Exploring linguistic complexity in readability analysis and L2 development

Detmar Meurers
Universität Tübingen

based on joint research with
Sowmya Vajjala and Xiaobin Chen

Oberseminar English Linguistics
Universität Frankfurt
January 9, 2017

What is readability analysis?

We want to determine how difficult it is to read
- a given text
- for a given purpose, e.g.,
  - skimming for information
  - answering comprehension questions
- for a given individual reader with
  - their knowledge of the topic domain
  - individual differences in cognition, affect, personality

⇒ Which characteristics of the texts can we consider?

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From tests to sentences
Wikipedia–SimpleWiki

Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Traditional approaches to readability

- Long history of readability formulas (DuBay 2004)
  - Developed for specific purposes, e.g., characterizing demands of military training manuals (Caylor et al. 1973)
- Formulas are based on shallow, easy to count features:
  - typically avg. sentence length and avg. word length, e.g., Flesch-Kincaid formula (Kincaid et al. 1975)
  - counts of words on specific word lists (Dale & Chall 1948)
- Problems of traditional readability formulas:
  - based on rough generalizations:
    - long words are rare, long sentences are difficult
  - formulas are domain dependent
  - provide only a quantitative measure, not a characterization of the language aspects involved in readability
What can we observe about a given text?

I. Which language **forms** are used, how are they combined?
   - **type** of forms in the **linguistic system**
     e.g.: complex NPs per sentence
   - **use** of forms in terms of personal **language experience**,
     evidence via proxy of representative language records
     e.g.: word frequency, average AoA

II. What type & amount of **meaning** is encoded by the forms,
    and how is it organized into a coherent discourse?
    e.g.: concreteness, lexical density, referential cohesion

III. What are its demands on **human processing**?
    e.g. memory load for referents, surprisal

SLA measures of proficiency development

- Second Language Acquisition research has developed a
  rich inventory of measures for monitoring development.
- Skehan (1989) characterized proficiency in terms of the
  three dimensions Complexity, Accuracy, und Fluency
  (CAF, Wolfe-Quintero et al. 1998; Housen & Kuiken 2009)
- **Complexity:**
  *The extent to which the language produced in*
  *performing a task is elaborate and varied.*
  (Ellis 2003, p. 340)

Connecting Readability and L2 Complexity

- How about making use of
  - SLA measures of the **complexity** of learner language
    for determining the **readability** of native texts?
- **Motivation:**
  - profit from rich set of SLA measures operationalizing
    complexity at all levels of linguistic modeling
  - using the same features to characterize reading texts
    and language proficiency can make it possible to
    - tailor complexity of input to learner proficiency (i+1)
- **Putting the idea to the test:**
  - Chen & Meurers (2016a,b)
Testing how well the idea works
A supervised machine learning setup as experimental sandbox

- Take a corpus of texts for which reading levels are known.
- Spell out hypotheses which properties matter as features.
- Train a machine learning model.
  - classification: discrete levels (e.g., beg., int., adv.)
  - regression: continuous levels (e.g., age)
  - ranking: relative level (which of two is easier)
- Evaluate model by predicting levels of unseen texts.

Features from SLA research

Lu (2010, 2011, 2012) surveyed complexity features used in SLA research, which we select many of our features from.

**Lexical Features**

- **Lexical Variation**
  - Type-Token Ratio = Typ/Tok
  - influenced by text length
- **Measure of Textual Lexical Diversity (MTLD, McCarthy 2005)**
  - average number of words needed to reach stable TTR point
- **Lexical Density** = TokLex / Tok
- Lex = open lexical classes (N, Adj, Adv, V)
- Overall we use: 19 lexical features (16 SLA, 3 others)

Syntactic complexity features

- analyze different units: sentences, T-units, clauses
  - mean length per unit
    - e.g., mean length of sentences
  - number of occurrences per unit
    - e.g., number of clauses per sentence
  - ratios of different subtypes (subordination, coordination)
    - e.g., dependent clauses per clause, . . .
  - specific constructions
    - e.g., complex nominals per clause, . . .
- Overall we use: 25 syntactic features (14 SLA, 11 others)

Corpus

- Needed: a corpus with gold-standard labels
- Previous work: graded reading material in WeeklyReader
- We compiled the WeeBit corpus (Vajjala & Meurers 2012):

