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





Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

5 Corpus Visualization

6 Summary

Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

6 Corpus Visualization

6 Summary

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

Introduction Overview

- 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 ¹
 - 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: Fürstinnenkorrespondenzkorpus²

 $^{^{1} {\}tt http://sfs.uni-tuebingen.de/langbank/de/people.html}$

²Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm. http://dhl.handle.net/11022/0000-0000-82A0-7





- Register in Diachronic German Science
- Designed for research purposes with a variationist approach studying diachronic register
- Version 6.0³: 50 texts about herbology (1482-1914)
- Only ENHG texts are used for LangBank (1482-1652: 24 texts, 80,095 dipl-token)

³https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt 🕨 🚊 🗠 🔍 🖓

Introduction RIDGES: Annotations

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

dipl	Fůr	die	giefftigen	thier	(1)
clean	Für	die	giefftigen	thier	1
norm	Für	die	giftigen	Tiere	1.

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

6 Corpus Visualization

6 Summary

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

- Texts need to be segmented into sentences to make Natural Language Processing (NLP) possible
- Graphematical sentence defnition in most contemporary european languages:

My mother went to work and I did my homework.

 \rightarrow One sentence or two sentences?

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

- Inconsistent systematic graphematical sentence marking in ENHG problematic
- \rightarrow No markers at all
- \rightarrow Differing set of markers (cross, virgel)
- $\rightarrow\,$ Lack of consistent functional distribution

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

• Example: A dot could be used to seperate verbal arguments

das Wasser [...] braucht der hocherfahrene Hieronymus von Braunschweig für das Abnehmen. Für den Hauptschwindel. Denen so Blut speien. 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

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

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Guidelines: T-Unit Oriented Approach and general principles

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.' (Weiß and Schnelle 2016)

- Based on pragmatic considerations: facilitating NLP
 - ightarrow Produce sentences as short as possible in the case of ambiguity
 - $\rightarrow~$ Using the position of the verb as a marker of subordination
- Based on linguistic considerations: map peculiar ENHG constructions

Sentence Segmentation Peculiar ENHG constructions: Examples

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 $Rhag {\it or1639: Pflantzgart}$

Inter-annotator agreement

- \pm 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 ($\kappa \ge 0.80$) (Landis and Koch 1977)

Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

6 Corpus Visualization

6 Summary

- ◆ □ ▶ → 個 ▶ → 注 ▶ → 注 → のへぐ

Natural Language Processing of ENHG

- 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

Natural Language Processing of ENHG

LangBank Pipeline

<ロト < 同ト < 回ト < 回ト = 三日 > 二日



Figure: LangBank processing pipeline: From raw data to visualization.

Natural Language Processing Evaluation of Analyses

- 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

Natural Language Processing

Preliminary Impressions

TrEd ver. 2.5049 Default(1/1): C:\Users\ling_schnelle\Desktop\Ar\ptb-berkeley-edge\ptb-berkeley-edge\AlchymistischePractic_1603_Libavius.conll.ptb.pmi

<u>File N</u>ode <u>Tree View M</u>acros <u>S</u>etup <u>H</u>elp



denn die Wasser sind besser und stärker, wenn sie fein gemachsam destilliert werden,



Natural Language Processing

Preliminary Impressions



▲□▶ ▲圖▶ ▲圖▶ ▲圖▶ ▲圖 • のへで

Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

6 Corpus Visualization

6 Summary

Linguistic Complexity LangBank Pipeline



Figure: LangBank processing pipeline: Complexity Analysis.

Linguistic Complexity Motivation

- Restrict queried document space, e.g.
 - \rightarrow Query only documents with high amount of nouns
- Access document level based on linguistic characteristics, e.g. \rightarrow 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

Linguistic Complexity General Aspects

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

- Measures of L2 performance: complexity, accuracy, and fluency (CAF) (Bulté and Housen 2014; Housen, Vedder, and Kuiken 2012; Kyle 2016)
- 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

Linguistic Complexity Transfer to Early New High German

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

- 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

▲□▶ ▲圖▶ ▲≣▶ ▲≣▶ = 差 = のへで

◆□> ◆□> ◆注> ◆注> □注□



Figure: LangBank processing pipeline: Visualization of Annotations in ANNIS.

