Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR GEISTESWISSENSCHAFTEN

Melanie Andresen & Heike Zinsmeister
{firstname}.{lastname}@uni-hamburg.de

THE BENEFIT OF SYNTACTIC VS. LINEAR N-GRAMS FOR LINGUISTIC DESCRIPTION

Our Claims

1. A purely linear approach to language is inadequate.

Our Claims

1. A purely linear approach to language is inadequate.
2. Syntactic n-grams are an alternative representation of language.

Our Claims

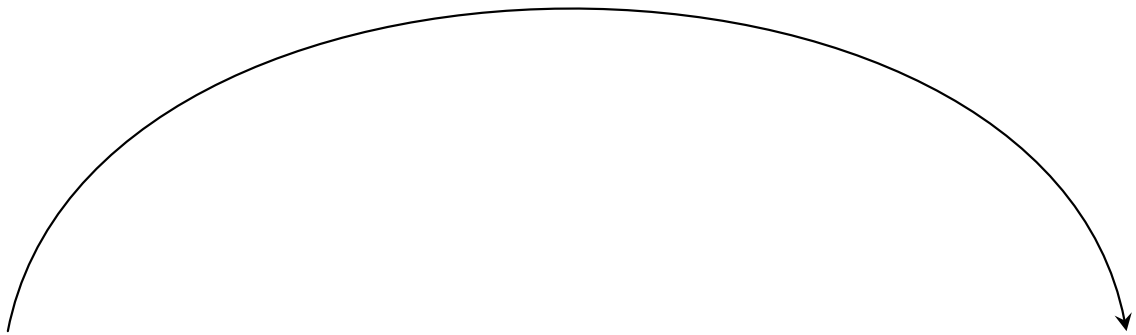
1. A purely linear approach to language is inadequate.
2. Syntactic n-grams are an alternative representation of language.
3. Syntactic n-grams can contribute to stylistic analysis.

Claim One

A purely linear approach to language is inadequate.

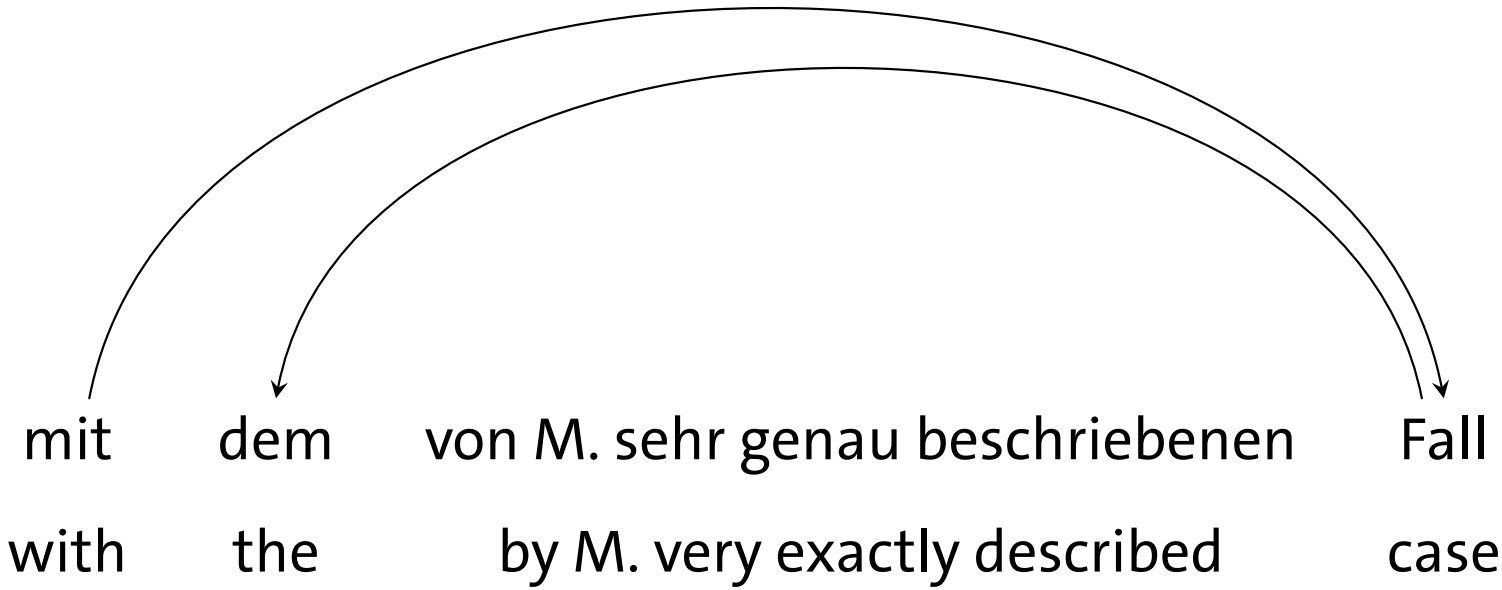
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