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Age in Years</th>
<th>Number of Articles</th>
<th>Avg. Number of Sentences/Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>from WeeklyReader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>7–8</td>
<td>629</td>
<td>23.41</td>
</tr>
<tr>
<td>Level 3</td>
<td>8–9</td>
<td>801</td>
<td>23.28</td>
</tr>
<tr>
<td>Level 4</td>
<td>9–10</td>
<td>814</td>
<td>28.12</td>
</tr>
<tr>
<td>from BBCBitesize</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KS3</td>
<td>11–14</td>
<td>644</td>
<td>22.71</td>
</tr>
<tr>
<td>GCSE</td>
<td>14–16</td>
<td>3500</td>
<td>27.85</td>
</tr>
<tr>
<td>Results on WeeBit (5 classes)</td>
<td>Ten Best Features (Information Gain)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>WeeklyReader: Previous Work</strong></td>
<td>▶ Half of the best features are SLA complexity measures:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feng (2010)</td>
<td>▶ mean length of a sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>122</td>
<td>▶ dependent clause to clause ratio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74.0%</td>
<td>▶ complex NPs per clause</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50.9%</td>
<td>▶ modifier variation (proportion of adjectives &amp; adverbs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ adverb variation (proportion of adverbs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petersen &amp; Ostendorf (2009)</td>
<td>▶ The other features in the Top 10:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>▶ avg. num. characters per word</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63.2%</td>
<td>▶ avg. num. syllables per word</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. &amp; O. syntactic features only</td>
<td>▶ proportion of words on Academic Word List</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>▶ num. co-ordinate phrases per sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50.9%</td>
<td>▶ Coleman-Liau score</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Word lists: Academic Word List</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WeeklyReader (Vajjala &amp; Meurers 2012)</td>
<td>▶ SLALEX + SLASYN:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication P. &amp; O. syntactic feat.</td>
<td>▶ 82.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>▶ 0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64.3%</td>
<td>▶ SLALEX:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>▶ 81.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>84.1%</td>
<td>▶ SLASYN:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>▶ 0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91.3%</td>
<td>▶ All our Features:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>▶ 0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68.1%</td>
<td>▶ Best10Features:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>▶ 0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71.2%</td>
<td>▶ 0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>▶ 0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>82.3%</td>
<td>▶ 0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>▶ 0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>93.3%</td>
<td>▶ 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>▶ 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>89.7%</td>
<td>▶ 0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploring
linguistic complexity
in readability analysis
& L2 development
Detmar Meurers

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit

Realization of the extended feature set
(Vajjala & Meurers 2014a)

▶ Resources:
- Celex Lexical Database (http://celex.mpi.nl)
- Kuperman et al. (2012)’s AoA ratings
- MRC Psycholinguistic database (http://ota.oucs.ox.ac.uk/headers/1054.xml)
- Wordnet (http://wordnet.princeton.edu)

▶ Tools:
- Features computed using:
  - Stanford Tagger (Toutanova, Klein, Manning & Singer 2003)
  - Berkeley Parsec (Petrov & Klein 2007)
  - Tregex Pattern Matcher (Levy & Andrew 2006)
- Machine Learning using WEKA
  - SMORReg algorithm (modeling readability as regression) trained on WeeBit corpus

Results on standard CCSS corpus

▶ Common Core State Standards reading initiative of the U.S. education system (CCSSO 2010)
▶ Reference corpus: 168 texts for grade levels 2–12
▶ Results (Spearman’s rank correlation since scales differ):

<table>
<thead>
<tr>
<th>System</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelson et al. (2012):</td>
<td>0.54</td>
</tr>
<tr>
<td>REAP</td>
<td>0.59</td>
</tr>
<tr>
<td>ATOS</td>
<td>0.53</td>
</tr>
<tr>
<td>DRP</td>
<td>0.50</td>
</tr>
<tr>
<td>Lexile</td>
<td>0.69</td>
</tr>
<tr>
<td>Reading Maturity</td>
<td></td>
</tr>
<tr>
<td>ETS SourceRater</td>
<td>0.75</td>
</tr>
<tr>
<td>Vajjala &amp; Meurers (2014a)</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Do the results generalize?
(Vajjala & Meurers 2014c)

▶ Does the WeeBit model generalize to other datasets?

<table>
<thead>
<tr>
<th>System</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonCore</td>
<td>0.69</td>
</tr>
<tr>
<td>TASA corpus</td>
<td>0.86</td>
</tr>
</tbody>
</table>

▶ Impact of genre differences in CommonCore data:

<table>
<thead>
<tr>
<th>Genre in CommonCore</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informative</td>
<td>0.76</td>
</tr>
<tr>
<td>Misc.</td>
<td>0.69</td>
</tr>
<tr>
<td>Literature</td>
<td>0.51</td>
</tr>
<tr>
<td>Speech</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Is the feature set informative enough for spoken language?
Readability analysis of TV subtitles
(Vajjala & Meurers 2014b)

- We used our feature set to train a model that identifies age-specific TV programs.
- Data: subtitles of BBC TV channels (Van Heuven et al. 2014)
- Classification into three age groups:
  - less than 6, 6–12, adult
  \[ \Rightarrow \] 96% classification accuracy (SMO, 10 fold CV)
  - single most predictive feature: average AoA of words, but accuracy is not reduced if this feature is removed
  - Classification is informed by a wide range of linguistic elaborateness, variedness, and cognitive characteristics.

