

Melanie Andresen
melanie.andresen@uni-hamburg.de

**NUMBERS AND RESULTS VS. REALITY
AND FATE: THE ACADEMIC LANGUAGE OF
LINGUISTICS AND LITERARY STUDIES**

#dlbb

Linguistics and Literary Studies

Linguistics and Literary Studies are often two parts of one study program or department.

Linguistics and Literary Studies

Linguistics and Literary Studies are often two parts of one study program or department.

- Similar disciplines?
- Similar academic writing conventions?
- Common scholarly identity?

Linguistics and Literary Studies

Linguistics and Literary Studies are often two parts of one study program or department.

- Similar disciplines?
- Similar academic writing conventions?
- Common scholarly identity?

Rather not.

Objectives

- Raise awareness for interdisciplinary differences (students and faculty)
- Get a more differentiated picture of disciplinary variation in the humanities

Research Questions

- How do the academic languages of linguistics and literary studies differ?

Research Questions

- How do the academic languages of linguistics and literary studies differ?
- How can we identify those differences in a data-driven way?

Methodological Considerations

Data-Driven Research

Now possible because of the new availability of

- 1 digital data
- 2 computational power
- 3 analytical procedures exploiting 1 and 2.

“rather than testing a theory by analysing relevant data, new data analytics seek to gain insights ‘born from the data’.”

(Kitchin 2014)

Corpus-Driven Research (CDR)

corpus-based

(theory-driven)

base the categories of your analysis on a theory

use the corpus “mainly to expound, test or exemplify theories and descriptions” (Tognini-Bonelli 2001, p. 65)

deductive

corpus-driven

develop the categories of your analysis from the data

“descriptions aim to be comprehensive with respect to corpus evidence” (ibid., p. 84)

inductive

CDR and Annotations

Annotations are disapproved, because

- the step from messy primary data to annotation categories “will ensure that the data will finally fit the theory” (Tognini-Bonelli 2001, p. 73).
- in the abstraction to categories, information is lost.
- predefined categories are not corpus-driven themselves.

(Tognini-Bonelli 2001; Sinclair 1992)

CDR and Annotations

- The step from messy primary data to annotation categories “will ensure that the data will finally fit the theory” (Tognini-Bonelli 2001, p. 73).

→ Grouping data into abstract categories in order to arrive at generalizations is a key step of academic knowledge production.

CDR and Annotations

- In the abstraction to categories, information is lost.

→ The information is still there and can be accessed when needed. The relation between the categories and the primary data should be explored.

CDR and Annotations

- Predefined categories are not corpus-driven themselves.

→ True, but not a problem. In order to make academic progress, we have to build on things others have achieved.

Conclusion

- The present research is an example of data-driven research.
- The data used in this study are corpora.
- Still, the term ‘corpus-driven’ is not employed here, because of the additional assumptions made in the corresponding discourse.

My Position

- In data-driven research, I do not test hypotheses based on theories.
- In data-driven research, I generate hypotheses based on data.
- In data-driven research, I can enhance my data with established categories in order to build on previous research e. g. by annotation.
- I must relate the results of my data-driven research to existing theories.

Data

Data

- 60 German PhD theses
 - 30 from linguistics (1.4 million token)
 - 30 from literary studies (2.2 million token)

Data

- 60 German PhD theses
 - 30 from linguistics (1.4 million token)
 - 30 from literary studies (2.2 million token)
 - submitted at 15 German universities
 - submitted between 2003 and 2012
 - published Open Access
- no random sample, but balanced corpus (see Biber 1993)

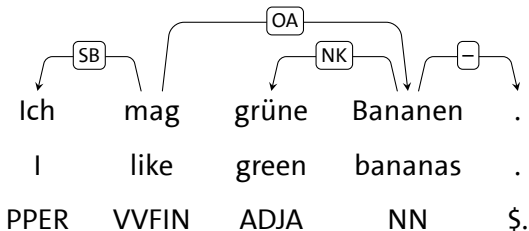
Preprocessing

- PDF → HTML → plain text
- semi-automatic deletion of
 - non-academic language (examples, citations, ...)
 - elements that disrupt text flow (footnotes, tables, ...)

