Annotation of an Early New High German Corpus: The LangBank Pipeline

Zarah Weiß and Gohar Schnelle

39. Jahrestagung der Deutschen Gesellschaft für Sprache:
AG 4: Encoding language and linguistic information in historical corpora

10.03.2017
Pipeline for the syntactical annotation of historical corpora in the framework of the LangBank-Project

Early New High German (ENHG) interesting for:
- Teaching of historical syntax
- Computational linguistics as a non-standard variety

Need for grammatically annotated data
Introduction
The LangBank-Project

- Cooperation project
  - Humboldt-Universität zu Berlin, Prof. Dr. Anke Lüdeling
  - Eberhard Karls Universität Tübingen, Prof. Dr. Detmar Meurers
  - Carnegie Mellon University Pittsburgh USA, Prof. Dr. Brian McWhinney
- Digital infrastructure to support the study of Latin and ENHG
- Extend existing corpora for teaching ENHG and non-linguistic research purposes
- Currently use RIDGES (Odebrecht et al. 2016)
- In planning: Fürstinnenkorrespondenzkorpus

---

1 http://sfs.uni-tuebingen.de/langbank/de/people.html
2 Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fürstinnenkorrespondenz (Version 1.1), Universität Jena, DFG. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm. http://hdl.handle.net/11022/0000-0000-82A0-7
Introduction

RIDGES-corpus

- **Register in Diachronic German Science**
- Designed for research purposes with a variationist approach studying diachronic register
- Version 6.0\(^3\): 50 texts about herbology (1482-1914)
- Only ENHG texts are used for LangBank (1482-1652: 24 texts, 80,095 dipl-token)

\(^3\)https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt
Annotations:

- Diplomatic transcription: **dipl** layer
- Normalization: layers **clean**, **norm**
- Also: lexical, graphical, and content annotations

Normalization

- Orthographical
- Phonological
- Morphological
- **Not** syntactical
Outline

1. Introduction

2. Sentence Boundary Annotation

3. Natural Language Processing

4. Linguistic Complexity

5. Corpus Visualization

6. Summary
Sentence Segmentation

Outline

• Texts need to be segmented into sentences to make Natural Language Processing (NLP) possible

• Graphematical sentence definition in most contemporary european languages:

  My mother went to work and I did my homework.

→ One sentence or two sentences?
Sentence Segmentation
Main issue

- Inconsistent systematic graphematical sentence marking in ENHG problematic
  → No markers at all
  → Differing set of markers (cross, virgel)
  → Lack of consistent functional distribution
Example: A dot could be used to separate verbal arguments


Megenberg1482: Buch der Natur

the highly experienced Hieronymus von Braunschweig uses this water against phthisis, dizziness and to heal those people, who vomit blood

Megenberg1482: Buch der Natur
Sentence Segmentation
Issues and Solution

Issues:
• Lack of systematic graphematical marking in ENHG
• No universal syntactical definition available (Schmidt 2016)

Solution:
• Sentence-segmentation guidelines for the special needs of ENHG
• Syntactical rather than graphematical approach
Definition t-unit (Hunt 1965):  

‘shortest grammatically allowable sentences into which (writing can be split) or minimally terminable unit’

Definition Early New High German t-unit (ENHG-TU):  

‘An ENHG-TU consists of a phrasal head and all of its arguments and adjuncts and nothing else.’ (Weiβ and Schnelle 2016)

• Based on **pragmatic considerations**: facilitating NLP  
  → Produce sentences as short as possible in the case of ambiguity  
  → Using the position of the verb as a marker of subordination

• Based on **linguistic considerations**: map peculiar ENHG constructions
Afinite constructions: covert finite auxilar or copula in periphrastic tenses

Und demnach ich [...] bei Apuleius Platonicus gesehen [habe], dass er etlichen Sternen Kräuter zugezählt [hat] von Bodenstein1557: Wie sich meniglich

And therefore I read in the writings of Apuleius Platonicus about the fact, that he used to attribute the herbs to the stars von Bodenstein1557: Wie sich meniglich

Semantically and syntactically differing set of subordination markers

[...] M. Cato Censorius, von dem L. Columella meldet/ dass er der erste gewesen/ so den Feldbau die lateinische Sprache gelehrt Rhagor1639: Pflantzgart