From texts to sentences

- Can we reliably analyze individual sentences?
- This would be useful
  - for text simplification
    - to identify targets for simplification
    - to evaluate aspects of simplification
  - to evaluate sentences in questionnaires
  - to rank candidates in generation systems
  \[ \Rightarrow \] Test model trained on WeeBit texts on individual sentences

Effect of Text Size on Classification Accuracy

- Training/testing with longer texts supports higher accuracy.
- But even with 100 words per text, one reaches >80%.
- Lex & Psych best in short texts, Syn more linear increase

Readability at the sentence level
(Vajjala & Meurers 2014a)

- Test on sentence-aligned Wiki–SimpleWiki (Zhu et al. 2010)
- Predictions of WeeBit text model:

  - Simplification is relative: A simplified sentence is simpler than its unsimplified version, but can be harder than another one.
  - Hard texts are not simply collections of hard sentences.
Dealing with the multi-level nature of evidence

Beyond averages

- To classify texts, we rely on evidence at different levels:
  - words, sentences, texts
- What is the best way to combine the evidence?
  - Is computing averages really preserving what is relevant?
- Explored for word frequencies in Chen & Meurers (2016a) using SUBTLEX Zipf scale (Van Heuven et al. 2014)

Dealing with the multi-level nature of evidence

Word frequencies in texts at different levels of granularity

- How about grouping tokens to obtain $n$ averages per text?
  1. $n$ frequency bands of the language
  2. $n$ clusters of words in document closest in frequency
  A text is represented by one avg. frequency feature per group.

  ⇒ This works well for 10-fold CV in WeeBit corpus:
  1. 67.5% accuracy with 90 frequency bands (by types)
  2. 54.6% accuracy with 100 clusters (by tokens)

- But does this generalize across corpora?
  → Compare WeeBit 10-fold CV with test on CommonCore, reporting Spearman's rank correlation coefficient ($\rho$)

Dealing with the multi-level nature of evidence

Spearman rank correlation within and across corpora

- Accuracy of 10-fold CV classification on WeeBit (5 levels):
  - Average frequencies baseline:
    - 24.2% with average token frequency as feature
    - 32.1% with average type frequency as feature
  - Adding Standard Deviation:
    - 39.9% with average token frequency + SD as features
    - 43.3% with average type frequency + SD as features

  → Let's explore different levels of granularity.
    - most informative: characterize a text through the vector of frequencies of every token in the text, but:
      - unlikely to generalize, and
      - texts differ in length
Using readability to rank web search results
(Vajjala & Meurers 2013)

- Are state-of-the-art readability models actually useful for classifying texts as found on the web?
  - Can we re-rank search results based on reading levels?

- Implementation details:
  - feature set from Vajjala & Meurers (2012)
  - trained model on WeeBit corpus
  - modeling: regression, since we want output on a scale

- We applied the readability model to search results obtained through BING search API.
  - took 50 search queries from a public query log
  - computed reading levels for Top-100 results

Dealing with the multi-level nature of evidence
Summary

- For aggregating word frequencies at the text level:
  - grouping by language frequency band better within-corpus
  - hierarchical clustering of words in text generalizes better

- Conclusion: We should put more thought into how to combine the multi-level nature of the readability evidence.

- Next idea to test:
  - How can the incremental process information provided by Surprisal (Boston et al. 2008) inform text difficulty?
  → Hierarchical clustering of Surprisal profiles of sentences.

Results: Reading levels of top search results
(Vajjala & Meurers 2013)

<table>
<thead>
<tr>
<th>Query</th>
<th>Result Rank:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>copyright copy law</td>
<td>1.8</td>
</tr>
<tr>
<td>halley comet</td>
<td>1.7</td>
</tr>
<tr>
<td>europe union politics</td>
<td>3.6</td>
</tr>
<tr>
<td>shakespeare</td>
<td>2.4</td>
</tr>
<tr>
<td>euclidean geometry</td>
<td>3.9</td>
</tr>
</tbody>
</table>

- Avg. reading level of search results quite high (5 = GCSE)

- The model identifies a range of reading levels among most relevant results returned by search engine.

