Introducing the Indiana Parsed Corpus of (Historical) High German

Christopher Sapp[1], **Elliott Evans**[1], **Rex Sprouse**[2], **Daniel Dakota**[3]
Indiana University
[1]Department of Germanic Studies
[2]Department of Second Language Studies
[3]Department of Linguistics
{csapp, evansell, rsprouse, ddakota}@iu.edu

Abstract

We outline the ongoing development of the Indiana Parsed Corpus of (Historical) High German. Once completed, this corpus will fill the gap in Penn-style treebanks for Germanic languages by spanning High German from 1050 to 1950. This paper describes the process of building the corpus: selection of texts, decisions on part-of-speech tags and other labels, the process of annotation, and illustrative annotation issues unique to historical High German. The construction of the corpus has led to a refinement of the Penn labels, tailored to the particulars of this language.

Keywords: Historical German, Treebanking, Annotations

1. Introduction

In the last 30 years, there has been an explosion of research on syntactic change in several languages, partially facilitated by the development of the Penn family of historical treebanks. This constituency-based annotation captures both linear and hierarchical relations between words, allowing for empirical and quantitative research into a variety of complex syntactic configurations. Key theoretical assumptions that underlie the Penn annotations are loosely based on the Principles and Parameters model (Chomsky (1981) and much subsequent work) of mainstream Generative syntax (with some simplifications, e.g. a flatter IP layer, as in 9). Penn-style corpora exist for many Germanic languages: English (Kroch, 2020; Taylor et al., 2003, 2006), Icelandic (Wallenberg et al., 2011), Yiddish (Santorini, 2021), and Low German (LG; Breitbarth et al., 2020). However, until now High German (HG), which includes not only Standard German but also southern and central dialects of the language, has only been represented in the Penn family by portions of a 16th-century New Testament translation (Light, 2011).

Other HG parsed corpora exist, e.g. Tiger (Brants et al., 2004), TüBa-DZ (Telljohann et al., 2015), and Baumbank.UP (Demske, 2019). However, their annotations violate key structural assumptions of Generative grammar (e.g. no crossing branches) or include as nodes topological fields, which are part of the German grammatical traditions but are not necessarily constituents. More importantly, none of these spans more than one historical stage of the language.

Here we introduce the Indiana Parsed Corpus of (Historical) High German (IPCHG), currently under development. We present the structure of the corpus, the selection of texts, and the parsing/annotation process. Because we are the first to adapt the Penn system to more than one HG text, we discuss our choice of part-of-speech and phrase labels, with a focus on the linguistic peculiarities of Early New High German (ENHG; 1350-1650) that call for unique solutions. We conclude with a use-case study that demonstrates the utility of the corpus for identifying morpho-syntactic as well as variationist features.

2. The Corpus and Texts

2.1. Corpus Structure

The complete corpus, containing over 1.4 million words from the years 1050-1950, consists of three subcorpora. The texts of each subcorpus have been extracted from different source corpora, each with its own annotation system:

- Middle High German (MHG; 1050-1350): 35 texts, approximately 250,000 words, selected from the Referenzkorpus Mittelhochdeutsch (ReM; Klein et al., 2016). ReM texts were manually divided into sentences, and automatic POS tagging, inflectional tagging, and lemmatization were completed/corrected manually.
- ENHG: 64 texts, approximately 520,000 words, selected from the Referenzkorpus Frühneuhochdeutsch (ReF; Wegera et al., 2021). Twenty-four of these, from ReF's sub-corpus Baumbank.UP (Demske, 2019), were syntactically annotated by hand, thus the POS tagging is very accurate. The remaining texts were processed similarly to ReM, although in ReF the tagging of many texts has not been manually verified.

LREC-COLING 2024, pages 9224–9233
20-25 May, 2024.

Figure 1: Regions represented in IPCHG

- New High German (NHG; 1650-1950): 66 texts, approximately 655,000 words, selected from the 300-million-word DTA (Deutsches Textarchiv, 2023). These are automatically divided into sentences (based on punctuation, which can be problematic for early texts) and automatically tagged for part of speech, inflection, and lemma. The tagging in DTA is reported to be of poor quality, at least for early NHG texts (Voigtmann and Speyer, 2023).

All three source corpora are available under CC licenses, and we have informed the creators of the corpora of our adaption of their annotated texts. All texts in our corpus are in the public domain.

We divide the German-speaking area into 12 regions, corresponding to the 10 HG regions of the Bonner Frühneuhochdeutschkorpus (Schröder et al., 2014) plus two northern regions, illustrated in Figure 1.[1] The northern regions are not represented in the early periods because the written language there was LG through the 16th century (but this time is covered by the Corpus of Historical Low German (CHLG; Breitbarth et al., 2020)).