L. Columella tells us about M. Cato Censorius, that he was the first person, whom taught the latin language in cultivation Rhagor1639: Pflantzgart
Sentence Segmentation
Inter-annotator agreement

- ± sentence boundary annotation by 3 annotators on 5 texts (1532 to 1639)
- 2,609 tokens with approximately 5% sentence boundaries
- Cohen’s $\kappa = 0.8151$ (Davies and Fleiss 1982)
- i.e. almost perfect agreement ($\kappa \geq 0.80$) (Landis and Koch 1977)
Natural Language Processing of ENHG
Approximation Strategy

• Need NLP analyses i) as annotation layers and ii) for complexity analyses
• Lack models for non-standard data and annotated data resources for training
• Use graphematic and morphological normalization of ENHG as proxy
  • + use available models while keeping syntactic structure
  • – requires normalization and looses graphematic and morphological information
Figure: LangBank processing pipeline: From raw data to visualization.
• Require satisfactory performance of NLP tools on normalized layer
• Currently annotate gold standard for dependency and constituency parsing, and morphological analysis
• Annotations by experts using TrEd annotation tool
• First evaluation of performance after 300 gold annotated sentences (April 2017)
• Continue gold standard annotation for entire LangBank Ridges subset
dann die Wasser sind besser und starker, wenn sie fein gemächsam destilliert werden.
Natural Language Processing
Preliminary Impressions
Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

5 Corpus Visualization

6 Summary
Figure: LangBank processing pipeline: Complexity Analysis.
Linguistic Complexity
Motivation

- Restrict queried document space, e.g.
  → Query only documents with high amount of nouns

- Access document level based on linguistic characteristics, e.g.
  → Find documents with high average integration cost, cf. Dependency Locality theory (Gibson 2000)

- Allow to compare texts by linguistic similarity, e.g.
  → Find texts that are syntactically similar to another

• Complexity: elaborateness, variedness, and interrelatedness of a system’s components (Rescher 1998)

• Applied to morphological, lexical, clausal, and sentential domain as well as to domains of textual cohesion, academic language, and cognitive load

• Operationalized to assess for example language proficiency, text readability, writing competence

• See e.g. Crossley, Kyle, and McNamara 2016; Kyle 2016; Lu and Ai 2015; Sheehan, Flor, and Napolitano 2013; von der Brück 2008
Based on contemporary German system (Hancke 2013; Weiß and Meurers Draft):

- 398 measures of elaborateness and variedness of
  - Morphology,
  - Lexicon,
  - Syntax,
  - Academic language, and
  - Correlates of cognitive load

ENHG: directly transfer 313 measures preserving indices from all domains

Lost mostly information on types of connectives and word frequencies
Outline

1. Introduction
2. Sentence Boundary Annotation
3. Natural Language Processing
4. Linguistic Complexity
5. Corpus Visualization
6. Summary
Figure: LangBank processing pipeline: Visualization of Annotations in ANNIS.
Figure: ANNIS Visualization: Startpage
Figure: ANNIS Visualization: Query
Figure: ANNIS Visualization: Constituency Tree
Corpus Visualization

ANNIS

**Figure:** ANNIS Visualization: Topological Field Tree
Corpus Visualization
ANNIS

Figure: ANNIS Visualization: Dependency Tree
Figure: ANNIS Visualization: Complexity Features as Meta
Figure: ANNIS Visualization: Query with complexity information
Outline

1. Introduction
2. Sentence Boundary Annotation
3. Natural Language Processing
4. Linguistic Complexity
5. Corpus Visualization
6. Summary
Summary

- LangBank provides systematic access to ENHG and Latin via
  - Rich linguistic annotation
  - Linguistic complexity characterization
- Access through basic and advanced search interfaces
- Analyze normalized ENHG texts with contemporary German NLP models
- Assume disambiguated sentence boundaries (candidate guidelines provided)
- Semi-automatic pipeline from raw data to annotated corpus
- Current & Future work:
  - Evaluation of NLP performance
  - Automation of normalization via RNNs
  - Simplified user-interface
Thanks for your attention!


Weiß, Zarah and Detmar Meurers (Draft). “Fine-Grained Linguistic Modeling of Textual Complexity Improves German L1 Grade Level Assessment”. In: