low German (CHLG; Booth et al., 2020), which we can use as a starting point. The Low German (LG) language subsumes northern dialects that preserve proto-Germanic **p, *t,* and **k*, while HG varieties partially or fully reflect the HG consonant shift (*p>pf, t>ts, k>x,* etc.) and include all central and southern dialects and Modern Standard German. Despite these phonological differences and the characterization of LG and HG as separate languages, they are highly similar in lexis and syntax (Salveit, 1970; Rösler, 1997), although this syntactic similarity is questioned by Booth et al. (2020).

We introduce the first stages of the Parsed Corpus of Early New High German, along with the current workflow for developing the treebank and the supporting rationale for our chosen annotation and methodology. We explore a strategy of training a parser on historical texts from the CHLG treebank to help facilitate and aid in creating an ENHG treebank. In addition to initial parsing experiments to provide basic insights into the effectiveness of a cross-variety parser, we perform a small pilot case study to highlight potential linguistic challenges and use-cases for the treebank.

2 Related Work

2.1 Historical Treebanking

The Penn system for historical corpora is refined and expanded from the Penn Treebank (Marcus et al., 1993). This constituency-based annotation captures both linear and hierarchical relations between words and allows a variety of complex syntactic configurations to be queried. There exist Penn-style historical corpora for several Germanic languages: three large corpora for historical English (Kroch, 2020; Taylor et al., 2003b, 2006), Icelandic Parsed Historical Corpus (Wallenberg et al., 2011), Penn Parsed Corpus of Historical Yiddish (Santorini, 2021), and CHLG, but not yet for any

[1] Neither Tiger (Brants et al., 2004) nor TüBa-DZ (Telljohann et al., 2015) are annotated with a PTB style framework.

Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 54 - 66

stage of High German.

Penn-style historical corpora are produced by an iterative process of automatic annotation and manual correction (Taylor et al., 2003a). If texts are already POS tagged, a typical parsing workflow is outlined by Booth et al. (2020): 1) basic/shallow rule-based parsing, 2) manual correction, embedding clauses and inserting empty categories, 3) rule-based validation and flagging of errors, and 4) manual correction of flagged errors. Manual correction is especially vital because medieval texts are not standardized, and researchers in diachronic syntax expect to query sentences that are accurately parsed.

2.2 Annotation Development

Each historical corpus in the Penn family slightly adapts the tagset of the Penn Parsed Corpora of Historical English (Kroch, 2020), either for language-specific reasons or to resolve inconsistencies in the tagsets of prior corpora. CHLG departs significantly from this (Booth et al., 2020): although it uses Penn-type tags for higher syntactic nodes, the POS tags are a variant of the the Stuttgart-Tübingen Tagset (STTS; Schiller et al., 1995, 1999). In CHLG, each terminal node is split into meta information and the wordform:

(1) *grotem*

```
(ADJA (META (CASE dat) (GEND neut)
            (LEMMA grōt) (NUM sg))
      (ORTHO grotem))
```

Our syntactic labels are largely as in CHLG, but for the heads, we were faced with the choice to adapt one of the Penn tagsets to historical German (making our corpus easily searchable by the diachronic generative syntax community) or keep the tagset of our source texts (making the corpus most similar to CHLG). We have chosen to use a modified form of the Penn tagset, because a) the STTS encodes some basic syntactic information, resulting in redundancy with higher constituents (Booth et al., 2020), b) researchers most likely to use our corpus are more familiar with Penn-type annotations, and c) most Penn corpora and many others (e.g. the SPMRL shared task (Seddah et al., 2013, 2014)) attach morphological information to the POS tag. Following HeliPaD (Walkden, 2016), we attach morphology and lemma to the POS tag and terminal, respectively:

(2) *grossem*

```
(ADJ^D^SG grossem=groß)
```

2.3 Historical Parsing

Some work exists on automatic syntactic analysis of German historical texts. Koleva et al. (2017) perform experiments with both a memory-based learning approach and a CRF model for POS tagging Middle Low German; a single mixed cross-genre, cross-city model yields the best results.

Ortmann (2020) shows that topological field identification models derived form modern German do not show good performance when applied to Early New High German, as the often extremely long sentences in ENHG are problematic. Follow up work in chunking (Ortmann, 2021b) and automatic phrase recognition (Ortmann, 2021a) yield similar findings, with increased sentence length causing additional errors, but including historical data in the training helps performance.

Full constituency parsing of Modern British English is performed by Kulick et al. (2014), obtaining results similar to that of the Penn Treebank. Kulick et al. (2022) develop the first parser for Early Modern English (1700-1914), noting that experiments using in-domain embeddings outperform those trained on Modern English.

Perhaps the most directly related work to ours is that of Arnardóttir and Ingason (2020), who build a single neural parsing pipeline for the Icelandic Parsed Historical Corpus. While achieving good performance when using a mix of data in the train, development, and test sets, they noted that performance drops when parsers were trained and tested on different time periods, with modern data showing more performance loss on older data than vice versa. One notable decision was the conversion of all historical texts to modern Icelandic spelling. We do not perform any such normalization and expect a large amount of dialectal and diachronic variation, but note that parsers have shown to be surprisingly adaptable to errors and inconsistencies in historical texts (Kulick et al., 2022).

3 Methodology

3.1 Treebanks

Historical treebanks are used to investigate changes that would be difficult to detect in a corpus that is only morphologically tagged. Treebanks in the Penn family can be analyzed using CorpusSearch 2 (CS2; Randall et al., 2004), a program whose query language is intuitive to generative syntacticians (e.g. `CP-SUB* dominates NP-OB1` returns direct objects in subordinate clauses).

Corpus of Historical Low German The CHLG treebank contains 20 Middle Low German (MLG) texts from 1279-1580, resulting in over 170,000 words. Phrase/clause labels are adapted from Kroch (2020). The tagset for terminal nodes is the Historisches Niederdeutsch-Tagset (HiNTS; Barteld et al., 2018), a variant of the Stuttgart-Tübingen Tagset (STTS; Schiller et al., 1995, 1999) adapted for historical Low German, see (1).

Parsed Corpus of Early New High German (PCENHG): currently consists of 5 texts with approx. 39,000 words.[2] Ultimately, this will be a structured corpus, aiming for one text from each of 10-12 regions for each 50-year time period between 1350 and 1650 (64 texts, approx. 600,000 words). Texts are adapted from the Referenzkorpus Frühneuhochdeutsch (ReF; Wegera et al., 2021); the texts come divided into sentences and POS-tagged using the Historisches Tagset (HiTS; Dipper et al., 2013), similar to the tagset of CHLG.

We selected three texts to be the first parsed and manually corrected texts for the PCENHG:

- *Neues Buch Köln*: chronicle of the city of Cologne from about 1360; Ripuarian dialect; 189 sentences = 10,027 words
- *Fierrabras*: fiction from 1533; Moselle Franconian; 401 sentences = 10,274 words
- *Wahrhaftig Historia*: Hans Staden's 1557 travel narrative; Rhine Franconian; currently 269 sentences = 4,251 words

These were chosen because they fall within the timespan of the CHLG and are from the north-west of the ENHG area, thus assumed to be lexically and grammatically closer than more southerly texts to the texts of CHLG. The three texts are Middle German, a dialect group of HG that retains some consonants of LG to varying degrees on a roughly north-south continuum. The dialect of Cologne (*Neues Buch*) shares the most features with LG, with fewer LG features in the Moselle Franconian *Fierrabras* and the fewest LG features in Rhine Franconian *Wahrhaftig Historia*. The locations of the texts vis-a-vis LG are illustrated in Figure 1.[3]

[2]The corpus can be found at https://ipchg.iu.edu

[3]Map adapted from Wiesinger et al.; labels are our own.

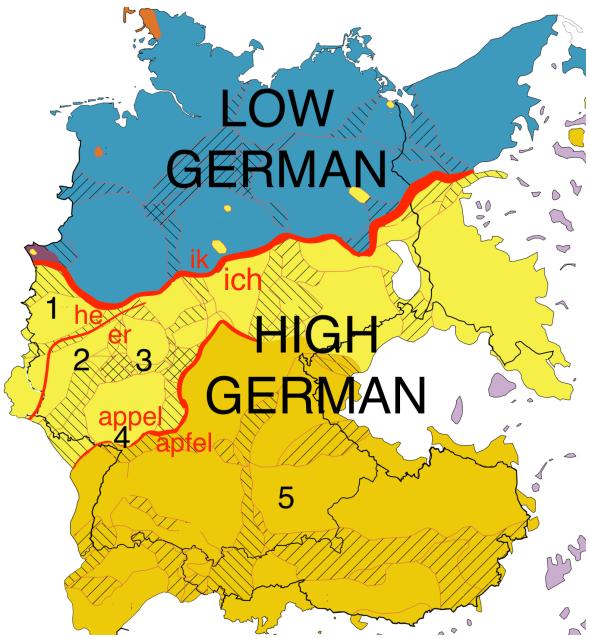

Figure 1: LG and HG. Texts in this study: 1=*Neues Buch*, 2=*Fierrabras*, 3=*Historia*, 4=*Karrenritter*, 5=*Geistliche Mai*

Trebank	Train	Dev	Test	Total
CHLG	9 997	999	833	11 829
Neues Buch	100	50	39	189
Fierrabras	200	101	100	401
Historia	140	68	60	268

Table 1: Treebank statistics for currently available gold annotated sentences with the train, development, and test splits.