To the extent possible, each region is represented by 1 text for each 50-year time bin, as shown in Figure 2.[2] At the time of submission, we have 30 of the texts gold annotated (shaded in green in Figure 2), containing about 10k tokens each, all from the ENHG subcorpus, i.e. 14th to 17th centuries. These texts can be downloaded or queried on the project's website.[3]

2.2. Text Selection

The structure of the corpus aims to capture as much syntactic variation as possible. Over 40 years of research, going back to Ebert (1980), have shown that many aspects of ENHG word order are sensitive to sociolinguistic factors such as the text's genre and the author's social class and gender. Thus, in addition to variation across time and regions, we attempt to balance the representation of genres (religious, legal, practical, academic, and literary texts) and include female authors when possible. In order to capture features of less formal, even stigmatized language (Schäfer, 2023), we try to include 1-2 dramas per time bin. However, we avoid poetic texts whenever a prose alternative exists for a time/region cell, on the assumption that meter and rhyme can have distorting effects on word order (see discussion in Fleischer and Schallert, 2011).

However, several factors make it impossible to construct a completely balanced historical corpus of HG:

1. Early MHG texts are difficult to localize precisely, so in the late 11th and early 12th centuries, we sometimes select two or three texts to represent a broader dialect area. Still, some MHG cells remain unfilled given the paucity of prose texts in early MHG. Other MHG cells are filled by very short texts.

2. Due to historical developments, many MHG texts are translations of the Bible and other religious works, whereas we avoid translations in ENHG and NHG.

3. Many of our MHG and ENHG texts are sermons, but we have no sermons represented in NHG.

4. The genre 'academic' changes substantively from MHG (mostly theology) to NHG (a variety of scientific disciplines). As a result, religious texts are over represented in the earliest centuries compared to the modern era.

5. Texts through the 15th century are largely dialectal, and in nearly every case the writer (if known) is from the region in which the

[1] Map adapted from https://upload.wikimedia.org/wikipedia/commons/5/55/German_dialect_continuum_in_1900_(according_to_Wiesinger,_Heeroma_%26_König).png. The 12 dialect labels are our own. Historically LG-speaking areas are in blue; HG dialects are yellow (Middle German) and orange (Upper German).

[2] Blank cells indicate that no text is available. Shaded texts are fully annotated and published.

[3] https://ipchg.iu.edu/

Cent.	West Middle German		West Upper German			East Middle German		East Upper German			(Low German region)	
	Cologne	Hesse	Alsace	Swabia	Switzerland	Saxony	Thuringia	Nuremberg	Bavaria	Austria	Northwest	Northeast
11.2			Älterer Physiologus & Gebete		Rheinau.			Williram	Otlohs Gebet			
12.1	Rheinfr. Interlinear & Bamb. Arzneibuch		Alkuins Traktat & Christi Geburt	Londoner Predigt &				Wiener Physiologus & Steinbuch		Prüler		
12.2	Schleizer Psalm	Trierer Interlinear	Lucidarius	Trud. Hohelied	Zuricher Pred.		Predigtfrag.	Schlierbach Ps.	Windberger Psalter	Spec.eccl.		
13.1	3 short texts	Mitteldeutsche Pred.	Millstätter Pred.	Hoffmann. Pred.	Zweifaltener Benediktine		Mülhäuser Rechtsbuch	Prager Predigtentw	St. Pauler Pred.	Leysersche Pred.		
13.2	Die Lilie	Salomons Haus	Freiburger Urkunden	Buch der Konige	Hugo v. S. Victor: Exp		Jenaer Martyrologiu		David v. Augsburg	Admonter Ben.		
14.1	Koelner Klosterpred.	Heligenleben	Nikolaus Pred.	Augsburger Urkunden	Hugo: Predigten	Leipziger Pred.	Berliner Evangelistar	Engelthaler Schwestern	Freisinger Rechtsb.	Klosterneub.		
14.2	Nuwe Buch Köln	Leiden Christi	Büchlein d. ew.	Rotes Buch Ulm	Naturlehre Mainau	Altdeutsche Pred.	Thüringer Spiele	6 Namen Fronleichna	Buch der Natur	Rationale		
15.1	Reimchronik	Karrenritter	Nebuchodonosor	Fuchsfalle	Appenweiler Chronik	Ältestes Stadtbuch	Eisenacher Chronik	Laien-doktrinal	Sendbrief von wahrer	Denkwürdig-keiten		
15.2	Koelhoff Chronik	Jerusalem	Chirurgie	Verbotene Kunst	Edlibach Chronik	Tauler Sermon	Stolle: Memoriale	St. Anselmi Fragen an	Bairische Chronik	Hystoria Troyana		
16.1	Junge fursten	Fierrabras	Butzers Predig	Franck: Weltbuch	Olvier und Artus	Bachmann: Martinus	Vonn gehorsam	Osiander: Grundliche	Geistliche Mai	Stiftbriefe		
16.2	Epitome Warhaftiger	Wahrhaftig historia	Nachbarn	Beschreib. der Reise	Gespenster	Chronica Marsburg	Thüringische Chronik	Dietriech: Summaria	Concordia	Moscouia	Sattler: 3 Predigten	Wahrhaftige Erklärung
17.1	Teutscher Nation	Hessische Chronik	Policeij Ordnung	Lichtkugel	Brun: Schiffarten	Opitz: Poeterey	Peckenstein: Theatri	Opus Theatricum	Kurtze vnd Nothwendig	Faber: Probstein	Das Friede Wünschend	Wahrem Christentum
17.2	Santa Clara: Grammatica	Becher: Psychosoph	Dannhauer: Catechismu	Zeller: Cenuria	Heidegger: Mythoscopi	Weise: Drey ärgsten Ertz-	Comoedia vom	Saar: Ost-Indianische	Furttenbach: Mannhaffter	Beer: Der verliebte	Hamburg: Statuta und	Siegemund: Wehe-
18.1	Niviandts: Güldenes	Bräuner: Pest-	Fassmann: Gelehrte		Scheuchzer: Natur-	Arnold: Secten-	Jacobi: Betrachtung	Die Curieuse	Decker Baumeister		Tschirnhaus: Getreuer	Gottsched: Pietisterey
18.2	Kortum: Jobsiade	Cancrin: Beschr. Der	Mesmer: Magnetismu	Schiller: Naïve und	Meyer: Grossen	Goeze: Zeitvertreib	Reichardt: Land- u.	Glück: Versuch	Sailer: Kurzgefasst	Kempelen: Maschine	Campe: Theophron	Forster: Ansichten
19.1	Hartwig: Die physische	Schopenhauer: J.v.Eyck	Kerner: Geschichte	Schmidlin: Ueber		Laube: Die Bernsteinhe	Niethammer: Streit	Hegel: Wissenscha	Steub: Drei Sommer in	Grillparzer: Ein treuer	Wienbarg: Aesthetisch	Fouqué: Frauen in
19.2	Fuhlrott: Neanderthal	Huber: Sieben	Laband: Das	Huber: Geschichtlic	Wedekind: Frulings	Nietzsche: Also sprach	Schleicher: Darwinische	Wundt: Handbuch	Pocci: Lustiges	Ernst: Training des	Brunn: Griechische.	Gercke: Torpedowaff
20.1	Egger: Christliche	Bremscheid: Christliche		Krukenberg: Frauenbew	Staiger: Grundbegriff	Hampe: Kaisergesch	Krueger: Wahlrecht		Weber: Wissenscha	Adler: Frauen der	Beck: Geschichte	Zetkin: Zur Frage des