Preprocessing

- PDF → HTML → plain text
- semi-automatic deletion of
 - non-academic language (examples, citations, ...)
 - elements that disrupt text flow (footnotes, tables, ...)
- automatic annotation using MATE (Bohnet 2010)
 - parts-of-speech
 - dependency syntax

Annotation Levels



Tagsets: Schiller et al. (1999) and Albert et al. (2003)

n-Gram Analysis

n-gram = sequence of n elements (words, parts-of-speech,...)

n-Gram Analysis

n-gram = sequence of n elements (words, parts-of-speech,...)

I like green bananas.

2-grams: *I like, like green, green bananas*

3-grams: *I like green, like green bananas*

n-Gram Analysis

n-gram = sequence of n elements (words, parts-of-speech,...)

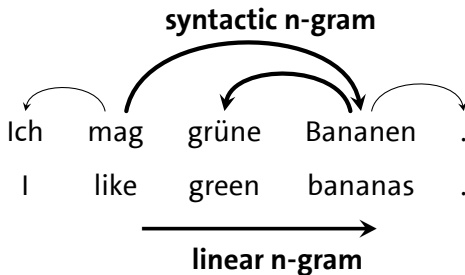
I like green bananas.

2-grams: *I like, like green, green bananas*

3-grams: *I like green, like green bananas*

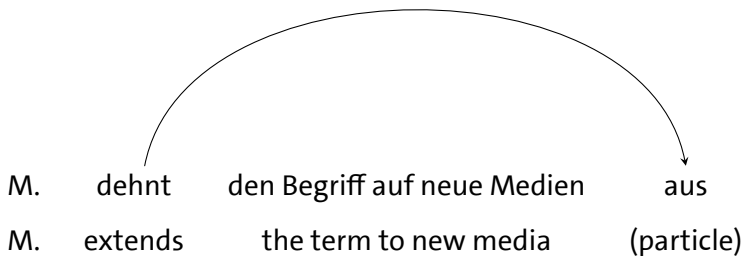
→ Which n-grams are much more frequent in linguistics than in literary studies and the other way around?

Linear vs. Syntactic n-Grams



see Goldberg & Orwant (2013) and Sidorov et al. (2012)

Linear vs. Syntactic n-Grams



Data Sets

Frequencies of:

- linear and syntactic sequences
- of tokens, part-of-speech tags and syntactic functions
- of length 1-5

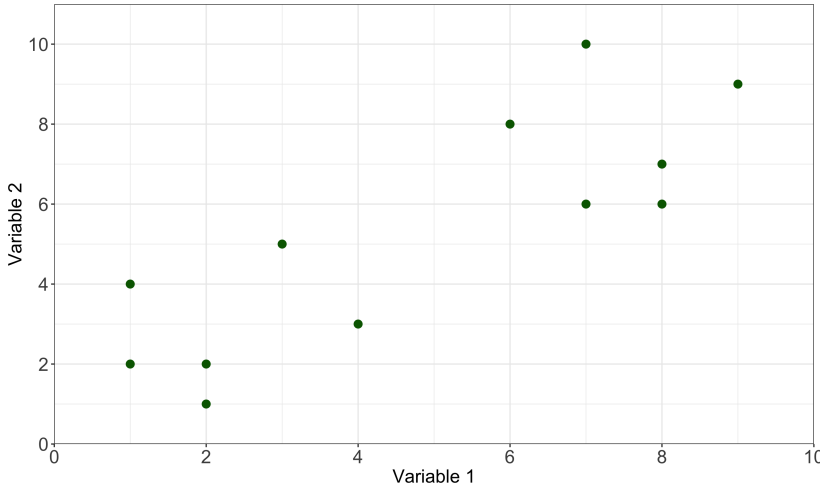
Method

Principal Component Analysis

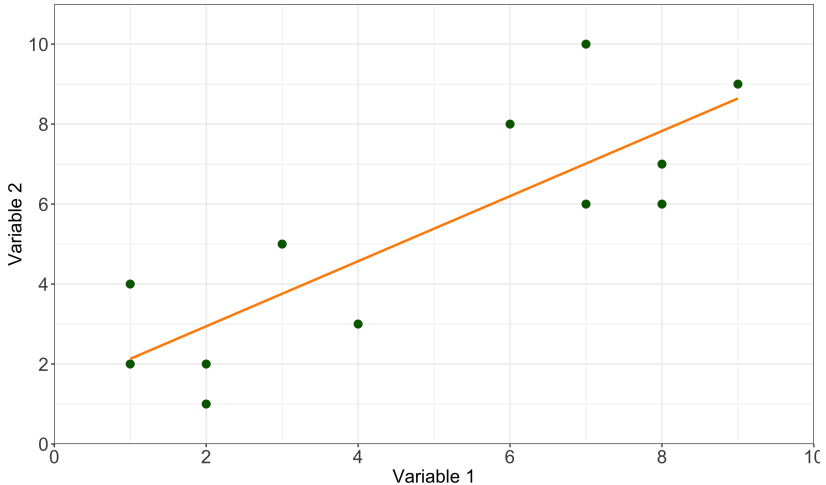
- Starting point: data set with many variables (= dimensions)
- e. g. relative frequencies of 44 parts-of-speech
- Aim: reduce number of dimensions and lose as little information as possible

see e. g. Binongo & Smith (1999)

Principal Component Analysis



Principal Component Analysis



Principal Component Analysis

The new dimension is defined in terms of the original variables:

$$PC1 = 2.5 \cdot \text{variable 1} + 1.3 \cdot \text{variable 2} + \dots$$

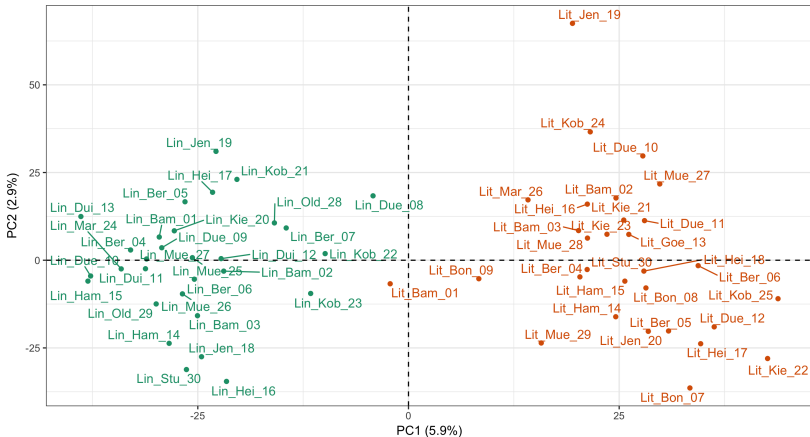
- Every text can be located in the new dimension.
- The loading scores of the variables correspond to their influence.

PCA: Advantages

- Works on unlabeled data
- Confirms the empirical importance of the variable 'discipline'
- allows for internal differentiation
(*Text A is 'more linguistic' than text B.*)

Results

Text Clustering by Tokens

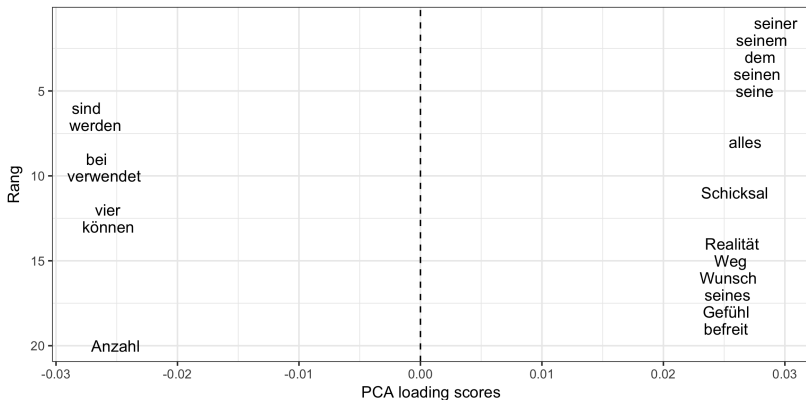


Fach • Linguistik • Literaturwissenschaft

Most Important Variables (PC1)

Linguistics

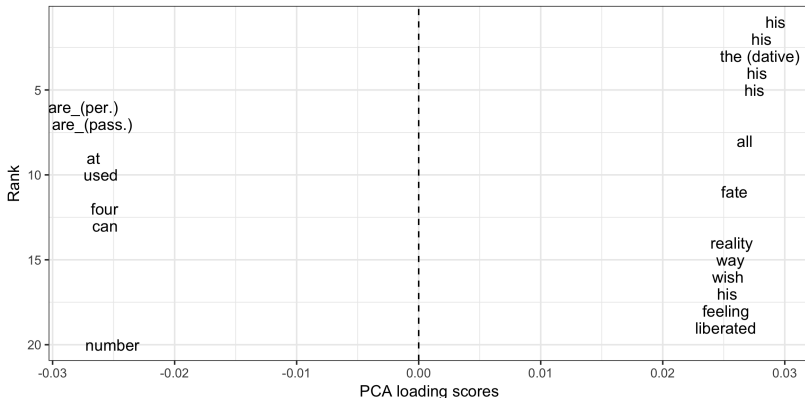
Literary Studies



Most Important Variables (PC1)

Linguistics

Literary Studies



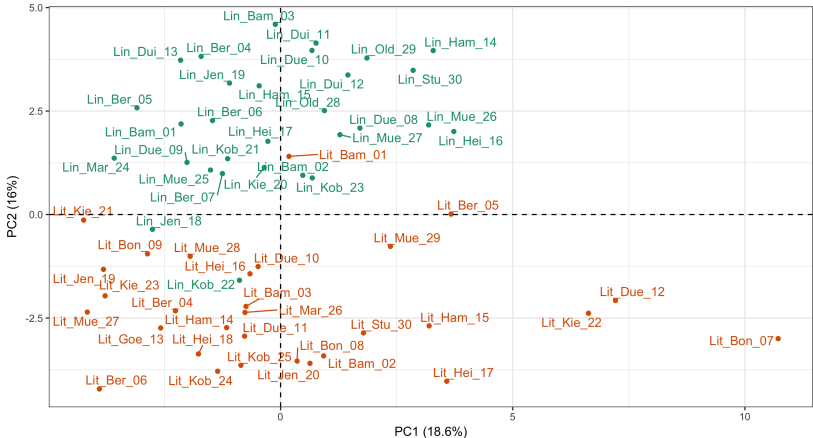
Most Distinctive Nouns

Linguistics		Literary Studies	
Anzahl	number	Schicksal	fate
Unterschiede	differences	Weg	way/path
Tabelle	table	Wunsch	wish
Untersuchungen	studies	Realität	reality
Auswertung	evaluation	Gefühl	feeling
Ergebnisse	results	Leben	life
Fällen	cases	Lebens	(of) life
Ergebnissen	results	Menschen	people
Daten	data	Nacht	night
Unterschied	difference	Gewalt	violence

Most Distinctive Verbs

Linguistics		Literary Studies	
können	can	empfindet	feels
lassen	let	sucht	searches
ergeben	result in	verliert	loses
verwendet	use	erinnert	remembers
auftreten	appear/occur	erfährt	experiences/learns
aufweisen	show/exhibit	begegnet	encounters
umfassen	comprise	verkörpert	embodies
beziehen	refer to	gerät	gets into
vorliegt	be available	fühlt	feels
könnten	could	zieht	pulls