◆□ ▶ < 圖 ▶ < 圖 ▶ < 圖 ▶ < 圖 • 의 Q @</p>

About ANNIS			Help us to make ANINS better!	logged in as "langbank	 Logou 			
		Help/Examples @ historisches-deutsch ×						
Frence areas man doard	9	document name	 cargus pach 	visualizer	info			
		AlchymistischePractic-VR_1603_Libawius	historisches-deutsch > AlchymistischePractic-VR_1600_Libavius	full text	0			
	100	AlchymiatischePractic_1603_Libawius	historisches-deutsch > AlchymistischePractic_1623_Libevius	fal.text	0			
		ArtzneyBuchleinDerKreutter-VR_1532_Tallat	historisches-deutsch > ArtzneyBuchleinDerKreutter-VR_1532_Tallat	ful test	0			
		ArtzneyBuchleinDerKreutter_1532_Tallat	historisches deutsch > ArtaneyBuchleinDerKreutter_1532_Tallat	ful test	0			
		BuchDerNatur_1482_vonMegenberg	historisches deutsch > BuchDerNatur_1482_vonMegenberg	full test	0			
O Search More + History +		ContrafaytKreuterbuch-CO00XVI-COXLVII_1532_Brunfels	historisches-deutsch > ContralayoKreuterbuch-CC000Wi-CC0UWI_1532_Brunfels	full test	0			
		ContrafaytRreuterbuch-VR_1532_Brunfels	historisches deutsch > ContralisytKreuterbuch-VR_1532_Brunfeis	full text	0			
the right side.		ContrafaytKreuterbuch_1532_Brunfels	buch_1532_Brunfels historisches-deutsch > Contralig/Kinsuterbuch_1532_Brunfels					
		GartDerGesundheit-WL1487_vonCuba	historisches-deutsch > GartDerGesundheit-VR_1487_vonCuba	fulltest	0			
angua unit search uppons		GartDerGesundheit_1487_vonCuba	historisches deutsch > GartDerGesundheit,1487_vonCuba	full test	0			
ter	1010	HortulusSanitatis_1609_Uffenbach	historisches deutsch > HortulusSanitatis_1609_Uffenbach	full test	0			
me * Texts Tokens		Kraeutterbuch_1609_Carrichter	historisches-deutsch > Kraeutterbuch, 1609, Carrichter	full text	0			
storisches-deutsch 24 80,095	0 8	NewKreueterbuch-VR_1563_Handsch	historisches-deutsch > NewKreueterbuch-VR_1563_Handsch	full text	0			
		Newl/reueterbuch_1563_Handsch	historisches-deutsch > NewKreueterbuch_1563_Handsch	ful text	0			
		NewKreuetterBuch-VR_1539_Bock	historisches-deutsch > NewKreuetterBuch-VR_1539_Bock	fulltest	0			
		NewKreuetterBuch_1539_Bock	historisches deutsch > NewKreuetterBuch_1539_Book	full test	0			
		NewKreuterbuch_1543_Fuchs	historisches-deutsch > NewKreuterbuch, 1543, Fuchs	full text	0			
		Paradelszgaertlein_1588_Rosbach	historisches-deutsch > Paradeiszgsertiein, 1588, Rosbach	full text	0			
		PfantzGart-VR_1639_Rhagor	historisches-deutsch > PflantzGart-VR_1639_Rhagor	full text	0			
		PflantzGart-c4_1639_Rhagor	historisches-deutsch > PflantzGart-c4_1639_8hagor	Sal text	0			
		PflantsGart_1639_Rhagor	historisches-deutsch > PflantzGart_1639_Rhagor	Sultest	0			
		WieSichMeniglich-VR_1557_vonBodenstein	historisches deutsch > WeSichWeniglich 48_1557_vonBodenstein	ful test	0			
		WieSichMeniglich_1557_vonBodenstein	historisches-deutsch > WeSichWeniglich, 1557, vonBodenstein	full text	0			
		Wund-Artzney 1652 Gmitt	historisches-deutsch > Wund-Artzney 1652 Greiff	ful test				