German Examples



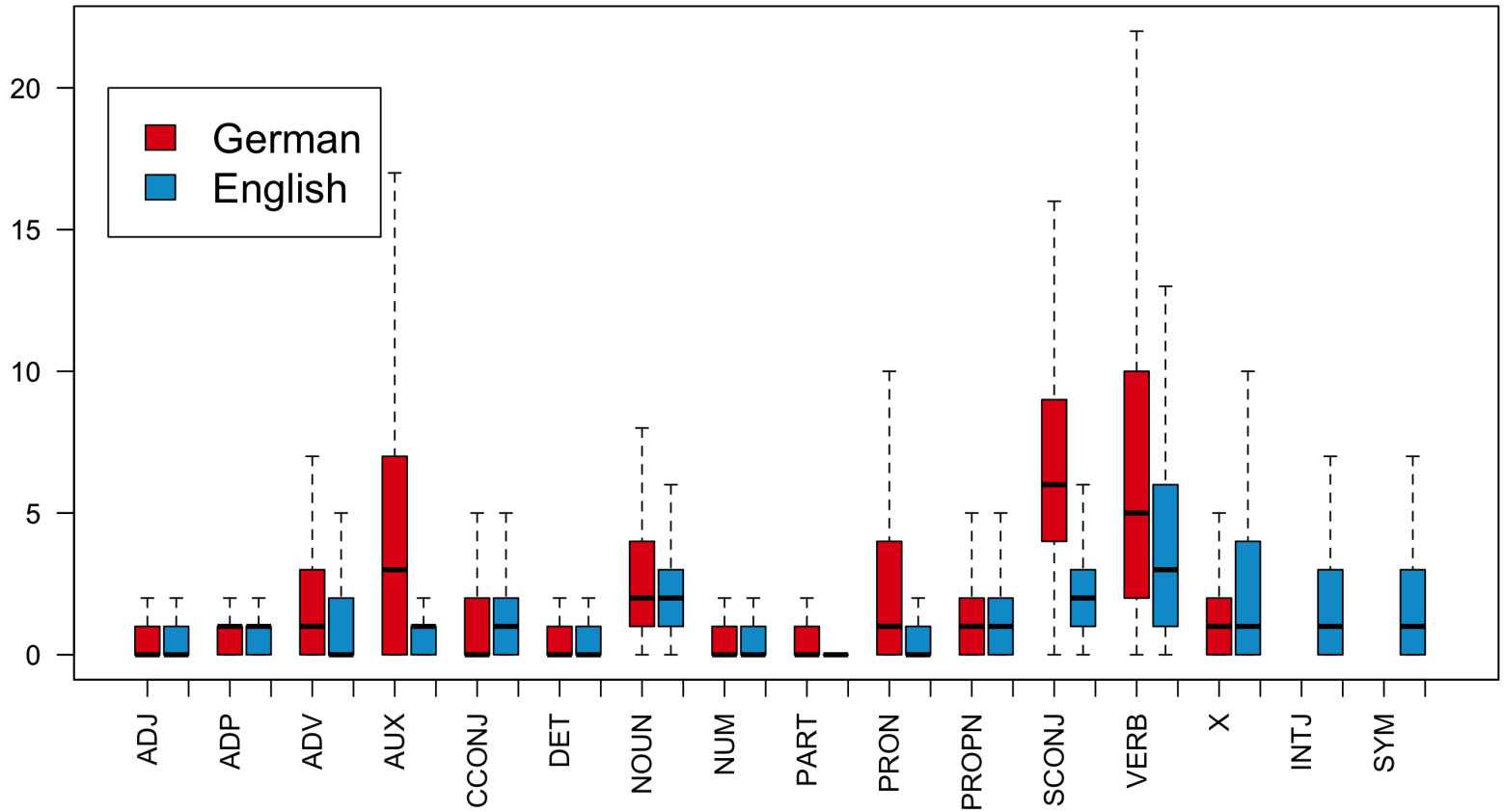
M.	dehnt	den Begriff auf neue Medien	aus
M.	extends	the term to new media	(particle)