Figure 2: Texts and structure of IPCHG

manuscript is produced. After the invention of the printing press, there is variation between local dialects and emerging super-regional printing languages, and texts from the 19th and 20th centuries are increasingly in the Standard language. Moreover, the place of publication for a selected text usually, but not always, corresponds to the author's region of birth. Nevertheless, we hope that the wide geographical basis will capture some syntactic variation, even within Modern Standard German.

6. Low German is the written language in the northern regions through the 16th century. Therefore, High German texts from these regions are represented in our corpus only from 1550 on.

We seek to follow current best practices for sampling for a corpus of this nature. From each selected text, we aim to annotate 10,000 tokens (although some texts, especially ones, are shorter than this.) Texts in ReM and ReF tend to be fewer than 20,000 words, so we simply annotate the first 10,000 tokens (excluding front matter, e.g. the table of contents, prologue, etc.), which is effectively sampling the beginning and middle of the body of the text. For NHG texts from DTA, which are generally much longer, we randomly select a 10,000-word sample from the beginning, middle, or end of the text, following the practice of the British National Corpus (Burnard, 2007).

3. Tag Set

Each corpus in the Penn family uses a slightly different tag set, to account for language-specific differences. The most closely related language to HG is LG, therefore, we often opt for phrasal labels corresponding most closely to those in the CHLG (Breitbarth et al., 2020). However, because the CHLG does not use a Penn tag set for word-level tags (instead keeping the tag set of their source corpus), our tag set is closest to that of the Old Saxon HeliPaD (Walkden, 2015).

3.1. IPCHG vs. STTS

Our source corpora ReM, ReF, and DTA use the Stuttgart-Tübinger Tagset (STTS; Schiller et al., 1995, 1999) or its offshoot Historisches Tagset (HiTS; Dipper et al., 2013). These tag sets are well suited to German grammar but somewhat counter-intuitive (e.g. tags for demonstratives begin with `PD`, while those for prepositions begin with `AP`). Moreover, they encode some basic syntactic information that is redundant in a parsed corpus that shows constituency, as the following examples of our equivalent tags illustrate:

- Instead of distinguishing demonstrative determiners (PDAT) from demonstrative pronouns (PDS), we label both D, and the difference is indicated by whether or not D has a sister in its NP.
- Prepositions (APPR) and postpositions (APPO) are in our system simply P and are distinguished by whether they precede or follow their NP complement.

On the other hand, some Penn-style tags include distinctions not made by STTS:

- We split STTS's VVFIN (finite verb) into VBPI (pres. ind. verb), VBDS (past subj. verb), etc.
- We distinguish STTS's VA (auxiliary verb) into BE for *sein* 'be', HV for *haben* 'have', and RD for *werden* 'become'.