3.2 Parsers

One unknown is whether a particular parser may be optimal for our workflow of post-correction, as different parsing strategies may produce different results given textual characteristics. We choose to perform preliminary experiments with two parsers that have yielded state-of-the-art results, the Berkeley Neural Parser (Kitaev et al., 2019) and the SuPar Neural CRF Parser (Zhang et al., 2020).

Berkeley Neural Parser decouples predicting the optimal representation of a span (i.e. input sequence) from predicting the optimal label, requiring only that the resultant output form a valid tree. This not only removes the underlying grammars found in traditional PCFG parsers, but also direct correlations between a constituent and a label (Fried et al., 2019). A CKY (Kasami, 1965; Younger, 1967; Cocke and Schwartz, 1970) style inference algorithm is used at test time. The parser

uses a self-encoder and can use BERT embeddings for word representations while additionally allowing POS tag prediction to be used as an auxiliary loss task.

SuPar Neural CRF Parser is a two-stage parser, that, similarly to the Berkeley parser, produces a constituent and then a label, and uses a BiLSTM encoder to compute context-aware representations by employing two different MLP layers indicating both left and right word boundaries. Each candidate is scored over the two representations using a biaffine operation (Dozat and Manning, 2017), and the CKY algorithm is used when parsing to obtain the best tree.

Experimental Setup Treebanks have both traces and empty categories removed before training—standard preprocessing for PTB-style treebanks. Features experimented with include: word+char, word+dbmdz BERT embeddings (Devlin et al., 2019)[4], and word+char+dbmdz embeddings. Results are reported including grammatical functions (GFs) using the SPMRL shared task scorer (Seddah et al., 2013, 2014), unless otherwise noted.

3.3 Workflow

As shown in Figure 2, our production of a text involves an iterative process of machine parsing and hand correcting, illustrated here with a relative clause from *Fierrabras*, sentence 36. We first download a .negra file of the text from the ReF:

(3)
```
der PRELS - SB 508
mit APPR - AC 502
golt NA - NK 502
koestlich ADJV - MO 506
belegt VVPPD - HD 506
was VAFIN - HD 508
```

The text is then parsed with a neural parser. However, because texts in the ReF have gold POS tags, we replace the POS tags from the output of the parser with the original POS tags:

(4)
```
(WNP(PRELS der))(IP-SUB(PP(APPR
mit)(NP(NA golt)))(ADVP(ADJV
koestlich))(VVPPD belegt)(VAFIN
was)
```

[4] https://github.com/dbmdz/berts; We also experimented with deepset AI embeddings, but found they consistently yielded worse performance than dbmdz embeddings, most likely due to WordPiece differences (see Reimann and Dakota (2021) for discussion).

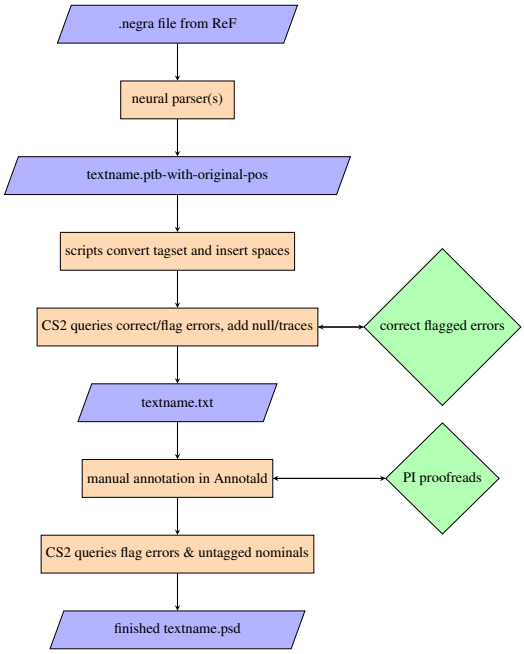

Figure 2: Parsing and correcting workflow

This serves a) to ensure POS tags that are more accurate than the parser output and b) as a check on the syntactic parsing in case of a mismatch between the (gold) POS tag and (machine-parsed) higher constituents.

We then execute several scripts on the parsed texts. An R script converts the STTS-style tags to an intermediate version of our tagset. The intermediate tags are Penn-style tags but maintain some of the fine-grained distinctions of the STTS that aid manual annotation, e.g. distinguishing relative pronouns from determiners:

(5)
```
(WNP(D-relative der))(IP-SUB(PP(P
mit)(NP(N golt)))(ADVP(ADV
koestlich))(VBN-adj-pred-adv
belegt)(AUX-finite was)
```

Sed scripts insert spaces between nodes, making the sentences readable by CS2. CS2 corpus revision queries correct and/or flag errors and insert (when possible) null subjects and traces:

(6)
```
(WNP (D-relative der)) (IP-SUB
(NP-SBJ *pro*-CHECK) (PP (P
mit) (NP (N golt))) (ADVP (ADV
koestlich)) (VBN-adverbial?
belegt) (BEDI^3^SG was)
```

