

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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



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
- 5 Corpus Visualization
- 6 Summary

- 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

- 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: Fürstinnenkorrespondenzkorpus ²

¹<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://hdl.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)

³<https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt> ▶

dipl	Für	die	gieftigen	thier	.
clean	Für	die	gieftigen	thier	.
norm	Für	die	giftigen	Tiere	.

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

Sentence Segmentation

Main issue: Example

- Example: A dot could be used to separate verbal arguments

*das Wasser [...] braucht der hocheifahrene 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 graphemata marking in ENHG
- No universal syntactical definition available (Schmidt 2016)

Solution:

- Sentence-segmentation guidelines for the special needs of ENHG
- Syntactical rather than graphemata approach

Sentence Segmentation

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
 - 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 auxiliar 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: Pflanzgart

*L. Columella tells us about M. Cato Censorius, that he was the first person,
whom taught the latin language in cultivation* Rhagor1639: Pflanzgart

Sentence Segmentation

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 ($\kappa \geq 0.80$) (Landis and Koch 1977)

Outline

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

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 loses graphematic and morphological information

Natural Language Processing of ENHG

LangBank Pipeline

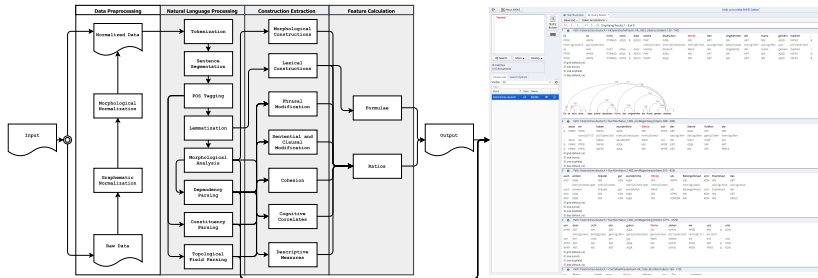


Figure: LangBank processing pipeline: From raw data to visualization.

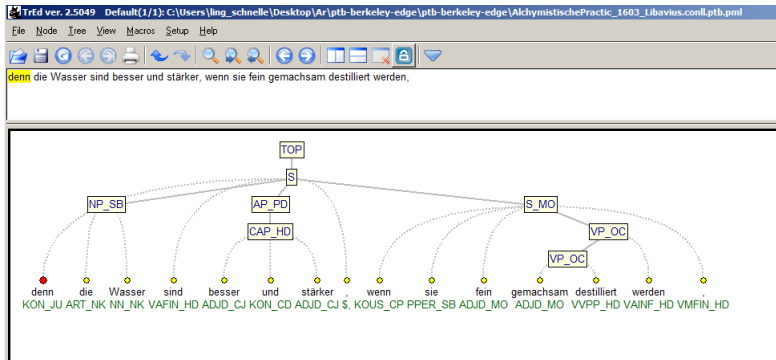
Natural Language Processing

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

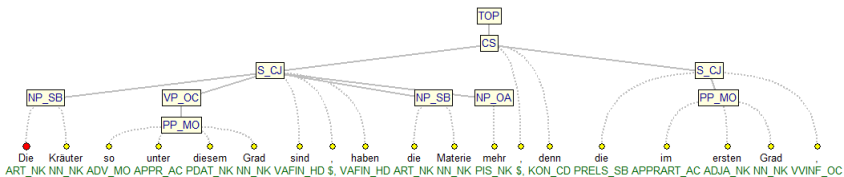
Natural Language Processing

Preliminary Impressions



Natural Language Processing

Preliminary Impressions



Outline

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

Linguistic Complexity

LangBank Pipeline

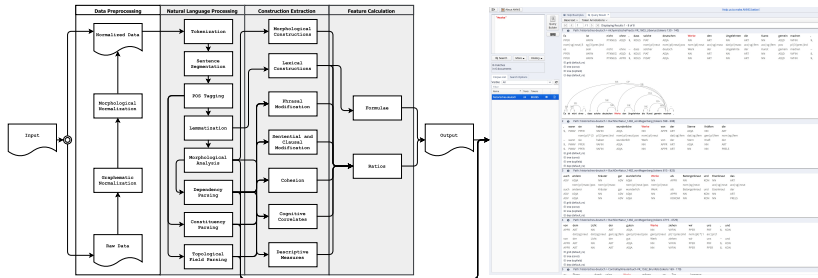


Figure: LangBank processing pipeline: Complexity Analysis.