Nevertheless, we exploit STTS's syntactic information as a check on our parsing, as discussed in 4.2.2.

3.2. IPCHG vs. Other Penn Corpora

When we depart from the word-level tags of the older Penn corpora, our tags correspond most closely to HeliPaD.[4] Here we mention two major innovations of HeliPaD: the extension of POS tags with inflectional subtags and the attachment of the lemma to the word form. E.g., the attested word *grosem* 'big.DAT.SG' in our system is annotated as in (1):

(1) `ADJ^D^SG grosem=groß`

To the HeliPaD tag set we have added:

- ADV+P for the prepositional adverbs (e.g. *dabei* 'thereby') and ADV+ADV for bimorphemic adverbs (e.g. *hierhin* 'towards here').
- INDPRO for the indefinites *niemand*, *jemand*, *nichts*, and *etwas*, which given their complex diachronic development call for a label other than PRO or Q.

In addition, we spell some labels slightly differently, e.g. BEG and BEN for HeliPaD's BG and BN (pres. and past participle of 'be') so that all forms of 'be' consistently begin with BE*. For the same reason, we use HVG, HVN, VBG, VBN, etc.

4. Annotation Process

4.1. Text Extraction

All 165 texts have been downloaded from the website of the source corpora and extracted into a format that can be parsed. The texts in the source corpora are tokenized and POS tagged, and some are lemmatized and tagged for inflection.

Texts in the ReF subcorpus Baumbank.UP are treebanks in the Negra format. We use treetools,[5] which extracts the parsed sentences and converts them to a single-line sentence in a Penn-style bracketed format. This is illustrated by the relative clause in (2):

(2) `(S(PRELS der) (AP(PP(APPR mit) (NA golt)) (ADJV koestlich) (VVPPD belegt)) (VAFIN was))`
'which with gold richly covered was'
(1533 *Fierrabras*, 36)

The remaining ReF texts and all ReM and DTA texts are .xml files. We use C6C[6] to convert these to the CoNLL-U Plus format.[7] A series of python scripts[8] extracts the sentences from CoNLL-U Plus into two formats: a Penn-style bracketed format without morphological extensions and lemmata, to be parsed (3), and the same format with these extensions (4), which can be utilized as described in 4.2.2:

(3) `(VROOT (PPER Wir) (AVD aber) (META <,>) (ADJA lieben) (NA bruoder) ($_ /) (META <,>) (VVFIN spricht) (NE Paulus) (META <,>) ...`
' "But we, dear brothers," says Paul ...'

(4) `(VROOT (PPER^Pl^Nom Wir=wir) (AVD aber=aber) (META <,>) (ADJA^Pos^Nom^Pl lieben=lieb) (NA^Nom^Pl bruoder=bruder) ($_ /=_) (META <,>) (VVFIN^3^Sg^Präs^Ind^St spricht=sprechen) (NE Paulus=paulus) (META <,>) ...`

We manually convert certain manuscript abbreviations to letters (e.g. replacing the nasal mark with an *n* or *m* in brackets (*ẽ* to *e*<*n*>). Such replacements are documented on the project website). A script checks for sentences longer than 512 subword tokens, which can prevent the text from being parsed. The long sentence is then split into two, to be rejoined after parsing.

4.2. Parsing

In order to speed up annotations in the earliest stages of project, we chose to leverage the CHLG

[4] See Walkden (2015) for a detailed discussion.

[5] https://github.com/wmaier/treetools

[6] https://github.com/rubcompling/C6C

[7] https://universaldependencies.org/ext-format.html

[8] All relevant treebank creation scripts can be found at https://github.com/ddakota/IPCHG

texts to train and develop a parser to parse ENHG texts (Sapp et al., 2023), which were then manually corrected. Once we had eight gold ENHG texts (approx. 80,000 tokens), we began training and developing on only ENHG sentences. Gold ENHG texts will also eventually form the initial training set for parsing the MHG and NHG subcorpora.

4.2.1. Parser

The Berkeley Neural Parser (Kitaev et al., 2019) was chosen as it is a state-of-the-art parser and can be configured to include an auxiliary task of predicting POS tags, which is particularly beneficial as many of the NHG texts do not contain gold POS tags. We use a combination of word, character and dbmdz embeddings[9] as input. The dmbdz embeddings are created on Modern Standard German (MSG), which exhibits noticeable orthographic differences from earlier stages of German; however, we are not aware of any available embeddings for historical German.

4.2.2. Use of Source Corpora's Tagging

The resulting, parsed text of the clause in (2) appears as (5):

(5) `(WNP (D der)) (IP-SUB (PP (P mit) (NP (N golt))) (ADVP (ADV koestlich)) (VBN belegt) (BEDI was))`

Our initial parse differs from the original annotated structure of the source corpus in several ways:

1. unlike (2), the parser did not form a constituent of the whole relative clause–it has correctly built `IP-SUB` and `WNP` in line with the Penn standard but failed to span `CP-REL`;
2. *mit golt koestlich belegt* is an `AP` (Adj/Adv Phrase) in (2) but is (correctly) not a constituent in our parse, as we treat participles adjacent to 'be' as verbal (i.e. in a statal passive) rather than adjectival (in a predicative AdjP);
3. the parser has correctly (according to our standard) made both *mit golt* and *koestlich* phrases attached at the clause level.