Figure: ANNIS Visualization: Startpage

EN About ANNS	Help us to melos AVMS better!	
and a second sec	Help/bampiles Q, Query Result *	
	Bise ted Token Annotations	
Query	K C 1 /29 X H Deplaying Results 1- 10 of 286	Result for: constrait-"NP" & constrait-"NP & #1 > #2
Faile	1 Pitti: edge.exemple > Addymittedreficetto/W,1603,LBavius liberes 114 - 133	left context: 0 right context: 11
	des Nächsten sonderlich deutscher Nation , Nutzen wie es angesehen , geneiche . Es ist nicht ohne , dass solche	
	ART NN AQD AQA NN 5, NN KOKON PPER VVPP 5, VVNF 5, PPER WARN PTXNEG APPE 5, KOLS PAT	
	projection of the second secon	
	der nildstar sonderlich deutsch. Nation – Mutarn wie es anschen – gereichen – es sein micht ohne – dass solder	
	ART NN AQD ADA NN 8, NN 8, NN 8, NPPER VYPP 8, VYPP 8, PPER WAYN PTINES ADD 8, RDUS PAR	
Q Search More + History +	ART AN ADV ADA NV 5, WEN EDIDAN PER VVPP 5, VVDU 5, FPER WEN FINED APR 5, KDUS PEAT	
285 matches	a ja janut ni	
in 23 documents		
	B deploided.ye	
Lorges Loc Search Uppons	2 o Fath: edge-exemple > Addyminister/#Factic, 1603_Lbavius (bokens 196 - 210)	left context: 0 right context: 11
Left Context 0	sette par profe Giber worden in Hesen gemecht, und eit im Deptilleren gebrecht .	
mitrourse 11	PRT ADV ADJA NN NN'N APPR NE VAPP 8, KON ADV APPRART NN VAPP 8,	
	sonipitest sonipitest provide the sonipitest provide the sonipitest provide the sonipitest provided th	
Show contast in tokens (defa., ?	solther ger goli Gias werden in Hessen machen - und oft in detaillieren brauchen -	
Results Per Page 10	TAT ADV ADJA NN WATN APPR NE WAPP 5, KEN ADV APPRART NN WAPP 5.	
and a second sec	PEDIT ADV ADJA NN NAVN APPR NE VHP 6, KEN ADV APPRAT NN VHP 6.	
one accord	8 gol blead, rel	
	a vergester.	

Figure: ANNIS Visualization: Query

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ



Figure: ANNIS Visualization: Constituency Tree



Figure: ANNIS Visualization: Topological Field Tree

◆□ ▶ < 圖 ▶ < 圖 ▶ < 圖 ▶ < 圖 • 의 Q @</p>

Path: edge-example > AlchymistischePractic_1603_Libavius (tokens 196 - 210)														
solche	gar	große	Gläser	werden	in	Hessen	gemacht		und	oft	im	Destillieren	gebraucht	
PIAT	ADV	ADJA	NN	VAFIN	APPR	NE	VVPP	\$,	KON	ADV	APPRART	NN	VVPP	\$.
nom pl neut		nom pl neut pos	nom pl neut	pl 3 pres ind		dat sg neut					dat sg neut	dat sg neut		
solcher	gar	groß	Glas	werden	in	Hessen	machen		und	oft	in	destillieren	brauchen	
PIAT	ADV	ADJA	NN	VAFIN	APPR	NE	VVPP	\$,	KON	ADV	APPRART	NN	VVPP	\$.
PIDAT	ADV	ADJA	NN	VAFIN	APPR	NE	VVPP	\$,	KON	ADV	APPRART	NN	VVPP	\$.
⊞ grid (default	ns)													
🕀 tree (const)														
🕀 tree (topfield)													
🖯 dep (default_	ns)													
				CI										
NK														
		MO		MO	\									
мо	NK	SB NK		NK	1) =									
+ + + +	14	$) \downarrow \bigcirc$	110 111	47 ¥	11C	4								
solche gar groß	e Glás	er werden in Hessen o	pemacht , und of	t im Destillieren ar	sbraucht									