Text Clustering by Parts-of-Speech

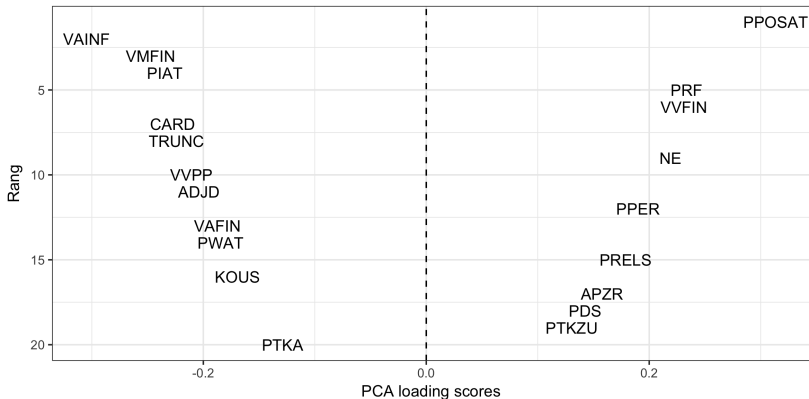


Fach • Linguistik • Literaturwissenschaft

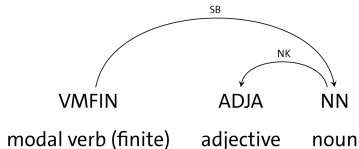
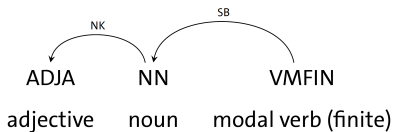
Most Important Variables (PC2)

Linguistics

Literary Studies



Most distinctive pattern for Linguistics:



- 3861 occurrences in the corpus
- 3747 unique with respect to the token level
- 2260 in the linguistic subcorpus

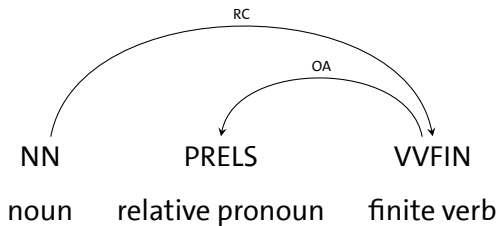
Examples:

*negative Konnotation kann
redaktionelle Bearbeitung soll
syntaktische Prozesse können
vorliegende Arbeit soll*

negative connotation can
editorial editing should
syntactic processes can
present study shall

Syntax

Most distinctive pattern for literary studies:



Examples:

Rolle, die ... spielt

Gott, den ... gibt

Wissen, das ... besitzt

Mann, den ... liebt

role that ... play(s)

god that .. exists

knowledge that ... possess(es)

man that is loved by ...

Back to Theory

How are the differences in language related to theories we have about what defines a discipline?

- topic
- method
- epistemological interest
- additive vs. substituting knowledge
- ...

Conclusions

Conclusions

Linguistics vs. Literary Studies:

- words related to quantitative methods in linguistics
 - complex verbs (modal and auxiliary verbs) in linguistics
 - words related to people in literary studies
 - relative clauses in literary studies
- mapping to theory-derived categories

Conclusions

Methodology:

- use of annotations enhances analysis
- syntactically informed sequences capture long-distance relations and variability in word order
- open question: Influence of substructures
- PCA is not helpful if the dimensions do not correspond to the disciplines

Thank you!

- Albert, Stefanie et al. (2003). *TIGER Annotationsschema*. URL:
https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tiger_annot.pdf.
- Biber, Douglas (1993). "Representativeness in Corpus Design". In: *Literary and Linguistic Computing* 8.4, pp. 243–257.
- Binongo, José Nilo G. & M. W. A. Smith (1999). "The Application of Principal Component Analysis to Stylometry". In: *Literary and Linguistic Computing* 14.4, pp. 445–466. DOI: 10.1093/llc/14.4.445.
- Bohnet, Bernd (2010). "Very High Accuracy and Fast Dependency Parsing Is Not a Contradiction". In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.
- Goldberg, Yoav & Jon Orwant (2013). "A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books". In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 241–247.
- Kitchin, Rob (2014). "Big Data, New Epistemologies and Paradigm Shifts". In: *Big Data & Society* 1.1, p. 2053951714528481. DOI: 10.1177/2053951714528481.
- Schiller, Anne, Simone Teufel, Christine Thielen & Christine Stöckert (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Stuttgart, Tübingen.
- Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh & Liliana Chanona-Hernández (2012). "Syntactic Dependency-Based N-Grams as Classification Features". In: *Advances in Computational Intelligence*. Ed. by Ildar Batyrshin & Miguel González Mendoza. Lecture Notes in Computer Science 7630. Springer, pp. 1–11. DOI: 10.1007/978-3-642-37798-3_1.
- Sinclair, John (1992). "The Automatic Analysis of Corpora". In: *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*. Ed. by Jan Svartvik. Trends in Linguistics 65. Berlin [u.a.]: Mouton de Gruyter, pp. 379–397.
- Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Studies in Corpus Linguistics 6. Amsterdam [u.a.]: Benjamins.