Some flagged errors are manually corrected between queries. The result is passed to an annotator, who using Annotald (Ingason et al., 2018) corrects the parse, assembles higher-level constituents if

Test	Features	SuPar			Berkeley		
		R	P	F	R	P	F
Neues Buch	word+char	39.15	38.71	38.93	28.75	44.53	34.94
	word+dbmdz	40.46	38.43	39.42	39.65	44.45	41.91
	word+dbmdz+char	43.47	41.34	42.38	40.29	46.39	**43.12**
Fierrabras	word+char	31.31	29.91	30.59	22.22	36.75	27.69
	word+dbmdz	36.11	33.39	34.70	42.51	45.78	44.08
	word+dbmdz+char	37.60	34.82	36.15	43.59	47.13	**45.29**

Table 2: Results for SuPar and Berkeley parsers using CHLG trained model and testing on 100 sentences of two different texts from ENHG

necessary, and ensures that GFs are correct. For example, because (6) was not automatically parsed as a relative clause, the CS2 query inserted a null subject instead of a trace; the annotator must embed the clause in CP-REL with null C, add a trace coindexed with the relative pronoun, and delete the extraneous null subject. All sentences are proofread by an expert annotator and returned to the annotator for further correction. Finally, more CS2 queries flag remaining errors and spread case/number tags to any untagged nominals; any flagged errors are again manually corrected. Final gold parse:

(7)
```
(CP-REL (WNP-SBJ-2 (D^N^SG der))
(C 0) (IP-SUB (NP-SBJ *T*-2) (PP
(P mit) (NP (N^D^SG golt))) (ADVP
(ADV koestlich)) (VBN belegt)
(BEDI^3^SG was)))
```

4 Parsing Experiments

4.1 CHLG on ENHG

It is unclear how many sentences we need to build an ENHG-only parsing model, and given that developing a large-scale treebank is a costly and timely process, we treat our current ENHG treebank as a low-resource language and aim to determine how we can facilitate faster annotations. One approach is to leverage the closely related CHLG, given its linguistic relatedness and much larger size. We are not aware of any standard train/development/test splits for the CHLG treebank, and with the limited number of sentences for ENHG, all experiments should be viewed as exploratory and with caution, as different chosen splits may yield noticeably different performance metrics (Dakota and Kübler, 2017), particularly as treebanks may scale in size in the future.

We first trained a parser only with the CHLG treebank using the numbers specified in Table 1 and parsed *Neues Buch* and *Fierrabras*. We then hand-corrected the first 100 sentences from each of the texts and used these sentences as an initial test set to determine to what extent we can use the CHLG treebank to parse the ENHG texts, results of which are presented in Table 2.

Results show, unsurprisingly, that a combination of word+char+dmbdz embeddings yields the best performance for both parsers. However, we see different trends between the parsers. One is that SuPar seems to favor recall over precision, while Berkeley is favoring precision over recall, which is particularly noticeable in the word+char experiments. The large discrepancy is diminished greatly for Berkeley once dmbdz embeddings are utilized, but still precision is favored. We also see that while both parsers achieve simlar performance on *Neues Buch*, Berkeley is significantly better than SuPar once dbmdz embeddings are utilized for *Fierrabras*. Additionally, *Fierrabras* seems to benefit more from the addition of the dmbdz embeddings than *Neues Buch*. One reason may be that dmdbz's embeddings are based on Modern Standard German, and *Fierrabras* is closer to Modern Standard German both temporally (by almost 200 years) and dialectally (i.e., it exhibits fewer Low German characteristics, see section 5 for additional analyses).

4.2 CHLG and ENHG

Based on Table 2, we choose the Berkeley parser for all additional experiments, as it slightly outperforms SuPar. Another rationale is the auxiliary task that predicts POS tags. In our experimental setup, SuPar uses only lexical information (i.e., different word representations), meaning it is more sensitive to lexical variation. The auxiliary task employed by Berkeley may help with such variation more effectively due to including POS information via the auxiliary task. Given that the data is

Train	Dev	Test	R	P	F	POS
CHLG	CHLG	50ENHG	40.89	45.66	43.15	00.00
100ENHG	50ENHG	50ENHG	38.62	63.67	48.07	73.37
CHLG+100ENHG	50ENHG	50ENHG	57.29	67.03	61.77	86.90
CHLG	CHLG	197ENHG	41.68	46.16	44.25	00.03
450ENHG	211ENHG	197ENHG	53.60	68.66	60.20	87.31
CHLG+450ENHG	211ENHG	197ENHG	61.24	70.60	65.69	90.88

Table 3: Results of ENHG and concatenated CHLG+ENHG parsers compared to a base CHLG parser. The number of ENHG sentences is prefixed in each column (e.g., 450ENHG is 450 ENHG sentences).