Figure: ANNIS Visualization: Dependency Tree

◆□ ▶ < 圖 ▶ < 圖 ▶ < 圖 ▶ < 圖 • 의 Q @</p>

III About ANNS				Elector	us to mak	e ANNS betteri	logged in as "langbani	C & Logout
Tiesse enter M	and the second s	1.00	Help/txamples Distorisches-deutsch ×					
FIGURE ADD		9	document name			corpus path	visualizer	info
		Query Builder	AldrymistischePractic-VR_1603_Libavius			historisches deutsch > AlchymistischePractic-VR_1603_Libawius	full text	0
		201	AldrymistischePractic_1603_Libavius			historisches deutsch > AlchymistischePractic_1603_Libavius	full text	0
		_	ArtzneyBuchleinDerKreutter-VR_1532_Tallat			historisches deutsch > ArtzneyBuchleinDerKreutter-VR, 1532, Tallat	fulloest	0
			ArtmeyBuchleinDerKreutter_1532_Tallat			historisches-deutsch > Artmey/BuchleinDerKreutter_1532_Tallat	full text	0
			BuchDerNatur_1482_vorWegenberg			historisches-deutsch > BuchDerNatur_1402_vonMegenberg	full text	0
O County Many Minterer			Contrafeyt/Xeuterbuch-CCXXVII-CCXLVII_1532_Brunfels			historisches-deutsch > ContraligeKreuterbuch-CCX0011-CCXUII_1532_Brunfels	full text	0
			Contralayt/veuterbuch-VR_1532_Brunfels			historisches deutsch > ContrafisytKreuterbuch-VR_1532_Brunfels	full text	0
on the right side.	utonal is available		Controlation-techarity 1522 Boundair			historisches deutsch > ContrafaytKreuterbuch_1532_Brunfels	full text	0
Corpus List Search Options Visible: All		Name	oc withymatistic nemacol-wit_1005_bbawlis	Volum		historisches deutsch > GartDerGesundheit/VR_1487_vonCuba	full text	0
		SEN_TU_P	o_EL_Phr_VPsPerTUnit	0.78		historisches-deutsch > GartDerGesundheit, 1487, vonCuba	fullops	0
		SEN_TU_F	o_El_Phr_VZsPerTUnit	0.16		historisches-deutsch > HortulusSanitatis_1609_Uffenbach	fullows	0
Norre	Nome * Texts Tokens		o_E_Phr_complexNominalsPerTUnit	0.47		historisches-deutsch > Kraeutterbuch 1609 Carrichter	full text	0
historisches-deutsch	24 80,095	SEN_TU_F	o_B_Sin_sentenceInfinitivesPerTUnit	0.09		historischer deutsch 5 Neutronaterbuch 50 1000 Maniferb	6.6 mm	
		SEN_TU_F	o_El_Tre_sumNonTerminalNodesPerTUnit	20.59			Dallanda -	
		SEN_TU_F	o_El_Tre_sumParseTreeHeightsPerTUnit	Highlightfolton 4.1 Patholic Section 3-Meetingethod, 1955, Hallon, 2 Hypothyle 2.2 National Section 3- Meetingethod, 1959, Book Hill 2.2 MatrixInte decision 3- Meetingethod, 1959, Book Phil 0.00 MatrixInte decision 3- Meetingethod, 1959, Book Phil 0.00 MatrixInte decision 3- Meetingethod, 1959, Book		historisches deutsch > NewKreueterbuch_15e3_Handsch	141.003	0
		SENUTU	o_El_Tun_complexTUnitsPerTUnit			historisches deutsch > NewKreuetterBuch-VR_1539_Bock	full text	0
		SEN_TU_F	o_EL_Vps_VerbsPerTUnit			historisches-deutsch > NewKreuetterBuch_1539_Bock	fullows	0
		SEN_TU_A	le_ELPer.passivesPerTUnit			historisches-deutsch > NewKreuterbuch_1543_Fuchs	fullows	0
		SEN_TU_A	le_El_Per_quasiPassivesPerTUnit			historisches-deutsch -> Paradeiszaaertieln 1588 Rosbach	full text	0
		SEN_VP_F	o_El_Ou_auxiliaryVerbClustersPerVerbOuster	1.90		historischer deutsch v Blandef ad VB 1020 Blancer	641000	
		SEN_VP_F	Fo_EL_Ou_mainVerbOustersPerVerbCluster 1.50 Fo_EL_Ou_modalWerbOustersPerVerbCluster 1.50			internet and a second se	Dallanda -	
		SEN_VP_F				historisches-deutsch > PflantzGart-64_1639_Rhagor	full text	0
		SENJAPJ	o_El_Clu_muWerbClusterSize	2.50		historisches deutsch > PflantzGart_1639_Rhagor	full cost	0
		SIN W.J.	o_El_Cvp_adjectiveAndAdverbVerbModifiersPerVP	0.31		historisches deutsch > WieSichMeniglich-VR_1557_vonBodenstein	full text	0
		SEN_VP_F	o. FLOvp. participieVerbMod NersPerVP	0.00		historisches deutsch > WeSichMeniglich, 1557, vonBodenstein	fulloest	0
			Wund-Artmey_1652_Greff			historisches-deutsch > Wund-Artaney_1652_Greff	full text	0
			Wund-Antoney_1652_Greiff			historisches-deutsch > Wund-Artaney_1652_Greff	fulltext	

