Exploring CEFR classification for German based on rich linguistic modeling

Julia Hancke  Detmar Meurers
Universität Tübingen

Learner Corpus Research Conference (LCR 2013)

Introduction

The Common European Framework of Reference for Languages (CEFR) is an increasingly used standard for characterizing the foreign language ability of a learner based on functional abilities to use language in different domains (public, private, occupational, etc.).

But there is a lack of

- authentic learner data illustrating CEFR levels and insight into the precise linguistic characteristics correlating with the proficiency levels.

Data used: German portion of MERLIN corpus

- 1027 German learner texts
  - about 200 texts per exam type (A1–C1)
  - range of lengths (6–366 words) with average 122 words
  - texts also vary in other parameters:
    - written for different tasks (one of three tasks per level)
    - written by learners with different native languages (> 12)

- Each text was graded in terms of CEFR levels
  - by multiple trained human raters at TELC, a major language test provider in Germany
  - reliability of ratings externally validated (Univ. Leipzig)
  - most common rating: B1
Features to be investigated

- Goal: richer linguistic modeling of CEFR levels
  - explore potentially relevant language features
  - test their impact on predicting CEFR class of each essay

- We explored:
  - lexical features
  - syntactic features
    - statistical language model
    - constituency-based
    - dependency-based
  - morphological features

Features explored

Lexical features

- Lexical density (Lu 2012)
  - ratio of number of lexical words to total number of words

- Lexical diversity:
  - TTR variants, MTLD, lexical word variation
    (McCarthy & Jarvis 2010; Crossley et al. 2011a; Lu 2012)

- Depth of lexical knowledge
  - lexical frequency scores (Crossley et al. 2011b)

- Lexical relatedness
  - hypernym & polysemy scores (Crossley et al. 2009)

- Shallow measures
  - spelling errors per number of words, word length

Syntactic features: 1. Statistical Language Models

- inspired by readability assessment research
  (Schwarm & Ostendorf 2005; Petersen & Ostendorf 2009; Feng 2010)

- used SRILM Language Modeling Toolkit (Stolcke 2002)

- trained on two data sets (Hancke, Meurers & Vajjala 2012)
  - easy: 2000 texts, German kid news website News4Kids
  - hard: 2000 texts, German news channel NTV website

- 12 features: unigram, bigram and trigram perplexity for
  - easy or hard text models based on
  - word or mixed (word+POS) representations
Features explored
Syntactic features: 2. Data-driven constituency features

- Is the frequency of common rules characteristic? (Briscoe et al. 2010; Yannakoudakis et al. 2011)
- Extracted all rules in the parse trees assigned by Stanford Parser in 700 articles from the NTV corpus

\[
\begin{align*}
S & \\
| & \\
NP & VP \\
| & \\
NNP & VPZ ADJP & | & \text{Norway} & | & \text{is} & | & \text{beautiful}
\end{align*}
\]

- Given a learner text, for each rule, we use as feature: \textit{rule frequency in text / number of words in text}

Features explored
Syntactic features: 3. Theory-driven constituency features (Hancke, Meurers & Vajjala 2012)

Syntactic properties assumed to be characteristic of complexity or difficulty in SLA proficiency and readability research:

- number and length of clauses, sentences, T-units
- NPs, VPs, PPs
- dependent clauses and coordinated phrases
  - per clause, sentence, T-unit
- interrogative, relative, conjoined clause ratios
- nonterminals per sentence
- parse tree height

Features explored
Morphological features

Linguistic properties based on dependency analysis used in SLA proficiency and readability assessment research:

- number of words between head and dependent
  - maximum
  - average number per sentence
- avg. number of dependents per verb (in words)
- number of dependents per NP (in words)
NLP used for automatic feature identification

- **Preprocessing**
  - sentence segmentation, tokenization (Apache OpenNLP)
  - spelling correction (Java API for Google Spell Check)

- **Lexicon**
  - lexical semantic relations (GermaNet, Hamp & Feldweg 1997)
  - lexical frequencies (dlexDB, http://dlexdb.de)

- **Part-of-Speech Tagging**
  - POS and lemmatization (TreeTagger, Schmid 1995)
  - fine-grained POS (RFTagger, Schmid & Laws 2008)

- **Parsing**
  - constituents (Stanford PCFG Parser, Rafferty & Manning 2008)
  - dependencies (MATE, Bohnet 2010)

---

**Experimental Setup**

- We divided the MERLIN data into
  - training set (721 essays)
  - test set (302 essays)

- We classify into five CEFR classes (A1, A2, B1, B2, C1).