- 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

- 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

- 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

Corpus Visualization Pipeline

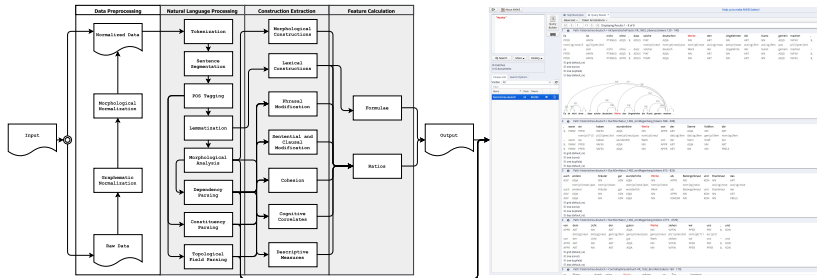


Figure: LangBank processing pipeline: Visualization of Annotations in ANNIS.

Corpus Visualization

ANNIS

The screenshot displays the ANNIS web interface. On the left, there is a search bar with the text "Please enter SQL query" and a "Query Builder" button. Below the search bar, there are buttons for "Search", "More", and "History". A message states "Welcome to ANNIS! A tutorial is available on the right side." Below this, there is a "Corpus List" section with a search box and a dropdown menu set to "Alle". A table below the search box shows search results for "historisches-deutsch" with 26 texts and 80,095 tokens.

The main area of the interface shows a table of search results. The table has four columns: "document name", "corpus path", "viewer", and "info". The results are as follows:

document name	corpus path	viewer	info
AlchymistischePractice-VfL_1603_Libavicus	historisches-deutsch > AlchymistischePractice-VfL_1603_Libavicus	Full text	0
AlchymistischePractice_1603_Libavicus	historisches-deutsch > AlchymistischePractice_1603_Libavicus	Full text	0
ArtzneyBuchleinDerKreutter_VfL_1532_Tallat	historisches-deutsch > ArtzneyBuchleinDerKreutter-VfL_1532_Tallat	Full text	0
ArtzneyBuchleinDerKreutter_1532_Tallat	historisches-deutsch > ArtzneyBuchleinDerKreutter_1532_Tallat	Full text	0
BuchDerNatur_1482_vonMeegenberg	historisches-deutsch > BuchDerNatur_1482_vonMeegenberg	Full text	0
CentralayfKreutterBuch-CCXXXVII-CCXLVII_1532_Brunfels	historisches-deutsch > CentralayfKreutterBuch-CCXXXVII-CCXLVII_1532_Brunfels	Full text	0
CentralayfKreutterBuch-VfL_1532_Brunfels	historisches-deutsch > CentralayfKreutterBuch-VfL_1532_Brunfels	Full text	0
CentralayfKreutterBuch_1532_Brunfels	historisches-deutsch > CentralayfKreutterBuch_1532_Brunfels	Full text	0
GartDerGesundheit-VfL_1487_vonCuba	historisches-deutsch > GartDerGesundheit-VfL_1487_vonCuba	Full text	0
GartDerGesundheit_1487_vonCuba	historisches-deutsch > GartDerGesundheit_1487_vonCuba	Full text	0
HortulusSanitatis_1609_Uffenbach	historisches-deutsch > HortulusSanitatis_1609_Uffenbach	Full text	0
Kraeutterbuch_1609_Carriochter	historisches-deutsch > Kraeutterbuch_1609_Carriochter	Full text	0
NewKreutterbuch-VfL_1563_Handuch	historisches-deutsch > NewKreutterbuch-VfL_1563_Handuch	Full text	0
NewKreutterbuch_1563_Handuch	historisches-deutsch > NewKreutterbuch_1563_Handuch	Full text	0
NewKreutterBuch-VfL_1539_Bock	historisches-deutsch > NewKreutterBuch-VfL_1539_Bock	Full text	0
NewKreutterBuch_1539_Bock	historisches-deutsch > NewKreutterBuch_1539_Bock	Full text	0
NewKreutterbuch_1543_Fuchs	historisches-deutsch > NewKreutterbuch_1543_Fuchs	Full text	0
Paradisgaertlein_1588_Rosbach	historisches-deutsch > Paradisgaertlein_1588_Rosbach	Full text	0
PflantzGart-VfL_1639_Rhagor	historisches-deutsch > PflantzGart-VfL_1639_Rhagor	Full text	0
PflantzGart-e4_1639_Rhagor	historisches-deutsch > PflantzGart-e4_1639_Rhagor	Full text	0
PflantzGart_1639_Rhagor	historisches-deutsch > PflantzGart_1639_Rhagor	Full text	0
WieschMenglich-VfL_1557_vonBodenstein	historisches-deutsch > WieschMenglich-VfL_1557_vonBodenstein	Full text	0
WieschMenglich_1557_vonBodenstein	historisches-deutsch > WieschMenglich_1557_vonBodenstein	Full text	0
Wund-Artzney_1652_Greifff	historisches-deutsch > Wund-Artzney_1652_Greifff	Full text	0