Figure: ANNIS Visualization: Complexity Features as Meta

About ANNIS												Helps	is to make ANI	NIS better!		
	Help/Examples Q, Query Result ×															
construct="NP" &	Base text To	Base text Token Annotations														
entence=/([4-9]+ [1-3][0-	K < 1	753 X X	Displaying Re	sults 1 -	10 of 528											
9]+)\.[0-9]+/	1 O Path: ed	ge-example > C	ontrafaytKreu	terbuch	VR_1532_8	Brunfels (t	okens 8 - 23)									
22223	eine al	Izemeine	Einleitung	20	Lob		ursprünglicher	Alterfahmis		Gebrauch		und	Erkenntnis	der	Kräuter	
	ART AI	DJA AL	NN	APPR	NN	\$,	ADJA	NN	s,	NN	s	KON	NN	ART	NN	s.
	acc sg fem_ac	x sg fem pos	acc sg fem		dat sg m	lasc.	dat sg fem pos	dat sg fem		dat sg m	asc		acc sg fem	gen pl masc	gen pl masc	
	ein al	gemein	Einleitung	zu	Lob		ursprünglich	Alterfahrnis		Gebrauch		und	Erkenntnis	der	Kräuter	
	ART AI	ALC	NN	APPR	NN	\$,	ADJA	NN	\$,	NN	\$	KON	NN	ART	NN	\$.
Q Search More + History +	ART AI	Alc	NN	APPR	NN	\$,	ADJA	NN	\$,	NN	\$	KON	NN	ART	NN	\$.
528 matches	⊕ grid (default_n	🗄 grid (default_ns)														
in 5 documents	tree (const)	I tree (const)														
	tree (topneid) dep (default o	n														
Corpus List Search Options	2 0 Path: ed	»/ ige-example > C	ontrafavtKreu	terbuch	VR 1532 8	Brunfels (t	okens 14 - 26)									
Left Context 0	ursprünglicher	Alterfahrnis	Gebrauc	h .	und 8	rkenntnis	der	Kräuter		Durch (Otto		Brunfels			
Ninha Garana	ADJA	NN	\$, NN	s	KON I	4N	ART	NN	\$.	NN	NE		NE			
Ingrit Contain	dat sg fem por	s dat sg fem	dat[sg]r	nasc		icc sg fer	n gen pl masc	gen pl masc			nom s	g masc	gen sg mas	c		
Show context in tokens (defa ?	ursprünglich	Alterfahrnis	- Gebrauc	h	und 8	Erkenntnis	der	Kräuter		durch (Otto		Brunfel			
Besults Per Page 10	ADJA	NN	\$, NN	\$, KON I	4N	ART	NN	\$.	APPR	NE		NE			
	ADJA	NN	\$, NN	\$, KON I	4N	ART	NN	\$.	APPR	NE		NE			
Order ascending	⊕ grid (default_n	5)														
	tree (const)															
	tree (topfield)	-														
	In the ded idealute is	-ci														

