Introducing the Indiana Parsed Corpus of (Historical) High German

Christopher Sapp, Elliott Evans, Rex Sprouse, Daniel Dakota

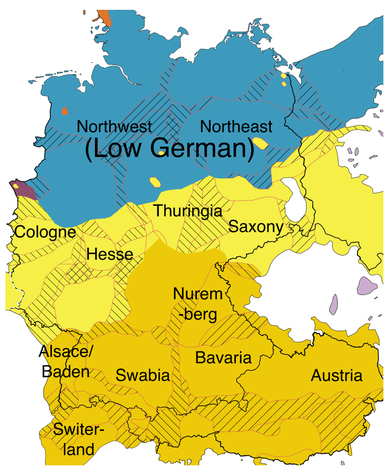
Indiana University

Overview

- Penn family of parsed corpora has not represented High German (HG) until now.
  - Consituency parsed; loosely based on Principles and Parameters grammar.
  - Existing Germanic corpora: English (PPCHE), Icelandic (IcePaHC), Yiddish (PPCHY), and Low German (CHLG).
- We fill this gap with the Indiana Parsed Corpus of (Historical) High German.

Corpus Structure

- 165 texts up to 10,000 words each (total 1.4 million words) from 1050-1950. Three subcorpora:
  - Middle High German (MHG; 1050-1350): 35 texts, selected from the Referenzkorpus Mittelhochdeutsch (ReM). Manually divided into sentences, manually corrected POS and inflection.
  - Early New High German (ENHG; 1350-1650): 64 texts, from the Referenzkorpus Frühneuhochdeutsch (ReF). Some manually, some automatically tagged.
  - New High German (NHG; 1650-1950): 66 texts, from the Deutsches Textarchiv (DTA). Automatic and often inaccurate sentence divisions and tagging.
- 12 German-speaking regions (map below):
  - 10 HG regions (yellow and orange) are represented at every time period.
  - 2 Low German regions (blue) represented beginning in the late 16th century.