German Examples



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

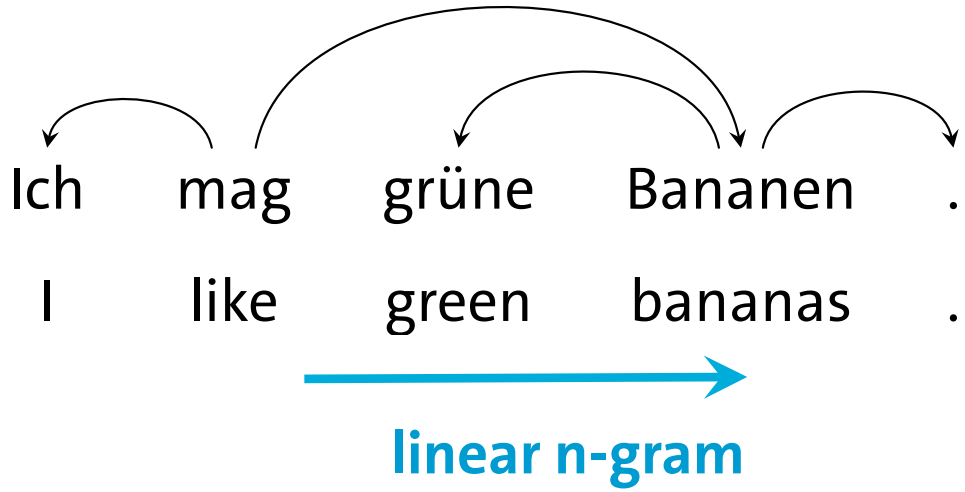
German vs. English (UD)



Claim Two

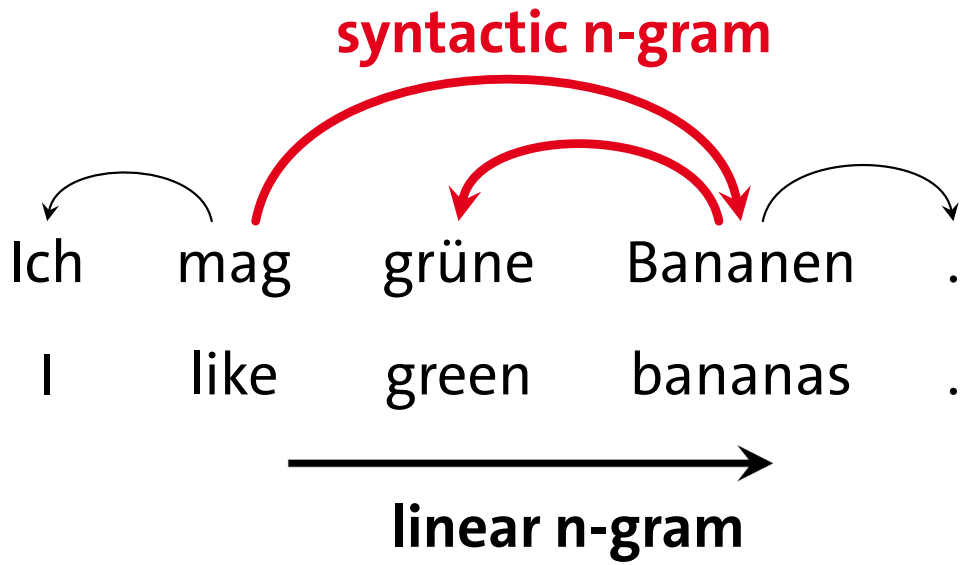
Syntactic n-grams are an alternative representation of language.

Syntactic N-Grams



(Sidorov et al. 2012, Goldberg and Orwant 2013)

Syntactic N-Grams



(Sidorov et al. 2012, Goldberg and Orwant 2013)

Claim Three

Syntactic n-grams can contribute to stylistic analysis.

Case Study

How does the German academic language of linguistics and literary studies differ stylistically?

