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

Melanie Andresen & Heike Zinsmeister
{firstname}.{lastname}@uni-hamburg.de
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.
A purely linear approach to language is inadequate.
German Examples

M. dehnt den Begriff auf neue Medien aus

M. extends the term to new media (particle)
German Examples

mit dem von M. sehr genau beschriebenen Fall
with the by M. very exactly described case
German vs. English (UD)
Claim Two

Syntactic n-grams are an alternative representation of language.
Syntactic N-Grams

Ich mag grüne Bananen.
I like green bananas.

linear n-gram

(Sidorov et al. 2012, Goldberg and Orwant 2013)
Syntactic N-Grams

Ich mag grüne Bananen.

I like green bananas.

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

Syntactic n-grams can contribute to stylistic analysis.
How does the German academic language of linguistics and literary studies differ stylistically?
<table>
<thead>
<tr>
<th>Subcorpus of Linguistics</th>
<th>Subcorpus of Literary Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 PhD theses</td>
<td>30 PhD theses</td>
</tr>
<tr>
<td>1.4 million tokens</td>
<td>2.2 million tokens</td>
</tr>
</tbody>
</table>
Data

subcorpus of linguistics
30 PhD theses
1.4 million tokens

subcorpus of literary studies
vs.
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

![Diagram showing the comparison between linear and syntactic models with different interpretability levels: None, LEX-P, LEX, LEX-PGRAM-P, GRAM-P. The diagram illustrates how interpretability increases from the top (None) to the bottom (GRAM-P).]
The Benefit of Syntactic vs. Linear N-Grams for Linguistic Description

Melanie Andresen & Heike Zinsmeister