Informing ICALL Reading System Design by Linking Text Complexity and Learner Proficiency with Textual Feature Vector Distance

Xiaobin Chen,
xiaobin.chen@uni-tuebingen.de

Detmar Meurers,
dm@sfs.uni-tuebingen.de

Tübingen University

August 24, 2017
Comprehensible input

Reading texts of appropriate difficulty levels to the learner’s language proficiency.

Providing learners with opportunities to practice being competent readers and motivating them to read more (Milone and Biemiller, 2014).

Factors affecting the appropriateness of reading input:

- Complexity or readability of text
- Reader-related factors: purpose of reading, the reader’s abilities, prior knowledge, interest and so on
Selecting Appropriate Reading Texts—Readability Assessment

- Readability: the sum of all elements of a text that affects a reader’s understanding, reading speed, and level of interest in the text (Dale and Chall, 1949).
- Qualitative and quantitative assessment
- Quantitative assessment: more objective and easier to automatize
- Multiple regression, machine learning approaches

**Problem:** The interaction between the reader and the reading text is often overlooked.
The Optimal Scenario for An ICALL System for Reading

- Learner modeling: proficiency, interests, prior knowledge, learning strategies...
- Assessment of text complexity/readability
- Adaptive assignment of reading input based on text complexity and learner factors
A Framework of ICALL for Reading

System elicits composition from student.

System extracts student text features.

System selects from repository texts of appropriate levels for student.

System periodically re-elicits compositions from user and decides whether student has made progress.

Student reads the assigned articles and does reading tasks.
The Proposed Method

Representing learner proficiency and text readability within the same vector space and using the vector distance between them as a measure of reading text appropriateness.
RQ: Can the distance between feature vectors of learner-produced texts and authentic reading texts be used to decide which readings are appropriate for the reader?
Usage of Textual Features

Assessment of

- text readability (Crossley et al., 2007; Flor et al., 2013; Lu et al., 2014; François and Watrin, 2011; Hancke et al., 2012; Heilman et al., 2007), and

- student writings for proficiency placement (Lu, 2010; Attali and Burstein, 2006)

There has been no attempt to use textual feature vectors to unify the readability and learner proficiency spaces.
Hypotheses

1. Vector distance should be positively correlated with level difference of authentic texts, i.e., greater level difference would result in greater vector distance and vice versa.

2. For linking learner produced and authentic texts: Given an authentic text supposedly appropriate for the learner, the distance between the authentic text and a text produced by a learner of lower proficiency level should be smaller than that between the authentic text and a text produced by a more proficient learner.
Test of Hypotheses: Corpora

- 30 articles (each offered in 5 different reading levels) randomly selected from Newsela
- 96 English continued stories written by 48 Chinese EFL students after reading stories whose endings had been removed (Wang and Wang, 2015).

### Table: Details of the Newsela and CW Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th># Levels</th>
<th># Texts</th>
<th>Words/Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newsela</td>
<td>5</td>
<td>150</td>
<td>763</td>
</tr>
<tr>
<td>CW</td>
<td>2</td>
<td>96</td>
<td>641</td>
</tr>
</tbody>
</table>
Following Vajjala and Meurers’s (2012) feature schemes, 102 lexical, syntactic, and discoursal features were extracted from each text, forming a 102-dimension vector to represent the text. Examples of features (see Appendix for full list):

- Corrected type token ratio
- Lexical density
- Mean length of clause
- Number of Dependent Clauses per T-unit
- Mean MRC Age of Acquisition
- Global/Local content word overlap
- ...
Euclidean n-space distance between \( p \) and \( q \) can be calculated with the Pythagorean formula:

\[
d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}
\]
Results: the Newsela Corpus

Figure: Feature Vector Euclidean Distance on Text Level Difference

- The greater the level differences, the further the vector distances.
- One-way ANOVA $F(3, 296) = 403.1, p < .001$. Post hoc TukeyHSD tests significant for all level difference pairs (all adjusted $p < .001$).