Data

subcorpus of linguistics
30 PhD theses
1.4 million tokens

vs.

subcorpus of literary studies
30 PhD theses
2.2 million tokens

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Data

subcorpus of linguistics
30 PhD theses
1.4 million tokens

vs.

subcorpus of literary studies
30 PhD theses
2.2 million tokens

automatic annotation of lemma, pos and dependencies (*MATE*, Bohnet 2010, trained on *TIGER Corpus*, Seeker and Kuhn 2012)

N-Gram Analysis

Data sets:

- linear n-grams (token, size 2-5)
- syntactic n-grams (token, size 2-5)

N-Gram Analysis

Data sets:

- linear n-grams (token, size 2-5)
- syntactic n-grams (token, size 2-5)

Questions:

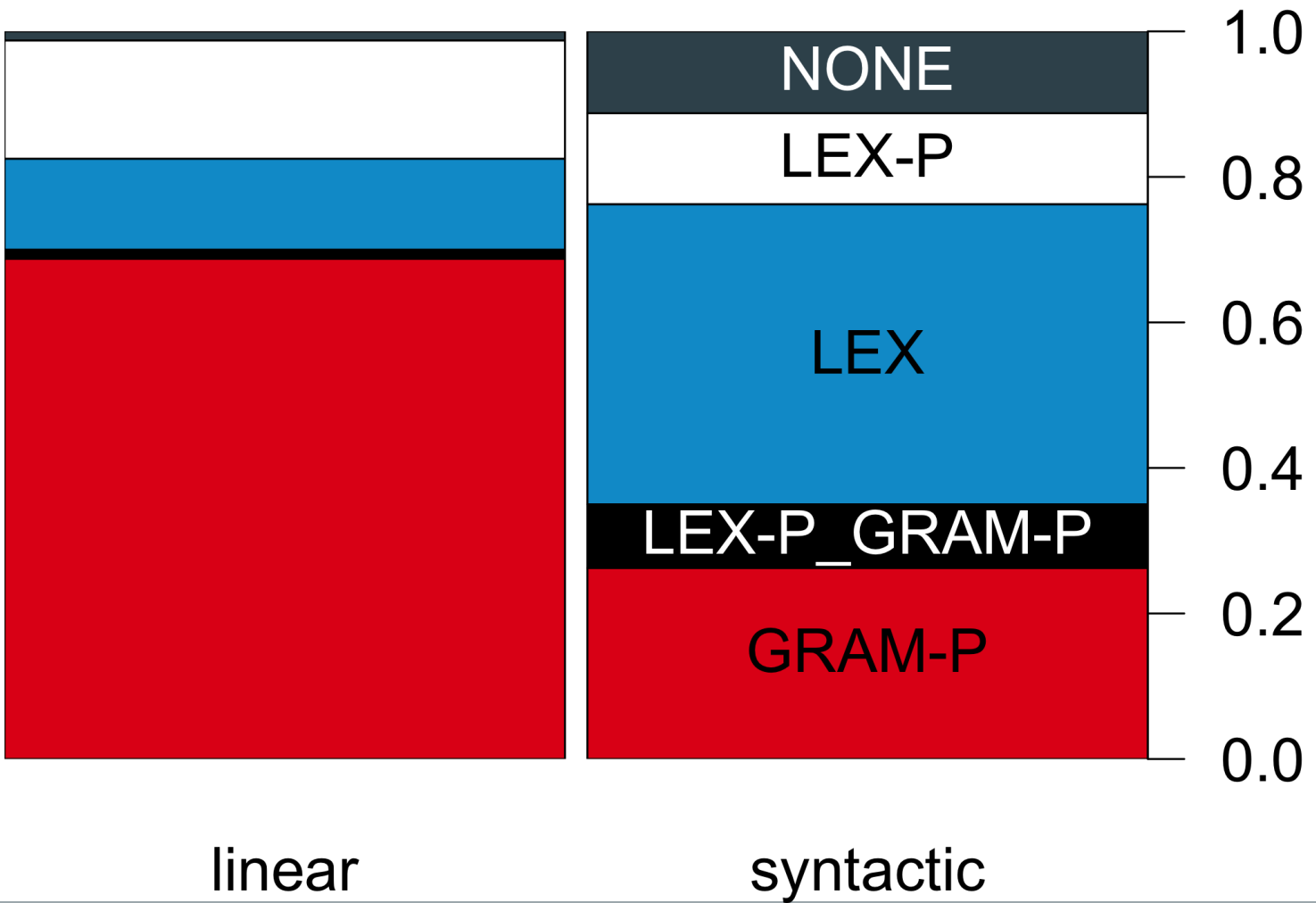
- For which of the n-grams can we attest a significant difference between the two corpora? (based on Welch's t-test)
- Which phenomena are captured by syntactic n-grams only?

Results

Syntactic n-grams capture many phenomena missed by linear n-grams:

- complex verbs (passive voice, modal verbs, particle verbs)
- light verb constructions
- ...

Quantifying Interpretability



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Meet us at our poster!

The Benefit of
Syntactic vs. Linear N-Grams
for Linguistic Description

Melanie Andresen &
Heike Zinsmeister