non-standardized and shows both lexical and syntactic changes, a parser that is potentially more robust to such changes is advantageous. Additionally, while currently annotated texts have gold POS accessible, this will not always be the case going forward. Having the parser still predict POS tags is then optimally beneficial, since many state-of-the-art parsers may choose not to use them or predict them, and it eliminates the need to train a separate POS tagger for future non-POS tagged texts.[5]

Due to a limited number of initial sentences, we perform a set of experiments in which we randomly divide the 200 sentences five times, selecting 100 training sentences, 50 development sentences, and 50 test sentences respectively in each case. We perform two experiments, one in which we only use the 100 sentences for training and another in which we concatenate the 100 sentences with the full CHLG treebank, while in both we use the 50 development and test sentences respectively. Such concatenation setups have proven beneficial in various dependency parsing experiments between dialects and related languages (Velldal et al., 2017; Mompelat et al., 2022). We compare the results against using the initially trained CHLG-only model from Table 2 and report averages over the five runs.

Results show that even 100 trained and 50 development sentences can outperform the CHLG-only model. However, we also see that concatenating the CHLG sentences with the 100 ENHG train sentences results in a substantial boost in performance, in particular to recall and POS accuracy. This is somewhat surprising given that CHLG has a different tagset, but it may be that the parser is able to recognize different lexical items as belonging to a specific language variety, and both treebanks use a similar phrase level annotation scheme, which helps identify higher-level projections.

After parsing and correcting an additional 659 sentences from *Neues Buch*, *Fierrabras*, and *Historia*, we perform a repeat of the same three experiments we did on our initial 200 gold annotated sentences, only now with 450 train sentences, and development and test sets of 211 and 197 respectively. We find that the performance of the CHLG-only model shows no significant change compared to when it is tested on the original 50 sentences, suggesting it can parse the available texts with a high degree of stability due to its linguistic relatedness and likely large number of higher-level projections relevant to ENHG, albeit still yielding suboptimal parses.

While the concatenation of the CHLG and 450 ENHG sentences still yields the best performance, the gap between the concatenation model and the ENHG model is substantially reduced both in terms of F-score and POS accuracy, and is still driven mostly by an increase in recall. This suggests we are approaching a threshold of ENHG sentences needed to build ENHG-only models that can yield results similar to that of models concatenated with the CHLG treebank and will no longer benefit substantially from the CHLG.

4.3 Post-Correction Annotation

One desired advantage of training a parser is to facilitate faster human annotations via post-correction. In order to examine if we can improve the rate of manual annotation, we collect statistics from a single expert annotator from two additional texts, initially parsed using different approaches.

For the shallow parse, we begin not with the output of a neural parser but simply with the terminals and POS tags from ReF. From there, the process

[5]We note that SuPar can utilize tag embeddings as features; however, they are not internally predicted, rather the POS tags must be provided at both train and test time. Thus we would still need to train an external POS tagger for any future data without gold POS tags when using this feature as input.

Text	Model	Words/Hr
Karrenritter	shallow	392
Karrenritter	GFs	340
Geistliche	shallow	273
Geistliche	GFs	352
Geistliche	noGFs	361

Table 4: Words annotated per hour on additional texts using different models: shallow (CorpusSearch queries), noGFs (model without grammatical functions), GFs (model with grammatical functions).

is much like that outlined in 3.3: convert the POS tags, insert spaces, and run CS2 corpus revision queries. These rule based-queries (e.g. build NP out of any adjacent D, ADJ, and/or N; build PP out of adjacent P and NP; build subordinate clause after subordinator, relative clause after a relative pronoun, etc.) function together as a basic parser.

For a parsing model, we have two variations, one with grammatical functions (GFs) and one without (noGFs), both of which are trained using the full CHLG treebank and adding all 859 gold annotated sentences from the first three ENHG texts, while still optimizing on CHLG.

The two additional texts are:

- *Karrenritter*: fiction; ca. 1430; South Rhine Franconian; 540 sentences = 10,041 words
- *Geistliche Mai*: medidation on the crucifix; 1529; Bavarian; currently 100 sentences = 3,045 words

The results in Table 4 suggest that, when the text is syntactically simple like *Karrenritter* (mean sentence length of 18.6 words), correcting from the output of the neural parser is no faster than correcting from a minimally parsed text. However, in a syntactically more complex text such as *Geistliche Mai* (mean length 30.5 words), manual correction is much faster when the text was parsed by a neural parser, either with or without GFs. Example sentences from each text (see Appendix A, Fig 3 and 4) illustrate that sentences from *Karrenritter* are not only shorter but also structurally less complex than those from *Geistliche*.

5 Dialectal differences: case study of *he/er*

There are several exploratory uses for historical treebanking, predominantly the ability to identify and analyze diachronic changes in syntactic structures. A more computational use is to examine how effectively we can develop parsers that cover a range of historical changes, as well as dialectal variation, in a language.