Figure: ANNIS Visualization: Query with complexity information

Outline

1 Introduction

2 Sentence Boundary Annotation

3 Natural Language Processing

4 Linguistic Complexity

6 Corpus Visualization





Summary

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

- 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

Summary

◆□ ▶ < 圖 ▶ < 圖 ▶ < 圖 ▶ < 圖 • 의 Q @</p>

Thanks for your attention!

References I



Bulté, Bram and Alex Housen (2014). "Conceptualizing and measuring short-term changes in L2 writing complexity". In: Journal of Second Language Writing 26, pp. 42–65.



Crossley, Scott A, Kristopher Kyle, and Danielle S McNamara (2016). "The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality". In: *Journal of Second Language Writing* 32, pp. 1–16.



Davies, Mark and Joseph L. Fleiss (1982). "Measuring agreement for multinomial data". In: *Biometrics* 38.4, pp. 1047–1051.

Gibson, Edward (2000). "The dependency locality theory: A distance-based theory of linguistic complexity". In: *Image, language, brain*, pp. 95–126.

Hancke, Julia (2013). "Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language". MA thesis. Eberhard Karls Universität Tübingen.



Housen, Alex, Ineke Vedder, and Folkert Kuiken (2012). "Document Viewing Options: Title: Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA". In: vol. 32. Language Learning & Language Teaching. Amsterdam, Philadelphia: John Benjamins Publishing. Chap. 1–2.

Hunt, Kellogg W. (1965). "Grammatical Structures Written at Three Grade Levels". In: NCTE Research Report 3.

References II

Kyle, Kristopher (2016). "Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication". PhD thesis. Georgia State University.



Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174.



Lu, Xiaofei and Haiyang Ai (2015). "Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds". In: *Journal of Second Language Writing* 29, pp. 16–27.



Odebrecht, Carolin et al. (2016). "RIDGES Herbology - Designing a Diachronic Multi-Layer Corpus". In: Language Resources and Evaluation.



- Rescher, Nicholas (1998). *Complexity: A philosophical overview*. Transaction Publishers.
- Schmidt, Karsten (2016). "Der graphematische Satz. The graphematic sentence.Vom Schreibsatz zur allgemeinen Satzvorstellung. From the written sentence to a notion of the sentence in general." In: Zeitschrift für germanistische Linguistik 44(2), pp. 215–265.
- Sheehan, Kathleen M, Michael Flor, and Diane Napolitano (2013). "A two-stage approach for generating unbiased estimates of text complexity". In: Proceedings of the 2th Workshop on Natural Language Processing for Improving Textual Accessibility. Association for Computational Linguistics. Atlanta, Georgia, pp. 49–58.

References III

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <



von der Brück, Tim (2008). "A Readability Checker with Supervised Learning Using Deep Indicators". In: *Informatica* 32, pp. 429–435.

- Weiß, Zarah and Detmar Meurers (Draft). "Fine-Grained Linguistic Modeling of Textual Complexity Improves German L1 Grade Level Assessment". In:
- Weiß, Zarah and Gohar Schnelle (2016). "Early New High German Sentence Segmentation Annotation Guidelines. Version 4.0." In: LangBank Homepage.