- We use the WEKA machine learning toolkit (Hall et al. 2009) for classification, specifically
  - SMO to train support vector machines (linear kernel)

- Many further experiments → Hancke (2013)

---

**Performance of different feature groups**

<table>
<thead>
<tr>
<th>Name</th>
<th>#</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>-</td>
<td>20.0</td>
</tr>
<tr>
<td>Majority Baseline</td>
<td>-</td>
<td>33.0</td>
</tr>
<tr>
<td>TENSE</td>
<td>230</td>
<td>38.5</td>
</tr>
<tr>
<td>ParseRules</td>
<td>3445</td>
<td>49.0</td>
</tr>
<tr>
<td>LanguageModel</td>
<td>12</td>
<td>50.0</td>
</tr>
<tr>
<td>SYN</td>
<td>47</td>
<td>53.6</td>
</tr>
<tr>
<td>MORPH</td>
<td>41</td>
<td>56.8</td>
</tr>
<tr>
<td>LEX</td>
<td>46</td>
<td>60.5</td>
</tr>
</tbody>
</table>

- Informative – but for this data set:
  - Text Length as a single feature: 61.4% accuracy

---

**Feature Groups Combinations**

The best two, three, and four class combinations:

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEX_MORPH</td>
<td>61.1</td>
</tr>
<tr>
<td>LEX_TEN</td>
<td>59.8</td>
</tr>
<tr>
<td>LEX_LM</td>
<td>59.4</td>
</tr>
<tr>
<td>LEX_LM_MORPH</td>
<td>61.1</td>
</tr>
<tr>
<td>SYN_Lex_MORPH</td>
<td>58.5</td>
</tr>
<tr>
<td>LEX_LM_TEN</td>
<td>57.8</td>
</tr>
<tr>
<td>SYN_Lex_LM_MORPH</td>
<td>58.8</td>
</tr>
<tr>
<td>SYN_Lex_LM_PR</td>
<td>57.8</td>
</tr>
<tr>
<td>LEX_LM_MORPH_TEN</td>
<td>57.8</td>
</tr>
<tr>
<td>ALL Features</td>
<td>57.2</td>
</tr>
</tbody>
</table>

- not particularly exciting, but lexical features help
Feature Selection

▶ How can we identify the best features?
▶ The features we use are not independent, so taking the best features using Information Gain is problematic.
▶ CfsSubsetEval: correlation-based feature selection
  ▶ Features that correlate highest with the class but have a low inter-correlation are preferred (Witten & Frank 2005).
▶ Results:

<table>
<thead>
<tr>
<th>Name</th>
<th>#</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CfsSubsetEval(LEX,LM,MORPH)</td>
<td>30</td>
<td>61.7</td>
</tr>
<tr>
<td>CfsSubsetEval(SYN,LEX,LM,MORPH)</td>
<td>34</td>
<td>62.7</td>
</tr>
<tr>
<td>CfsSubsetEval(ALL)</td>
<td>88</td>
<td>61.8</td>
</tr>
</tbody>
</table>

Qualitative analysis of the 34 selected features

Syntax

▶ sophistication of production units
  ▶ avg. sentence length, length of a t-unit
▶ embedding
  ▶ dep. clause with conj. to dep. clause ratio
▶ verb phrase complexity
▶ coordination
▶ passive voice
▶ text length

Lexicon

▶ spelling errors
▶ lexical richness (TTR, MTLD)
▶ verbal/nominal style (verb variation, noun token ratio)
▶ lexical sophistication (frequency, easy unigrams, length)
▶ but: no lexical relatedness features were selected

Morphology

▶ use of derivation (derived nouns/nouns, specific suffixes)
▶ nominal case (genitive, nominative)
▶ verbal mood and person (subjunctive, 2. person forms)
▶ Automatic proficiency classification: a useful experimental sandbox for exploring the role of linguistic modeling
▶ Quantitatively difficult but possible to outperform the very high text-length baseline on the new MERLIN corpus.
▶ Qualitatively insightful analysis of features is possible.
▶ Feature selection helps improve classification results and identify qualitatively interpretable feature groups.

**Outlook:**
- reliable sentence segmentation for learner language needed, crucial for many complexity features
- analyze impact of learner errors on such analyses, possible using target hypotheses
- principled exploration of variationist linguistic features (→ talk on Saturday with Julia Krivanek)
## Qualitative analysis of selected features

### Detailed Lexicon

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical richness</td>
<td>type-token ratio, root type-token ratio, corrected type-token ratio, HDD, MTLD</td>
</tr>
<tr>
<td>lexical richness w. respect to verbs</td>
<td>squared verb variation 1, corrected verb variation 1</td>
</tr>
<tr>
<td>nominal style</td>
<td>noun token ratio</td>
</tr>
<tr>
<td>word length / difficulty</td>
<td>avg. num. syllables per word, avg. num. characters per word</td>
</tr>
<tr>
<td>lexical sophistication</td>
<td>annotated type ratio, unigram plain easy ratio of words in log frequency band two, ratio of words in log frequency band four</td>
</tr>
<tr>
<td>spelling errors</td>
<td>ratio of lex. types not in Dlex, Google spell check error rate</td>
</tr>
</tbody>
</table>

### Detailed Morphology

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominalization, use of derivational suffixes and words with Germanic stems</td>
<td>nominal case</td>
</tr>
<tr>
<td>nominal case</td>
<td>genitive-noun ratio, nominative-noun ratio</td>
</tr>
<tr>
<td>verbal mood and person</td>
<td>subjunctive-verb ratio, second person-verb ratio, third person-verb ratio</td>
</tr>
</tbody>
</table>

---

### NLP used for feature identification

- Lexical
- Syntactic
- Language Model
- Constituency
- Dependency
- Morphological
- NLP used for feature identification

---

### Summary

- Merlin
- LifeLong Learning Programs
- University of Tübingen