- Readability-based re-ranking of results potentially useful for language-aware search engines in Language Teaching (Ott & Meurers 2010; Chinkina & Meurers 2016)
Linking readability and L2 development

- Can we use the complexity features as a looking glass on both readability and L2 development?
  - Idea: Provide texts just above level of student \((i + 1)\)
  - EFCamDat corpus (Geertzen et al. 2013), prerelease 2:
    - 1.2 million assignments (70 million words)
    - written by nearly 175 thousand learners
    - across a wide range of levels (CEFR A1–C2)
Linking readability and L2 development

First conclusions

- Good face validity for using same measures for readability and L2 development.
- Various challenges that need to be addressed, e.g.:
  - impact of analyzing learner language: e.g., variable orthography should not be counted as lexical richness
  - systematic sentence segmentation difficult (both for learner language and certain text types, e.g., CVs)
- Questions to be addressed for selecting $i + 1$ text material:
  - How reliably can we determine reading ability based on measures of writing proficiency?
  - What does the “+1” amount to, for the different aspects of linguistic modeling (lexicon, syntax, . . .)?

Summary (cont.)

- Sentence-level analysis:
  - Documents at a given reading level contain sentences at a range of levels of complexity.
  - Readability analysis at the sentence level is feasible, but needs to take the relative nature of readability into account.
- Taking the multi-level nature of the evidence on readability (word, sent., text) into account can support significant gains.
- Approach is applicable to other languages:
  - German readability (Hancke, Meurers & Vajjala 2012) and proficiency (Hancke & Meurers 2013), readability for Bulgarian (Nikolova 2015) and Greek (Georgatou 2016)

Outlook on some current collaborations

- Analyzing linguistic complexity and accuracy in relation to task demands (Alexopoulou, Michel, Murakami & Meurers 2017)
- At interface with empirical educational science (LEAD):
  - Studying the cognitive correlates of readability using eye-tracking (Vajjala, Meurers, Eitel & Scheiter 2016)
  - Is the language in school books appropriate for grade level and school type? ReadingDemands with Berendes & Bryant
  - Linguistic and numerical factors contributing to the complexity of Word Problems (Daroczy et al. 2015)
  - Impact of linguistic complexity of questionnaires (Göllner et al. 2014)
  - CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis (Chen & Meurers 2016a)
    - http://www.ctapweb.com

Summary

- Measures of development from SLA research turn out to be excellent predictors for readability classification.
- Our approach outperforms previously published readability assessment results on WeeklyReader data.
  - best non-commercial readability model on CCSS data
- Feature set generalizes well to other data sets and genres
  - including spoken language transcripts and web data
- Readability reflected in a wide range of linguistic properties
  - Linguistic and cognitive features complement each other.
References


Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers

Introduction
How can we obtain evidence?

Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki

Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary

Outlook


Sentence-level analysis using ranking

Ranking experiments on Wiki-SimpleWiki (Vajjala & Meurers new)

- Employ a Ranking algorithm, as commonly used in information retrieval to rank search results.
- Setup: SVMRank on Wiki–SimpleWiki (10-fold CV)
- Result: 82.7% accuracy
  - accuracy = percentage of correctly ranked pairs
  - baseline: 72.3% accuracy for Flesch-Kincaid formula
- Do these results generalize to other data?
  → We compiled a new data set from OneStopEnglish.com
Sentence-level analysis using ranking
OneStopEnglish experiments (Vajjala & Meurers new)

- Texts from *The Guardian* manually rewritten at 3 levels.
- We extracted and aligned 3113 sentences at two levels (OSE2) and 837 across three levels (OSE3), e.g.:
  - Adv: *In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.*
  - Int: *In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.*
  - Ele: *In Beijing, people went to the Apple Store with flowers and candles.*

Sentence-level analysis using ranking
OSE and Cross-Corpus Results (Vajjala & Meurers new)

- Defined separate train and test sets for Wiki and OSE2.
- Flesch-Kincaid baselines:
  - Wiki
    - Test: 69.0%
  - OSE2
    - Test: 69.6%

- Same-Corpus Results of RankSVM model:
  - Wiki
    - Train: 81.8%
    - Test: 81.5%
  - OSE2
    - Train: 74.6%
    - Test: 77.5%

- Cross-Corpus Results of RankSVM model:
  - Wiki
    - Train: 81.8%
    - Test: 81.5%
  - OSE2
    - Train: 74.6%
    - Test: 77.5%

⇒ Rich features set supports reliable sentence-level analysis.