For texts with gold POS/inflection tags and lemmata (some texts in ReF and most in ReM), we use more scripts to replace the automatic POS tagging from the parser with the manually verified POS tags (and if available, morphological extensions and lemmata).[10] In our example, the parser's POS tags in (5) are replaced by the source corpus's original POS tags, while the phrasal labels from the parser remain:

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

If the original POS and morphological tagging are gold, this can serve as a check on the accuracy of the parse in the next steps. In this example, the gold tag `PRELS` for the relative pronoun *der* is more informative for the human annotator than the parser's tag `D`, even though the Penn standard calls for `D`.

4.3. Annotation

4.3.1. Rule-based Preprocessing

Before manual correction and annotation, we run a script that converts the original POS and morphological tags (in STTS) as in (6) to an intermediate version of our tag set as in (7), which maintains some of the basic syntactic distinctions of STTS that can aid manual annotation. For example, `PRELS` is converted to `D-relative`, using the Penn-type tag `D` but including a flag to the annotator to build a relative clause. As another example, the original, gold tag `VVPPD` (past part. used adverbially) is converted to `VBN-adverbial?`, flagging this as potentially needing to be in ADVP (although ultimately our standard calls for this to be treated as a verb at the clausal level).

(7)
```
(WNP (D-relative der))
(IP-SUB (PP (P mit)
            (NP (N golt)))
        (ADVP (ADV koestlich))
        (VBN-adverbial?  belegt)
        (AUX-finite was))
```

A series of corpus revision queries in CorpusSearch 2[11] use rule-based validation to remove these labels if they match the parse (so `D-relative` is replaced by `D` if it is at the beginning of a relative clause; otherwise the `-relative` flag remains.) Other CorpusSearch 2 queries identify and flag obvious errors (e.g. clauses with no verb, attributive adjectives that are not sister of N, heads that occur in the wrong phrase type, etc.). Still other queries insert null elements such as traces and *pro* subjects, as the Berkeley parser does not produce such annotation decisions in its parses. Final queries case/number tag determiners and pronouns whose case is unambiguous (e.g. *dem* can only be dative singular).

[9] https://github.com/dbmdz/berts

[10] The clause in (2), (5-9) is from a text that has gold POS but is not tagged for inflection or lemma.

[11] http://sourceforge.net/projects/corpussearch

In the example clause, because the parser did not successfully build a relative clause, the `-relative` flag remains in (8), and without a subject (or even a subject trace), the corpus revision query has incorrectly inserted a null subject. Moreover, since the `WNP` was not successfully marked as the subject, *der* was not case/number tagged:

(8)
```
(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))
```

4.3.2. Manual Correction and Revision

The result is passed to a human annotator, who using the GUI Annotald[12] corrects the parse and assembles higher-level constituents, guided by the project's extensive on-line annotation manual. The annotators, currently seven people, are Ph.D. holders or graduate students in Germanic linguistics with extensive coursework in German syntax and historical varieties of German. Graduate student annotators undergo a month of supervised annotation when joining the project, and after that initial month their work continues to be proofread by a PI or postdoc.

Each annotator is assigned one text. The main tasks of the annotator are to:

- manually correct flagged structures, e.g. by placing an example like (8) in a relative clause, removing the `-relative` flag, and replacing the *pro* subject with a subject trace (see (9));
- check that the POS (and morphological tagging if present) of the head matches the phrase assigned by the parser, and make the necessary correction (sometimes correcting the phrase label assigned by the parser, but sometimes correcting the original POS tag to match the parse)–in this case, we remove the flag `-adverbial?` from the participle because we treat such examples as statal passives;
- ensure that the attachment of modifiers, the structure of more complex clauses, etc. match the meaning of the sentence;
- add any missing inflectional tagging;
- leave notes in a separate document indicating any non-routine corrections or questions about the annotation, to be addressed by the PIs;
- make manual changes to the annotation after proofreading by a PI or postdoc.

[12] https://github.com/Annotald/annotald

Version	Precision	Recall	F-Score
V1 (manual)	88.27	89.77	89.01
V2 (corrected)	90.78	92.07	91.42
V3 (final)	91.06	92.35	91.70

Table 1: Graduate student annotation vs. gold

The final, gold version of the example sentence then looks like the following:

(9)
```
(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)))
```

4.3.3. Rule-based Postprocessing

Once an entire text has been manually annotated, proofread, and corrected, the annotator passes it to a PI for automatic post-checking. Post-checking further insures a high-quality annotation, by flagging remaining errors and inconsistencies (Booth et al., 2020). This involves another series of CorpusSearch 2 corpus revision queries. After each query, any flagged errors are manually corrected, and the corrected file is the input for the next corpus revision query. After post-checking, the final version is added to the existing available annotated texts.