X.B. Chen & D. Meurers (Tübingen)
Results: the Continuation Writing Corpus

**Figure:** Illustration of Hypothesis 2 with the CW Corpus

Dist. 1 > Dist. 2

**Figure:** Illustration of Hypothesis 2 with the CW Corpus
### Table: Results from the Continuation Writing Corpus

<table>
<thead>
<tr>
<th></th>
<th>Distance 1</th>
<th>Distance 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>mean</strong></td>
<td>16.66</td>
<td>14.37</td>
</tr>
<tr>
<td><strong>sd</strong></td>
<td>3.58</td>
<td>2.84</td>
</tr>
</tbody>
</table>

Paired sample t-test: $t = 3.35$, $df = 47$, $p \leq .001$
Summary

- It is highly important that language learners are provided with sufficient authentic target language input that suits their language ability.
- A commonly used method is to do readability assessment before assigning texts to readers.
- However, because of the great variety of language learners, readability assessment suffers from lack of account on learner factors, such as language proficiency, prior knowledge, interests, etc.
- We proposed using vector distance as a measure of text level difference (both for authentic texts and learner produced texts as well as between them).
- The proposed method is validated with an authentic corpus and a corpus of continuation writings, forming the baiss for designing ICALL system for reading text selection.
Future Directions

- Implementing and empirically testing a system designed with the proposed framework.
- Reduction of vector dimensions.
- Systems targeting specific linguistic constructs.
References


Appendix: List of Textual Features

Mean Bird et al.'s Age of Acquisition on Words
Mean Bristol’s Age of Acquisition on Words
Mean Cortese and Khanna’s Age of Acquisition on Words
Mean Kuperman et al.'s Age of Acquisition on Words
Mean Kuperman et al.'s Age of Acquisition on Lemmas
Referential Expressions: Number of Particles per Sentence
Referential Expressions: Percentage of Articles
Referential Expressions: Percentage of Personal Pronouns
Referential Expressions: Number of Personal Pronouns per Sentence
Referential Expressions: Number of Possessive Pronouns per Sentence
Referential Expressions: Percentage of Possessive Pronouns
Referential Expressions: Pronoun Noun Ratio
Referential Expressions: Number of Pronouns per Sentence
Referential Expressions: Percentage of Pronouns
Referential Expressions: Proper-Noun Noun Ratio
Global Argument Overlap
Global Content Word Overlap
Global Noun Overlap
Global Stem Overlap
Local Argument Overlap
Local Content Word Overlap
Local Noun Overlap
Local Stem Overlap
Mean MRC Age of Acquisition
Mean MRC Colorado Meaningfulness
Mean MRC Concreteness
Mean MRC Familiarity
Mean MRC Imagineability
Mean MRC Pavio Meaningfulness
Adjective Variation
Adverb Variation
Corrected Verb Variation 1
Modifier Variation
Appendix: List of Textual Features

Noun Variation
Number of Adjectives
Number of Adverbs
Number of Conjunctions
Number of Determiners
Number of Function Words
Number of Interjections
Lexical Density
Number of Modal Verbs
Number of Nouns
Percentage of Pronouns
Percentage of Prepositions
Number of Pronouns
Number of Proper Nouns
Number of Verbs
Percentage of Verbs
Number of Verbs in Past Tense
Number of Gerund or Verbs in Present Participle
Number of Past Participle
Number of Verbs not in 3-rd Person Singular Present
Number of Verbs in 3-rd Person Singular Present
Number of Wh-Pronouns
Squared Verb Variation 1
Verb Variation 1
Verb Variation 2
Number of Constituents per Clause
Number of Constituents per T-unit
Percentage of Complex T-unit
Percentage of Coordinate Clauses
Number of Coordinate Clauses per T-unit
Percentage of Dependent Clauses
Number of Dependent Clauses per T-unit
Mean Length of Clause
Appendix: List of Textual Features

Mean Length of T-unit
T-unit Complexity Ratio
Number of Verb Phrase per T-unit
Mean Parse Tree Height Per Sentence
Mean Sentence Length
Mean Number of Clauses per Sentence
Mean Number of Conjunction Phrases per Sentence
Mean Number of Constituents per Sentence
Number of Noun Phrases
Mean Number of Noun Phrases per Sentence
Number of Prepositional Phrases
Mean Number of Prepositional Phrases per Sentence
Mean Number of Reduced Relative Clauses per Sentence
Mean Number of S-bars per Sentence
Number of Sentences
Mean Number of Sub-trees per Sentence
Mean Number of T-units per Sentence
Number of Verb Phrases
Mean Number of Verb Phrases per Sentence
Mean Number of Wh-Pronouns per Sentence
The Automated Readability Index
The Coleman-Liau Readability Index
The Fog Readability Index
The Forecast Readability Index
The Flesch Readability Index
The Kincaid Readability Index
The LIX Readability Index
The SMOG Readability Index
Number of Characters
Number of Syllables
Bilogarithmic Type Token Ratio
Corrected Type Token Ratio
Mean Textual Lexical Density
Appendix: List of Textual Features

Root Type Token Ratio
Type Token Ratio
Uber Index