Readability Classification for German using Lexical, Syntactic, and Morphological Features

Julia Hancke  Sowmya Vajjala  Detmar Meurers
University of Tübingen, Germany

COLING 2012, Mumbai, India
Outline of the talk

- Introduction to Readability Classification
- German Readability Classification
- Approach: Features, Experiments & Results
- Conclusions & Future directions
Introduction

What is “readability assessment”?  

Measuring **how difficult it is to read a text**, 
- given a **purpose**, e.g., 
  - identifying appropriate texts for the readers 
  - evaluation of natural language generation systems 
- based on **properties of the text** using criteria which are 
  - theory-driven (e.g., difficult syntactic constructions) 
  - data-induced (e.g., corpora with graded texts) 
  - in-between (e.g., derived frequency information for words) 
- for a **user**, given some information about him (e.g., language ability, age, working memory) 
  - obtained directly (e.g., questionnaire), or 
  - indirectly (e.g., inferred from nature of a search query)
How do we measure text complexity?

Traditional Approaches – English

- Different aspects of linguistic complexity play a role in determining the readability of a text.

- **Traditional readability formulae** for English used shallow quantitative features
  - e.g., average sentence length.
  - Kincaid et al. (1975), DuBay (2004), . . .

- Some others were based on **lexical measures**
  - e.g., words belonging to specific word lists
How do we measure text complexity
Some recent CL Approaches

- Language n-gram models
  (Collins-Thompson & Callan 2004; Si & Callan 2001; Petersen & Ostendorf 2009)

- Machine learning using lexical and syntactic features
  (Heilman et al. 2007; Schwarm & Ostendorf 2005; Feng 2010)

- Language specific morphological features
  Italian: (Dell’Orletta et al. 2011); French: (François & Fairon 2012)

- Features motivated by cognitive perspective
  (Crossley & McNamara 2011)

- Features motivated by Second Language Acquisition (SLA)
  - measures of language development/proficiency can successfully be reused as measures of text complexity
    (Vajjala & Meurers 2012)
How do we measure text complexity
Non-English Readability Classification

- Multi-level classification using text-book corpora
  French: (François & Fairon 2012), Japanese: (Sato et al. 2008)

- Two level classification using publicly available corpora
  Italian: (Dell’Orletta et al. 2011); Portuguese: (Aluisio et al. 2010)
  Danish: (Klerke & Søgaard 2012)

- The DeLite readability checker for German
  (Vor der Brück & Hartrumpf 2007; Vor der Brück et al. 2008)
  - Human annotated corpus of 500 texts, classified into ten levels of reading difficulty
  - Lexical, syntactic, semantic and discourse features
  - Texts are specific to administrative and legal domains
  - Relatively small for building a classification model
Our aims in this paper

- Perform two-class readability classification for German
  - Creating a corpus of "easy" and "difficult" texts
  - Using lexical, syntactic and morphological features
- Explore the use of a rich feature set for German, including
  - features used in research on English
  - SLA measures of complexity
    - elaborateness
    - variability
  - features making use of German morphology
Our Approach
The Corpus

- We created a two class text corpus (*easy vs. difficult*) from two websites containing articles on similar topics:
  - Website for children: GEOlino (http://www.geolino.de)
  - Website for adults: GEO (http://www.geo.de)

- Educational monthly magazines covering articles related to technology, nature etc.

- GEOlino not a simplified version of GEO.
  - Content is created specifically for child readers.
Our Approach

Experimental Setup

- Classifier: Sequential Minimal Optimization Algorithm (SMO) as implemented in WEKA
- Dataset: 1800 texts
  - 900 from GEOlino website (easy)
  - 900 from GEO website (hard)
- Evaluation: 10-fold cross-validation
  - random baseline for binary classification: 50%
Features Used

Lexical Features

- Traditional lexical features
  - e.g., avg. num. characters per word
- Lexical richness features from SLA (Lu 2011)
  - Type-token ratio variants.
  - Lexical variation (e.g., noun variation)
- Other lexical features
  - e.g., verb token ratio $= \frac{Tok_{\text{Verb}}}{Tok}$
- German specific lexical features
  - e.g., sein to verb token ratio

⇒ Classification accuracy: 82.1%.
Features Used

Syntactic Features

- General parse tree based features
  - e.g., # NPs/sentence, avg. length of a PP
- German specific feature
  - # zu-marked infinitives per phrase.
- SLA based syntactic complexity features (Lu 2010)
  - Ratios that capture embedding, co-ordination
    - e.g., # co-ordinate phrases per sentence
  - Lengths of various production units
    - e.g., avg. length of a sentence, clause.
  - Relationships between specific structures
    - e.g., # verb phrases per T-Unit

⇒ Classification accuracy: 76.8%
Features Used

Language Modeling Features

- Used a separate dataset for building language models.
  - *Easy model*: News4Kids (www.news4kids.de)
  - *Difficult model*: NTV (www.n-tv.de)

- We trained two types of language models:
  - Word models
  - Mixed models: words + parts of speech
    - Replacing words with parts of speech: Information Gain

- Perplexity scores for uni-, bi- and tri-gram word and mixed models were used to train a classifier.

⇒ Classification accuracy: 77.6%
German Morphology

German morphology: 1) morphosyntax and 2) word formation.

1) Inflection
   - Verb endings indicate e.g., person and number
     - I buy vs. he buys
   - Case information is carried by articles/nouns/adjectives.