4.3.4. Annotation Accuracy

To verify the accuracy of the effectiveness of our process for annotating the corpus, we conducted an experiment in which we selected 10 sentences to be annotated by a graduate student annotator with 6 months of experience on the project. The student's annotations were evaluated against a PI's annotation (treated here as the gold standard) of the same sentences. Evaluation metrics, which we use a proxy for inter-annotator reliability, are presented in Table 1. We can see that the student's initial manual annotation (V1) were already very good with respect to the gold annotation. After proofreading by the postdoc and further manual correction (V2), followed by automatic post-checking (V3), the performance continued to incrementally improve. This demonstrates that the manual corrections and rule-based post-checking outlined above are both necessary and sufficient to ensure the highest possible accuracy.

Moreover, as the PIs query the corpus for research projects (see Section 6), any identified er-

rors or inconsistencies are corrected in subsequent versions of the corpus.

5. Annotation Issues

Applying an annotation scheme to a language for the first time often requires language-specific decisions. In some cases, HG structures are also found in the closely related LG, while other issues are unique to our corpus. When faced with a choice, we have two guiding principles. First, our guidelines must enable our team to carry out the annotations consistently and without arbitrary decisions. Secondly, the annotations should be transparent and useful to researchers who are querying syntactic structures in the corpus.

5.1. Issues that HG shares with LG

Booth et al. (2020) identify several syntactic structures of LG that required a language-specific solution. In many cases, we have adopted Booth et al.'s solution at the phrase label (while keeping a Penn-style tag for the head):

- We treat prepositional adverbs like *dabei* 'thereby' as the head of PP.
- For multi-word prepositions like *bis zu* 'until', we place the two Ps as heads of a single PP.
- Phrases like *von ... wegen* 'because of' (lit. 'by way of ...') and *um ... willen* 'for the sake of ...' that are grammaticalizing from PPs to circumpositions are treated conservatively, i.e. with *wegen/willen* as the object of P:

(10)
```
(PP (P vmb)
    (NP (NP-POS (D^G^SG der)
                (N^G^SG iunckfraun))
        (N^A^SG willen))))
```
'for the maiden's sake'
(1557 *Bairische Chronik*, 57)

- Asyndetic subordinate clauses representing indirect speech are treated like *that*-clauses with a null complementizer (C 0).
- A verb-second clause introduced by *wande* 'because', which is ambiguously a main clause or a subordinate clause with extraposition, is labeled not IP-MAT but IP-X.

In a few cases, we have faced the same challenges as Booth et al. (2020) but choose a slightly different strategy. For asyndetic conditional and concessive clauses, we have adopted the label CP-ADV but have chosen not to insert a null complementizer.

In some other cases, we maintain a label used in the Penn Corpora of historical English (Kroch, 2020; Taylor et al., 2003, 2006), rather than following an innovation by the CHLG. To name a few examples, we maintain the use of:

- `FOREIGN` for a string of foreign words, so that we do not have to parse e.g. Latin;
- `QTP` for a direct quotation that does not form a clause;
- `-RFL` on reflexive object NPs.

5.2. Issues Unique to HG

As the first large-scale corpus of HG in the Penn system, the texts in the IPCHG have presented a number of challenges to the Penn labels that call for HG-specific solutions. In addition to the larger issues addressed in this section, many lexeme-specific guidelines have been developed for certain individual words.[13]

5.2.1. Original vs. Annotators' Punctuation

The creators of ReF and ReM manually divided texts into sentences (and smaller clauses and phrases) by including a node like `<boundary_tag="(.)">` in the attributes of the sentence's last token. These are assigned based on the punctuation rules of Modern Standard German. In this example from ReF, the last token of the sentence is an original, manuscript punctuation mark '/':

(11)
```
<tok_anno trans="/" utf="/"
ascii="/" ...>
<boundary tag="(.)"/>
<pos tag="$_"/>
<posLemma tag="$_"/>
<annoType tag="manual"/>
<cora-flag name="boundary"/>
</tok_anno >
```

These boundary tags are useful in the manual correction phase, as our team's annotators can use them to verify decisions about where clause boundaries lie. However, we need to distinguish original, manuscript punctuation marks—which we give the Penn-style POS tags ', ' and '.'—from those added by the creators of ReM and ReF, which we tag `META` and place in angle brackets (parentheses being reserved to surround nodes). Thus the punctuations in (11) appear in our corpus as:

(12)
```
...  (.  /) (META <.>))
```

[13] See `https://ipchg.iu.edu/` for details.

5.2.2. Ambiguous Subordinators

ENHG has an especially large number of words that appear to introduce subordinate clauses. Our general strategy is to place such words in the embedded clause, regardless of etymology, if they appear as the first element of the clause.

