Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and Its Cross-Corpus Validation

Zarah Weiss & Detmar Meurers
University of Tübingen, ICALL Research Group (http://icall-research.de)

Introduction

We are presenting an empirically broad cross-corpus analysis of German educational media language targeting adults and children.

Readability assessment refers to the task of linking a text to the appropriate target audience based on its complexity.

Potential application domains include: the design and evaluation of education materials, information retrieval, and text simplification.

While for English reference corpora for cross-corpus testing are available (Common Core, WeeklyReader), there are no similar resources for German.

We illustrate how empirically broad models successfully generalize across German educational media corpora.

Automatic Complexity Analysis

We automatically extract 400 complexity measures covering clausal, phrasal, lexical, morphological, and discourse complexity, cognitive complexity, and language use.

The features are theoretically grounded in SLA research, where Complexity is defined in terms of elaborateness and variation of language (Housen et al., 2012).

This is to our knowledge currently the most extensive feature set for German; for details, see Weiss & Meurers (accepted).

The pipeline will be integrated into CTAP (Chen & Meurers, 2016) by the end of fall 2018.

Informativeness of Complexity Measures

We calculated the information gain (IG) of each feature for the target audience with 10-fold cross-validation (10-fold CV).

When we rank all features by IG and compare the top 20 features with an Pearson inter-correlation $r \leq 0.5$, we see that

- language use and cohesion measures cover overall 55% of the top 20 features for both data sets.
- also measures of phrasal, sentential, lexical, and morphological complexity are important.

Bilingual Readability Models

We train linear SMO classifiers using WEKA (Hall et al., 2009) and evaluate them with 10-fold CV and cross-corpus testing.

Training on GEO/GEOlino yields high performance within and cross-corpus testing, but training on Tagesschau/Logo does not generalize to GEO/GEOlino.

This difference is not due to the larger corpus size of GEO/GEOlino: also GEO/GEOlino generalizes well across corpora while Tagesschau/Logo does not.

Data

We compiled two corpora of German news media for adults and children:

- GEO/GEOlino: online articles from the leading German monthly educational magazine GEO and its children counterpart GEOlino (similar to Hancke et al. (2012)).
- Tagesschau/Logo: official subtitles of the German daily news broadcasts for adults (Tagesschau) and children (Logo).

All subtitles and all article links are available for research.

Readability Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Testing</th>
<th>Features</th>
<th>Acc.</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>GEO/GEOlino</td>
<td>50.0</td>
<td>GEO/GEOlino</td>
<td>50.0</td>
<td></td>
</tr>
<tr>
<td>10-fold CV</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td></td>
</tr>
<tr>
<td>Cross-Corpus</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td></td>
</tr>
<tr>
<td>Balanced CC</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td>GEO/GEOlino</td>
<td>89.4±0.9</td>
<td></td>
</tr>
</tbody>
</table>

Summary & Outlook

- We designed a highly accurate model which i) is based on a very broad coverage of linguistic features, and ii) successfully generalizes across corpora and text types.
- We show that German educational media language is successfully and broadly adapted towards their target audiences, unlike, e.g., German school textbooks (Berendes et al., 2017).
- Next steps include comparing the data to language produced by children belonging to the target group (linguistic complexity ≤ proficiency, see Weiss & Meurers (accepted))

References