2a) Derivation
   - Nominalizations with a suffix e.g.,
     - *regieren* – *regierung*
       govern government
   - or without a suffix e.g.,
     - *laufen* – *to run*
       der lauf the run

2b) Compounding
   - *Stadtlaufl* – City Run
     Stadt – city; lauf – run
Features Used

Inflectional Morphology Features

➤ Verbal tense/mood might be indicators of text difficulty.
  ➤ e.g., # infinitive verbs / # verbs.
  ⇒ Classification accuracy: 67.2%

➤ Nominal case information might also be useful.
  ➤ e.g., # accusative nouns / # nouns.
  ⇒ Classification accuracy: 74.3%

⇒ Classification accuracy: 79.0%
Features Used

Derivational Morphology Features

- We compiled a list of 25 native and foreign suffixes (Fleischer & Barz 1995).

- For each suffix, we included all of the different gender and number forms.

- Calculated the following ratios for each suffix:
  - Suffix token ratio (# instances of suffix / # tokens)
  - Suffix noun ratio (# instances of suffix / # nouns)
  - Suffix derived noun ratio (# instances of suffix / # derived nouns)

- Additional feature: derived nouns to nouns ratio.

⇒ Classification accuracy: 78.5%.
We considered two simple compounding features.

- Ratio of compound nouns to all nouns.
- Average number of words in a compound

⇒ Classification accuracy: 57.0%
### Features used

#### Morphological Features – Summary

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Num. Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivational</td>
<td>76</td>
<td>78.5%</td>
</tr>
<tr>
<td>Infl. Nouns</td>
<td>4</td>
<td>67.2%</td>
</tr>
<tr>
<td>Infl. Verbs</td>
<td>13</td>
<td>74.3%</td>
</tr>
<tr>
<td>Infl. Nouns + Infl. Verbs</td>
<td>17</td>
<td>79.0%</td>
</tr>
<tr>
<td>Compunding</td>
<td>2</td>
<td>57.0%</td>
</tr>
<tr>
<td>All Morph.</td>
<td>95</td>
<td>85.4%</td>
</tr>
</tbody>
</table>
Top-10 Predictive Features
Feature set determined by Information Gain

<table>
<thead>
<tr>
<th>Feature</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ Avg. word length</td>
<td>Lex/Trad</td>
</tr>
<tr>
<td>→ # 2nd person Vs / # finite Vs</td>
<td>Morph</td>
</tr>
<tr>
<td># syllables per word</td>
<td>Lex/Trad</td>
</tr>
<tr>
<td># 3rd person Vs / # finite Vs</td>
<td>Morph</td>
</tr>
<tr>
<td>→ Avg. length of a T-unit</td>
<td>Syn</td>
</tr>
<tr>
<td>Avg. length of a sentence</td>
<td>Syn/Trad</td>
</tr>
<tr>
<td>→ # Complex nominals per clause</td>
<td>Syn</td>
</tr>
<tr>
<td># Complex nominals per T-unit</td>
<td>Syn</td>
</tr>
<tr>
<td>→ # PPs per sentence</td>
<td>Syn</td>
</tr>
<tr>
<td>Avg. length of a clause</td>
<td>Syn</td>
</tr>
</tbody>
</table>
Traditional Features as a Baseline

Features from traditional readability formulae
- avg. num. characters per word
- avg. num. syllables per word
- avg. num. words per sentence

⇒ Classification accuracy : 82.2%.
## Overall result summary

A performance comparison of all five feature groups

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Num. Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional (baseline)</td>
<td>3</td>
<td>82.2%</td>
</tr>
<tr>
<td>Lexical</td>
<td>23</td>
<td>82.1%</td>
</tr>
<tr>
<td>Syntactic</td>
<td>26</td>
<td>76.8%</td>
</tr>
<tr>
<td>Morphology</td>
<td>95</td>
<td>85.4%</td>
</tr>
<tr>
<td>Language Modeling</td>
<td>12</td>
<td>77.6%</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td><strong>155</strong></td>
<td><strong>89.7%</strong></td>
</tr>
<tr>
<td><strong>TOP 10</strong></td>
<td><strong>10</strong></td>
<td><strong>84.3%</strong></td>
</tr>
</tbody>
</table>

⇒ The best result is 7.5% better than the baseline model.
⇒ Error reduction of 42%
Conclusions

- Our approach provides accurate readability classification for German with a broad range of features.
  - Features from English successfully transferred to German.
  - Measures from Second Language Acquisition research are useful as features for readability classification.
  - The German-morphology features proved to be good indicators as an individual feature group.
- Obtained very good accuracy figures, but as importantly:
  - supports comprehensive modeling of underlying properties
  - cannot easily be tricked by surface modifications
Future Directions

- Generalizability of the features
  - Will these readability models generalize across datasets?
    - e.g., classifying documents on the web.

- Towards simplification
  - Which features are useful at a paragraph or sentence level to identify targets for simplification?
Thank you for your attention!
Questions? :-)

Readability Classification for German using Lexical, Syntactic, and Morphological Features
Julia Hancke, Sowmya Vajjala, Detmar Meurers

Introduction
Outline
Background

Related Work
Traditional Formulae
Computational Approaches
Non-English Languages

Our Approach
Features
Lexical Features
Syntactic Features
Morphological Features
Top-10 Features
Traditional Readability Features

Results
Conclusions
Outlook

References
References