To demonstrate use-cases for both on the PCENHG, we train a parser on a single text and parse the other texts to 1) explore the difficulty in cross-textual parsing given diachronic, dialect, domain, and genre differences and 2) analyze parser outputs of a single, high-frequency function word which has two distinct forms but a single syntactic representation across the different dialects.

5.1 Cross-Text Parsing Results

Table 5 presents results for training and testing on the various texts. Perhaps most striking is the fact that the CHLG-only model produces better parsers on *Fierrabras* and *Historia* than on *Neues Buch*, although the latter is linguistically closest to LG. However, both *Fierrabras* and *Historia* have noticeably shorter sentences (mean sentence length 25.6 and 15.8 words, resp.) than *Neues Buch* (mean length 53 words), thus the parser may just be able to create more efficient trees on the shorter sentences, which are often syntactically simpler. Not only does *Neues Buch* have a noticeable issue in recall, while the other two texts show a better balance, it also has low scores in every experimental setup, except training on itself, with substantially lower scores when training on the other ENHG texts, which also have noticeably shorter sentences. This suggests that the characterization of dialects as similar on traditional (phonological) criteria does not guarantee ease of parsing, due to idiosyncratic properties of particular texts. We also face the challenges of both genre and domain differences in combination with diachronic changes, making efficient cross-textual parsing difficult, let alone building a single unified parsing model.

5.2 *he/er* Analysis

To further illustrate these capabilities, we perform a pilot investigation involving the pronouns *he* and *er* (both of which are Eng. 'he', the masculine nominative singular personal pronoun). The pronoun *he* is found in LG and in Cologne (*Neues Buch*) while *er* is used in the rest of HG (see Fig 1). The dialect of Cologne is transitional between LG and HG for this feature (and several others).

When training on CHLG, we see that the parser is effectively able to correctly project the lexeme *he* in *Neues Buch* to a NP-SBJ, but struggles noticeably with *er* found in *Fierrabras*. Instead, *er* is

Train & Dev	Test	R	P	F	POS
CHLG	Neues Buch	31.76	40.11	35.45	00.06
	Fierrabras	47.63	49.32	48.46	00.00
	Historia	58.55	59.89	59.21	00.00
Neues Buch	Neues Buch	38.83	56.32	45.97	76.53
	Fierrabras	35.29	60.16	44.49	66.01
	Historia	40.92	71.82	52.13	70.54
Fierrabras	Neues Buch	18.80	42.50	26.07	63.08
	Fierrabras	57.00	76.21	65.22	88.06
	Historia	45.66	64.10	53.47	81.65
Historia	Neues Buch	15.66	17.01	16.31	61.48
	Fierrabras	40.45	53.57	46.10	76.99
	Historia	52.76	67.51	59.23	84.29

Table 5: Results for training and testing on different texts.

Train & Dev	Test	Correct	Error
CHLG	Neues Buch	30	2
	Fierrabras	6	40
	Historia	1	2
Neues Buch	Neues Buch	31	1
	Fierrabras	43	3
	Historia	2	1
Fierrabras	Neues Buch	13	19
	Fierrabras	46	0
	Historia	3	0
Historia	Neues Buch	15	17
	Fierrabras	44	2
	Historia	3	0

Table 6: Phrase projection error counts for lexemes *he* or *er* when training on one text and testing on another.

often projected to a NP-OB2 (i.e. indirect object). This is probably due to the fact that *er* is never masc.nom.sg. 'he' in CHLG but rather fem.dat.sg. 'her' or even possessive 'her/their'. Such findings are in line with previous research, showing that increasing differences in lexicon and syntactic structure limit the effectiveness of cross-dialect parsing (see Chiang et al. (2006) for difficulties in Arabic dialect parsing within a PTB framework).

However, training on *Neues Buch*, from the transitional dialect of Cologne, does not show performance degradation seen from CHLG. On *Fierrabras*, it is able to correctly project most *er*, despite being trained where *he* is the realized form. This may be because Cologne and Middle German are HG dialects with similar pronoun systems, with the sole exception of *he/er*, and the parser is able to overcome this difference via POS and syntactic inferences. This is partially supported by the fact that more than half the time *er* is tagged as a determiner in *Fierrabras*, but this does not result in error propagation, since such instances still successfully project to a NP-SBJ. On the other hand, we see that models trained on *Historia* and *Fierrabras* are able to successfully parse *er* in *Fierrabras* and *Historia*, respectively, but show mixed results on *he* in *Neues Buch*.

While parsing on closely related languages or dialects can be successful, important factors, such as irreconcilable differences in function words, can limit the effectiveness. When differences between varieties are more superficial, however, a parser can more adequately overcome minor lexical and syntactic variation.

6 Conclusion

We have introduced the Parsed Corpus of Early New High German. Its introduction and continued development presents an additional resource for research both on diachronic syntax and on parsing.