CHLG and other Penn-style corpora treat adverbial clauses introduced by a word homophonously an adverb or complementizer (e.g. *do* 'then/when') as if the word heads an AdvP and takes the embedded clause as its complement, as in this (slightly modified) Middle Low German example from the CHLG web page:

(13)
```
(ADVP (KOUS Do)
      (CP-ADV (WADVP-1 0)
              (IP-SUB (ADVP *T*-1)
                      (NP-SBJ (PPER he))
                      (NP-OB1 (DPDS dit))
                      (VVPP gesproken)
                      (VAFIN hadde))))
```
'when he had spoken this . . . '

We, however, tag these homophonous adverb/complementizers as C (unless they have wh-moved, i.e. in questions and relative clauses) and make them the head of the clause:

(14)
```
(CP-ADV (C da)
        (IP-SUB
          (NP-SBJ (PROˆNˆSG er)
          (NP-OB1 (PROˆAˆSG das))
          ...
          (VBDIˆ3ˆSG sprach)
```
'when he said this . . . '

Similarly to (13), CHLG treats apparent prepositions followed by subordinate clause as a PP, headed by P with CP complement. However, we opt to treat these as if the 'preposition' and complementizer form a complex subordinator:

(15)
```
(CP-THT (C auff) (C das)
        (IP-SUB
              (NP-SBJ (PRO es))
              (NP-OB1 (D die)
                      (ADJ wilden))
              (VBDSˆ3ˆPL hoereten)
              ...
```
'so that the wild ones heard it. . . '

5.2.3. Particles

POS tagging of German particles is particularly tricky, because they often defy traditional POS categories and/or are homophonous with other POS.

Modal/discourse particles are difficult to distinguish from adverbs without direct access to intonation. We thus tag potential modal particles ADV.

Relative pronouns can be followed by *da*, which appears to be a holdover from OHG's doubly filled comp (*da* as a relative particle). However, because this is often ambiguous with adverbial *da* 'then/there', we treat all such cases as ADV:

(16)
```
(CP-REL
      (WNP-SBJ-1
            (WDˆNˆSG welches))
      (C 0)
      (IP-SUB (NP-SBJ *T*-1)
              (ADVP (ADV da))
              (RP an)
              (VBDIˆ3ˆSG kam)))
```
'. . . [a ship], which arrived (?there)'
(1557 Staden *Historia*, 242)

Similarly, *dann* 'then', when following a clear subordinator as in *eh dann* 'before (then)' is treated as an ADV within IP-SUB, rather than as a second C.

Finally, superlatives of predicative adjectives (*am kleinsten*, lit. 'on the smallest') appear on the surface to be PPs. Rather than annotating them as literal PPs, we treat *am* 'on the' as an unanalyzed particle that accompanies the adjective:

(17)
```
(ADJP (ADV am)
      (ADJS kleinsten))
```
'(the) smallest'

6. Use-case study

To illustrate the uses of the IPCHG, we present ongoing research on adnominal genitives in ENHG, which may precede (18) or follow (19) the head N:

(18) [$_{Gen}$ meins hrrn] eelicher sun
'my lord's legitimate son'
(1480 *Troyana*, 342)

(19) das haubt [$_{Gen}$ der heyligen jungfrauwen]
'the head of the holy virgin'
(1486 *Jerusalem*, 24)

A preliminary investigation of 17 of the texts annotated thus far demonstrates its usefulness for investigating variation in syntax. First, the corpus has revealed two structures un- or underreported in the literature on adnominal genitives: the 'split' genitive, in which a pre-N genitive is modified by a post-N phrase (20), and the 'embedded' genitive, in which the genitive follows some modifier of N but precedes the N itself (21):

(20) [$_{Gen}$ Josephs] sun [$_{PP}$ von aramathia]
'Joseph of Arimathea's son'
(1430 *Karrenritter*, 472)

(21) eyn besunder [$_{Gen}$ Rulands] streitgesel
'a certain combatant of Ruland'
(1533 *Fierrabras*, 196)

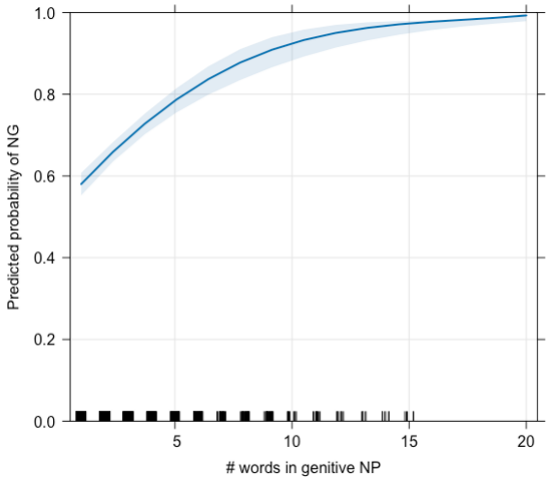

Figure 3: Effect of genitive length on N-Gen order

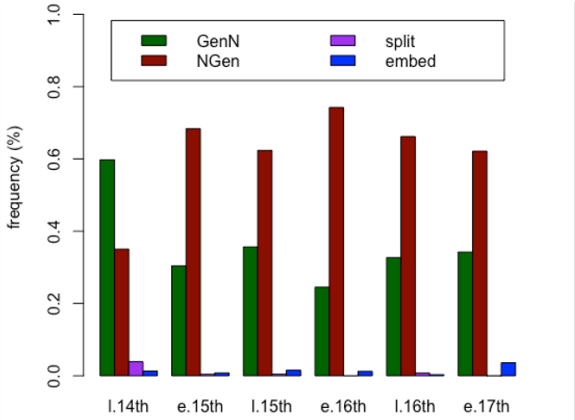

Figure 4: The four word orders over time

Secondly, the corpus allows researchers to test the effect of language-internal variables. For example, we used CorpusSearch 2 to query the effect of the length of the genitive on word order. One-word genitives favor the Gen-N order, while longer genitives increasingly favor the N-Gen order (p<0.001), illustrated in Figure 3.