Figure: ANNIS Visualization: Startpage

Corpus Visualization

ANNIS

ANNIS interface showing a query result for the query: `CONCAT("NP" & "CONCAT("NP" & #1 > #2`

Path: edge-example > AchyrisischePraxis_VII,1903_Libavus (tokens 114 - 133)

das	Nächster	sonderlich	deutscher	Notizen	-	Nutzen	was	es	angesehen	,	gerachte	-	es	ist	nicht	ohne	,	klass	sücher	
AET	NN	ADJD	ADJA	NN	&	NN	KONJW	PRIS	VVPP	&	VVPP	&	PPER	PRON	PFNEG	ADP	&	KONJ	PNF	
ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo	ganz[lg]muo
der	nächster	sonderlich	deutscher	Notizen	-	Nutzen	was	es	angesehen	-	gerachte	-	es	sein	nicht	ohne	-	klass	sücher	
AET	NN	ADJD	ADJA	NN	&	NN	KONJW	PRIS	VVPP	&	VVPP	&	PPER	PRON	PFNEG	ADP	&	KONJ	PNF	
AET	NN	ADJV	ADJA	NN	&	WVFIN	KONJW	PPER	VVPP	&	VVPP	&	PPER	WVFIN	PFNEG	ADP	&	KONJ	PNF	

Left context: 0 | Right context: 11

Path: edge-example > AchyrisischePraxis_1603_Libavus (tokens 194 - 210)

schick	gar	große	Güter	wenden	in	Praxen	gemacht	,	und	ist	in	Disciplinen	gebraucht	-
PNF	ADV	ADJA	NN	VAVN	ADP	NE	VVPP	&	KONJ	ADV	ADVPART	NN	VVPP	&
nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo	nicht[lg]muo
schick	gar	große	Güter	wenden	in	Praxen	gemacht	-	und	ist	in	Disciplinen	gebraucht	-
PNF	ADV	ADJA	NN	VAVN	ADP	NE	VVPP	&	KONJ	ADV	ADVPART	NN	VVPP	&
PKAT	ADV	ADJA	NN	VAVN	ADP	NE	VVPP	&	KONJ	ADV	ADVPART	NN	VVPP	&