Cent.	West Middle German		West Upper German			East Middle German		East Upper German			(Low German areas)	
	Cologne	Hesse	Alsace	Swabia	Switzerla	Saxony	Thuringia	Nuremberg	Bavaria	Austria	Northwes	Northeast
11.2			Älterer Physiologus & Rheinau. Gebete					Williram	Otlohs Gebet			
12.1	Rheinfr. Bamb.	Interlinear & Arzneibuch	Alkuins Predigt	Traktat & Christi	Londoner Geburt			Wiener Steinbuch	Physiologus &	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		Mülhäuser Rechtsbuc	Prager Predigtent	St. Pauler Pred.	Leyserscse Pred.		
13.2	Die Lilie	Salomons Haus	Freiburger Urkunden	Buch der Konige	Hugo v. S. Victor: Exp		Jenaer Martyrolog		David v. Augsburg	Admonter Ben.		
14.1	Koelner Klosterpre	Heligenleben	Nikolaus Pred.	Augsburger	Hugo: Predigten	Leipziger Pred.	Berliner Evangelist	Engelthaler	Freisinger Rechtsb.	Klosterneub.		
14.2	Nuwe Buch Köln	Leiden Christi	Büchlein d. ew.	Rotes Buch Ulm	Naturlehre Mainau	Altdeutsche Pred.	Thüringer Spiele	6 Namen Fronleichn	Buch der Natur	Rationale		
15.1	Reimchronik	Karrenritter	Nebuchodonosor	Fuchsfalle	Appenweiler Chronik	Ältestes Stadtbuch	Eisenacher Chronik	Laien-doktrinal	Sendbrief von	Denkwürdig-keiten		
15.2	Koelhoff Chronik	Jerusalem	Chirurgie	Verbotene Kunst	Edlibach Chronik	Tauler Sermon	Stolle: Memoriale	St. Anselmi	Bairische Chronik	Hystoria Troyana		
16.1	Junge fursten	Fierrabras	Butzers Predig	Franck: Weltbuch	Olvier und Artus	Bachmann: Martinus	Vonn gehorsam	Osiander: Grundlich	Geistliche Mai	Stiftbriefe		
16.2	Epitome Warhaftig	Wahrhaftig historia	Nachbarn	Beschreib. der Reise	Gespenster	Chronica Marsburg	Thüringische	Dietriech: Summaria	Concordia	Moscouia	Sattler: 3 Predigten	Wahrhaftige
17.1	Teutscher Nation	Hessische Chronik	Policeij Ordnung	Lichtkugel	Brun: Schiffarten	Opitz: Poeterey	Peckensteín: Theatri	Opus Theatricu	Kurtze vnd Nothwendi	Faber: Probstein	Das Friede	Wahrem Christentu
17.2	Santa Clara:	Becher: Psychoso	Dannhauer:	Zeller: Cenuria	Heidegger:	Weise: Drey	Comoedia vom	Saar: Ost-Indianisch	Furttenbach:	Beer: Der verliebte	Hamburg: Statuta	Siegemund: Wehe-
18.1	Niviandts: Güldenes	Bräuner: Pest-	Fassmann: Gelehrte		Scheuchzer: Natur-	Arnold: Secten-	Jacobi: Betrachtu	Die Curieuse	Decker Baumeiste		Tschirnhaus:	Gottsched:
18.2	Kortum: Jobsiade	Cancrin: Beschr.	Mesmer: Magnetis	Schiller: Naïve und	Meyer: Grossen	Goeze: Zeitvertrei	Reichardt: Land- u.	Glück: Versuch	Sailer: Kurzgefas	Kempelen: Maschine	Campe: Theophro	Forster: Ansichten
19.1	Hartwig: Die	Schopenhauer:	Kerner: Geschicht	Schmidlin: Ueber		Laube: Die	Niethammer: Streit	Hegel: Wissensch	Steub: Drei	Grillparzer: Ein	Wienbarg: Aesthetisc	Fouqué: Frauen in
19.2	Fuhlrott: Neanderth	Huber: Sieben	Laband: Das	Huber: Geschichtl	Wedekind: Frulings	Nietzsche: Also	Schleicher:	Wundt: Handbuch	Pocci: Lustiges	Ernst: Training	Brunn: Griechisch	Gercke: Torpedow
20.1	Egger: Christliche	Bremschei d:		Krukenberg:	Staiger: Grundbeg	Hampe: Kaiserges	Krueger: Wahlrecht		Weber: Wissensch	Adler: Frauen	Beck: Geschicht	Zetkin: Zur Frage

- When possible, 1 text per region per 50-year time bin (table above).
- Currently 34 ENHG texts gold annotated (shaded green) and available at https://ipchg.iu.edu/
- Balance sociolinguistic factors (genre/register, social class, gender) if possible.

Tagset

- We adapt the Penn tagset to German, mostly following the POS tags of the Old Saxon HeliPaD.
  - Source corpora use variants of the Stuttgart-Tübinger Tagset (STTS).
  - STTS is often redundant in a parsed corpus (e.g. pre- vs. postpositions).
  - STTS also lacks useful distinctions, e.g. collapsing tense & mood as 'finite.'

Annotation Issues

- German-specific adaptations, some following CHLG (Booth et al. 2020). Others are unique to HG, e.g.:
- Homophonous adverbs/complementizers tagged C and placed in the sub. clause.
- Exception: *da* following an overt relative pron. is treated as ADV:

(1)
```
(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)

- Discourse particles are difficult to distinguish from adverbs, thus tagged ADV.
- Superlatives of predicative adjectives (*am kleinsten*, lit. 'on the smallest') look like PPs, but we annotate *am* 'on the' as an unanalyzed particle:

(2)
```
(ADJP (ADV am) (ADJS kleinsten))
```
'(the) smallest'

Annotation Process

1. Extract each sentence from ReM/ReF/DTA into a single bracketed line, using treetools, C6C, and/or python scripts:

   (3)
   ```
   (S (PRELS der) (AP (PP (APPR mit) (NA golt))
   (ADJV koestlich) (VVPPD belegt)) (VAFIN was))
   ```
   'which with gold richly covered was' (1533 *Fierrabras*, 36)

2. Parse with the Berkeley Neural Parser using word+char+dbmdz_embeddings (predict POS tags as aux task), trained targeting gold ENHG:

   (4)
   ```
   (WNP (D der)) (IP-SUB (PP (P mit) (NP (N
   golt))) (ADVP (ADV koestlich)) (VBN belegt)
   (BEDI was))
   ```

3. Replace the POS tagging from the parser with ReM/ReF's gold POS tags (if available), but keep the phrasal labels from the parser:

   (5)
   ```
   (WNP (PRELS der)) (IP-SUB (PP (APPR mit) (NP
   (NA golt))) (ADVP (ADJV koestlich)) (VVPPD
   belegt) (VAFIN was))
   ```

4. Rule-based preprocessing to improve the parse: convert gold POS tags to a version of our tagset (keeping a few distinctions e.g. D-relative), correct or flag obvious errors, insert null elements (here, an erroneous null subject):

   (6)
   ```
   (WNP (D-relative der))
   (IP-SUB (NP-SBJ *pro*-CHECK)
           (PP (P mit) (NP (N golt)))
           (ADVP (ADV koestlich))
           (VBN-adverbial? belegt)
           (BEDIˆ3ˆSG was))
   ```

5. Hand correct the parse, build higher constituents (here, the CP-REL), remove flags (e.g. changing D-relative to D), add inflection:

   (7)
   ```
   (CP-REL (WNP-SBJ-2 (DˆNˆSG der))
           (C 0)
           (IP-SUB (NP-SBJ *T*-2)
                   (PP (P mit) (NP (NˆDˆSG golt)))
                   (ADVP (ADV koestlich))
                   (VBN belegt)
                   (BEDIˆ3ˆSG was)))
   ```

6. Rule-based postprocessing flags remaining errors; final hand correction.
7. Inter-annotator reliability 89.01 after step 5; improves to 91.70 after step 6.

Use Case Study

- Besides filling a gap in the Penn family, the IPCHG is already useful for research in diachronic syntax.
- Adnominal genitives in ENHG precede (8) or follow (9) the head N:

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

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

1. We discovered two additional structures: 'split' (10) and 'embedded' (11) genitives:

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

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

2. We can determine language-internal effects, e.g. the length of the genitive (Fig. 1):

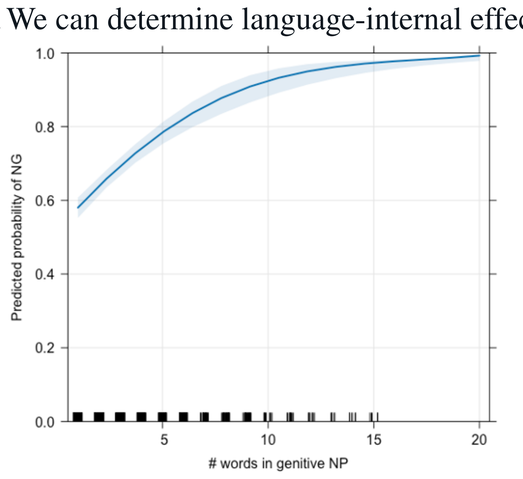


Figure 1: effect of length on order

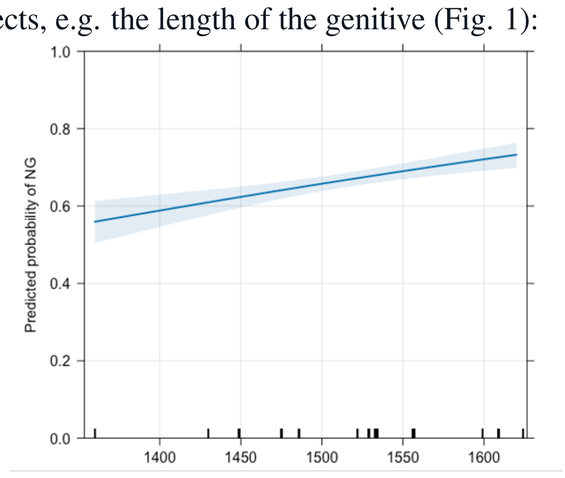


Figure 2: effect of year on order

3. We can determine external/variationist features; e.g. time (Fig. 2).