We have begun construction of the treebank by successfully utilizing a treebank from a closely related language to develop a base parsing system that helps speed up the annotation process. We use a cyclical process in which outputs are sent through a workflow that automates various post-correction requirements, before finally being hand-corrected by an expert annotator, with the new gold sentences able to be used to train a new parser.

As our gold treebank for ENHG continues to grow, we should be able to reduce our dependence on the CHLG treebank. However, we have also shown that while there are lexical and some syntactic differences between the texts, higher-level projections still benefit from the mixing since many

of the rules are applicable in both varieties, even in the presence of lexical and lower-level syntactic differences, as indicated by the case study of *he/er* variation.

Once the Parsed Corpus of Early New High German is complete, we expect to use it to train a model that can parse both Middle High German (1050-1350) and Modern German (1650-present). This will allow the completion of a parsed corpus of the whole history of HG (as well as providing a source of possible additional data for developing additional PTB-style treebanks of Modern Standard German). Such a timespan and the variation in texts will also allow us to contribute simultaneously to both cross-domain and diachronic parsing research, in particular using a single unified model.

References

Þórunn Arnardóttir and Anton Karl Ingason. 2020. A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser. In Costanza Navarretta and Maria Eskevich, editors, *Proceedings of CLARIN 2020*, pages 48–51.

Fabian Barteld, Sarah Ihden, Katharina Dreessen, and Ingrid Schröder. 2018. HiNTS: A Tagset for Middle Low German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3940–3945, Miyazaki, Japan. European Language Resources Association (ELRA).

Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. A Penn-Style Treebank of Middle Low German. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 766–775.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 369–376, Trento, Italy.

John Cocke and Jacob Schwartz. 1970. Programming Languages and Their Compilers. Technical report, Courant Institute of Mathematical Sciences, New York.

Daniel Dakota and Sandra Kübler. 2017. Towards Replicability in Parsing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2017*, pages 185–194, Varna, Bulgaria.

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 (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Muller, and Klaus-Peter Wegera. 2013. HiTS: ein Tagset fur historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics, Special Issue*, (28):85–137.

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

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-Domain Generalization of Neural Constituency Parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy.

Anton Ingason, Jana Beck, and Aaron Ecay. Annotald, v1.13.10 [online]. 2018.

Tadao Kasami. 1965. An Efficient Recognition and Syntax-Analysis Algorithm for Context-Free Languages. Technical report, AFCRL-65-75, Air Force Cambridge Research Laboratory.

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.

Mariya Koleva, Melissa Farasyn, Bart Desmet, Anne Breitbarth, and Veronique Hoste. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics*, 22:107–140.

Anthony Kroch. Penn Parsed Corpora of Historical English [online]. 2020.

Seth Kulick, Anthony Kroch, and Beatrice Santorini. 2014. The Penn Parsed Corpus of Modern British English: First Parsing Results and Analysis. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 662–667, Baltimore, Maryland.

Seth Kulick, Neville Ryant, and Beatrice Santorini. 2022. Penn-Helsinki Parsed Corpus of Early Modern English: First Parsing Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 578–593, Seattle, United States.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ludovic Mompelat, Daniel Dakota, and Sandra Kübler. 2022. How to Parse a Creole: When Martinican Creole Meets French. In *Proceedings of the The 28th International Conference on Computational Linguistics (COLING 2022)*, pages 4397–4406.

Katrin Ortmann. 2020. Automatic Topological Field Identification in (Historical) German Texts. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–18, Online.

Katrin Ortmann. 2021a. Automatic Phrase Recognition in Historical German. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 127–136, Düsseldorf, Germany.

Katrin Ortmann. 2021b. Chunking Historical German. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 190–199, Reykjavik, Iceland (Online).

Beth Randall, Anthony Kroch, and Beatrice Santorini. CorpusSearch 2 [online]. 2004.

Sebastian Reimann and Daniel Dakota. 2021. Examining the Effects of Preprocessing on the Detection of Offensive Language in German Tweets. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 159–169, Online.

Irmtraud Rösler. 1997. *Satz, Text, Sprachhandeln: Syntaktische Normen der mittelniederdeutschen Sprache und ihre soziofunktionalen Determinanten*. Heidelberg: Winter.

Laurits Salveit. 1970. Befehlsausdrücke in mittelniederdeutschen Bibelubersetzungen. In Dietrich Hoffmann, editor, *Gedenkschrift für William Foerst*, pages 278–89. Böhlau.

Beatrice Santorini. Penn Parsed Corpus of Historical Yiddish, v1.0 [online]. 2021.

Anne Schiller, Simone Teufel, Christine Stöcker, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset). Technical report, Universität Stuttgart and Universität Tübingen.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland.

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

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003a. The Penn treebank: An overview. In *Treebanks: Building and Using Parsed Corpora, ed. by Anne Abeillé*, pages 5–22. Kluwer: Dordrecht, Netherlands.

Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Parsed Corpus of Early English Correspondence [online]. 2006.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frans Beths. York-Toronto-Helsinki Parsed Corpus of Old English Prose [online]. 2003b.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2015. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD Parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10, Gothenburg, Sweden.

George Walkden. 2016. The HeliPaD: a parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21(4):559–571.

Joel C. Wallenberg, Anton Karl Ingason Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. Icelandic Parsed Historical Corpus (IcePaHC) Version 0.9 [online]. 2011.

Klaus-Peter Wegera, Hans-Joachim Solms, Ulrike Demske, and Stefanie Dipper. Referenzkorpus Frühneuhochdeutsch (1350–1650) Version 1.0 [online]. 2021.

Peter Wiesinger, Klass Heeroma, and Werner König. German dialect continuum in 1900. https://commons.wikimedia.org/wiki/File:German_dialect_continuum_in_1900_(according_to_Wiesinger,_Heeroma_%26_K%C3%B6nig).png. Accessed: 2022-11-05.

Daniel Younger. 1967. Recognition and parsing of context-free languages in n^3. *Information and Control*, 10(2):189–208.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and Accurate Neural CRF Constituency Parsing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4046–4053.

A Appendix

The following examples from our corpus illustrate the finished product of the annotation workflow outlined in section 3.3. They also exemplify that the shorter sentences from *Karrenritter* (mean sentence length 18.6 words; the illustrated sentence is 19 words) are generally less complex than relatively longer sentences from *Geistliche Mai* (mean length 30.5 words; the illustrated sentence is 33 words). Note that the example from *Geistliche Mai* not only has more embedded clauses (here: a relative clause inside an adverbial clause) but also more complex NPs, with many NPs containing possessives (NP-POS) and/or appositives (NP-PRN).

```
( (IP-MAT (QP-1 (Q^A^SG alles))
          (HVDS^3^SG het)
          (NP-SBJ (PRO^N^SG er))
          (NP-OB1 (PRO^A^SG s)
                  (QP *ICH*-1)
                  (CP-ADV *ICH*-2))
          (NP-OB2 (PRO^D^SG ir))
          (NEG nit)
          (VBN gesagt)
          (PP (P vmb)
              (NP (Q^A^SG alle) (D^A^SG diß) (N^A^SG welt)))
          (CODE <,>)
          (CP-ADV-2 (WADVP-3 (WADV wie))
                    (C 0)
                    (IP-SUB (ADVP *T*-3)
                            (NP-SBJ (PRO^N^SG es))
                            (PP (P zwuschen)
                                (NP (NP (PRO^D^SG im))
                                    (CONJP (CONJ vnd)
                                           (NP (PRO$^D^SG siner)
                                               (N^D^SG frauwen)))))
                            (VBDI^3^SG stund)))
          (CODE <.>))
(ID 1430.NN.Karrenritter.SRhFrk.,13))
```

Figure 3: Example of 19-word sentence from *Karrenritter*

```
( (IP-MAT (PP (P in)
              (NP (D^D^SG dem)
                  (ADJ^D^SG xij)
                  (N^D^SG spiegelein)
                  (NP-PRN (N^D^SG m))))
          (VBI^2^SG schau)
          (CODE <,>)
          (CP-ADV (WADVP-1 (WADV wye))
                  (C 0)
                  (IP-SUB (ADVP *T*-1)
                          (NP-SBJ (D^N^SG dys)
                                  (ADJ^N^SG hochwyrdig)
                                  (N^N^SG creucz))
                          (BEPI^3^SG ist)
                          (NP-PRD (D^N^SG dye)
                                  (ADJP (NP-POS (Q^G^PL aller) (N^G^PL halltum))
                                        (ADJS^N^SG reychest))
                                  (N^N^SG manstrancz)
                                  (CODE <,>)
                                  (CP-REL (WPP-2 (P in) (WNP (D^A^SG dye)))
                                          (C 0)
                                          (IP-SUB (PP *T*-2)
                                                  (VBN gefast)
                                                  (BEPI^3^SG ist)
                                                  (BEN gewossen)
                                                  (NP-SBJ (D^N^SG der)
                                                          (ADJ^N^SG heyllig)
                                                          (NP-POS (Q^G^PL aller)
                                                                  (ADJ^G^PL heylligen))
                                                          (CODE <,>)
                                                          (NP-PRN (D^N^SG der)
                                                                  (VBN^N^SG vergot)
                                                                  (N^N^SG mensch))
                                                          (NP-PRN (NPR^N^SG xpsen)))
                                                  (PP (P mit)
                                                      (NP (PRO$^D^SG seiner)
                                                          (ADJ^D^SG salligen)
                                                          (N^D^SG sell)))))))))
                  (CODE <.>))
      (ID 1529.Fridolin.GeistlicheMai.Bavaria.,98))
```

Figure 4: Example of 33-word sentence from *Geistliche Mai*