Left context: 0 | Right context: 11

Figure: ANNIS Visualization: Query

Corpus Visualization

ANNIS

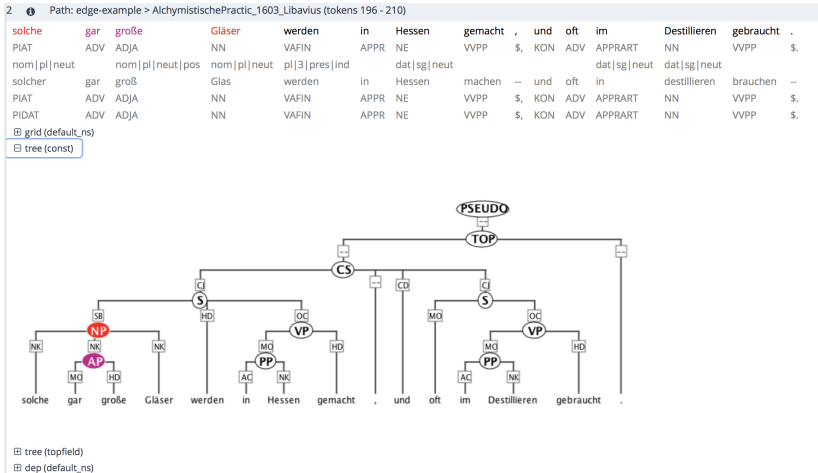


Figure: ANNIS Visualization: Constituency Tree

Corpus Visualization

ANNIS

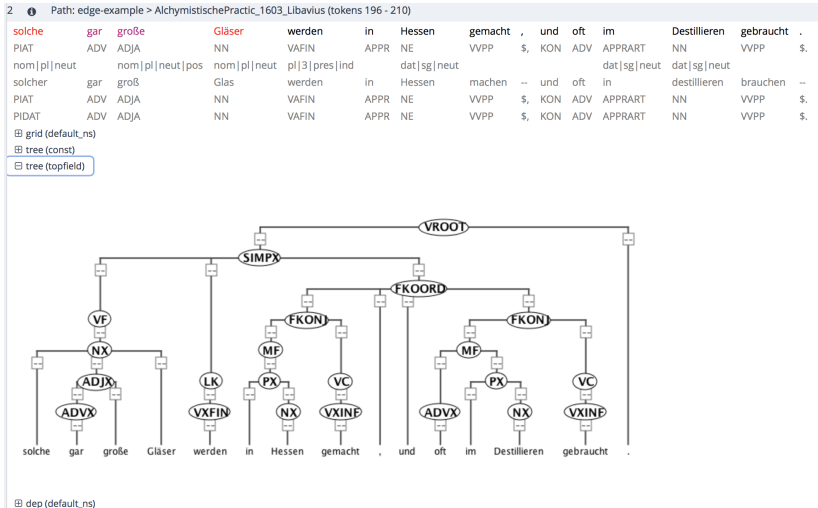


Figure: ANNIS Visualization: Topological Field Tree

Corpus Visualization

ANNIS

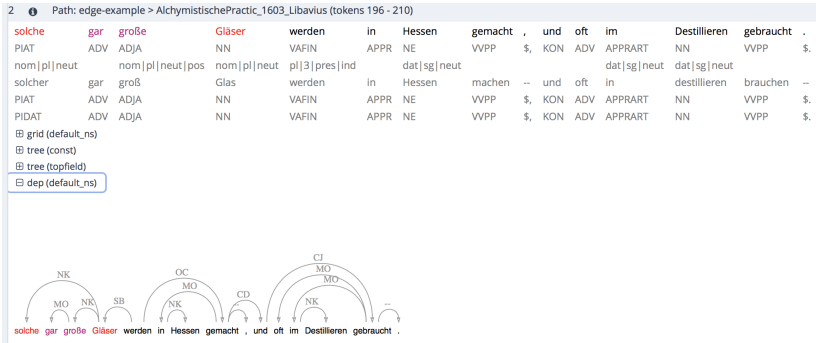


Figure: ANNIS Visualization: Dependency Tree

Corpus Visualization

ANNIS

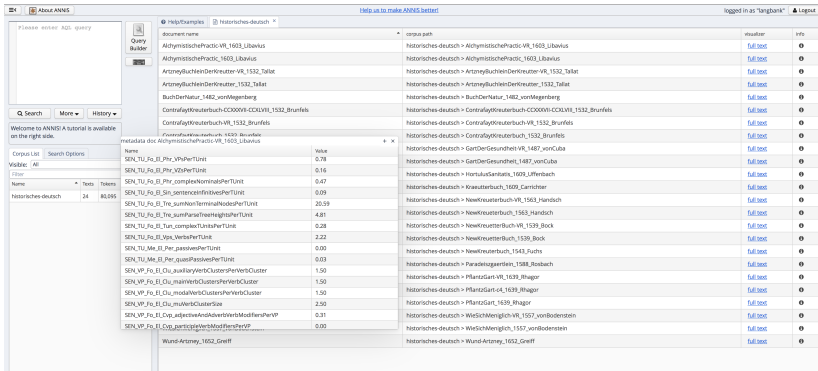


Figure: ANNIS Visualization: Complexity Features as Meta

Corpus Visualization

ANNIS

ANNIS About ANNIS [Help us to make ANNIS better!](#)

Help/Examples Query Result

Base text Token Annotations

1 Path: edge-exemple > ContrafaytKreuterbuch-VR_1532_Brunfels (tokens 8 - 23)

ene	allgemeine	Einleitung	zu	Lob	,	ursprünglicher	Alterfahris	,	Gebrauch	,	und	Erkenntnis	der	Kräuter	,
ART	ADJA	NN	APPR	NN	\$	ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$
acc sg fem	acc sg fem pos	acc sg fem	dat sg fem	dat sg fem	dat sg fem pos	dat sg fem	dat sg fem	dat sg fem	dat sg fem	dat sg fem	acc sg fem	gen pl fem	gen pl fem	gen pl fem	gen pl fem
ein	allgemein	Einleitung	zu	Lob	--	ursprünglich	Alterfahris	--	Gebrauch	--	und	Erkenntnis	der	Kräuter	--
ART	ADJA	NN	APPR	NN	\$	ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$
ART	ADJA	NN	APPR	NN	\$	ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$

grid (default.ms)
 tree (const)
 tree (topfield)
 dep (default.ms)

2 Path: edge-exemple > ContrafaytKreuterbuch-VR_1532_Brunfels (tokens 14 - 26)

ursprünglicher	Alterfahris	,	Gebrauch	,	und	Erkenntnis	der	Kräuter	,	Durch	Otto	Brunfels
ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$	NN	NE	NE
dat sg fem pos	dat sg fem	dat sg fem	dat sg fem	acc sg fem	gen pl fem	gen pl fem	gen pl fem	gen pl fem	nom sg fem	gen sg fem	gen sg fem	gen sg fem
ursprünglich	Alterfahris	--	Gebrauch	--	und	Erkenntnis	der	Kräuter	--	durch	Otto	Brunfel
ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$	APPR	NE	NE
ADJA	NN	\$	NN	\$	KON	NN	ART	NN	\$	APPR	NE	NE

grid (default.ms)
 tree (const)
 tree (topfield)
 elem (default.ms)

Query Builder

528 matches in 5 documents

Search Options

Left Context:

Right Context:

Show context in:

Results Per Page:

Order:

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

Summary

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*.