Thirdly, the structure of the corpus allows the effect of variationist features (dialect, time, and genre) to be tested. We illustrate this in Figures 4 and 5 (p<0.001) with time as the variable.

7. Conclusion

The IPCHG fills a major gap in the Penn family by providing a constituency-parsed corpus of High German. Careful selection of texts and manual correction paired with rule-based validation ensure a high-quality, representative corpus, an early version of which is already available for public use. The ENHG texts annotated and published thus far have led to a refinement of the Penn labels, tailored

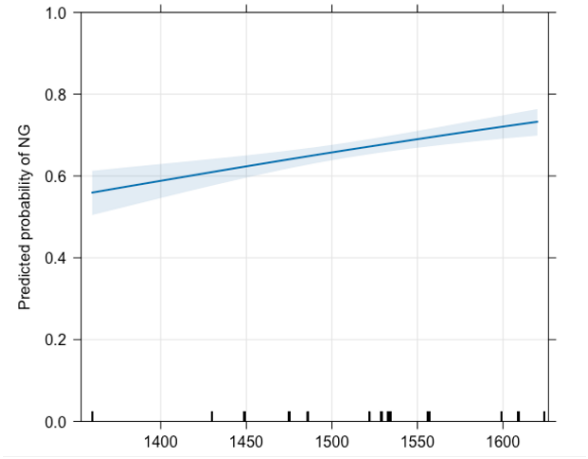

Figure 5: Effect of time on N-Gen order

to the idiosyncrasies of HG. Finally, the IPCHG is already proving itself a useful tool for investigating syntactic variation and change.

8. Acknowledgments

This project is possible thanks to grants from the National Science Foundation (BCS-2314522) and the Indiana University Faculty Research Support Program. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. We are indebted to the developers of the ReM, ReF, and DTA for making these corpora available for public use. Thanks also to Aaron Ecay for help with Annotald and to Beatrice Santorini for sharing her many CorpusSearch 2 scripts.

9. Bibliographical References

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*, 2004 (2):597–620.

Lou Burnard. BNC User Reference Guide: 1 Design of the corpus [online]. 2007.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

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, Special Issue*, 28:85–137.

Robert Peter Ebert. 1980. Social and stylistic variation in early new high german word order: The sentence frame (›satzrahmen‹). 102:357–398.

Jürg Fleischer and Oliver Schallert. 2011. *Historische Syntax des Deutschen: Eine Einführung*. Narr, Tübingen.

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.

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.

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.

Lea Schäfer. 2023. Dramatic texts as a source of stigmatization from below. In *International Conference on Historical Linguistics, Heidelberg, Germany, September 2023*.

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.

Sophia Voigtmann and Ausgustin Speyer. 2023. Where to place a phrase? An informational and generative approach to phrasal extraposition. *Journal of Historical Syntax*, (7):1–48.

10. Language Resource References

Breitbarth, Anne and Farasyn, Melissa and Booth, Hannah and Ecay, Aaron. 2020. *The (parsed) Corpus of Historical Low German (CHLG)*. University of Ghent.

Demske, Ulrike. 2019. *Referenzkorpus Frühneuhochdeutsch: Baumbank.UP*. Universität Potsdam: Institut für Germanistik.

Deutsches Textarchiv. 2023. *Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. Berlin-Brandenburgischen Akademie der Wissenschaften.

Klein, Thomas and Wegera, Klaus-Peter and Dipper, Stefanie and Wich-Reif, Claudia. 2016. *Referenzkorpus Mittelhochdeutsch (1050–1350)*. Ruhr-University of Bochum.

Kroch, Anthony. 2020. *Penn Parsed Corpora of Historical English*.

Light, Caitlin. 2011. *The [Penn] Parsed Corpus of Early New High German*. University of Pennsylvania.

Santorini, Beatrice. 2021. *Penn Parsed Corpus of Historical Yiddish, v1.0*.

Schröder, Bernhard and Wegera, Klaus-Peter and Solms, Hans-Joachim and Schmitz, Hans-Christian and Fisseni, Bernhard. 2014. *Bonner Frühneuhochdeutschkorpus (FnhdC)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

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

Walkden, George. 2015. *HeliPaD: the Heliand Parsed Database*. Zenodo. Full manual online at http://www.chlg.ac.uk/helipad/.

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

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