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

Reading and Language Learning

- Comprehensible input
- Reading texts of approrpiate 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

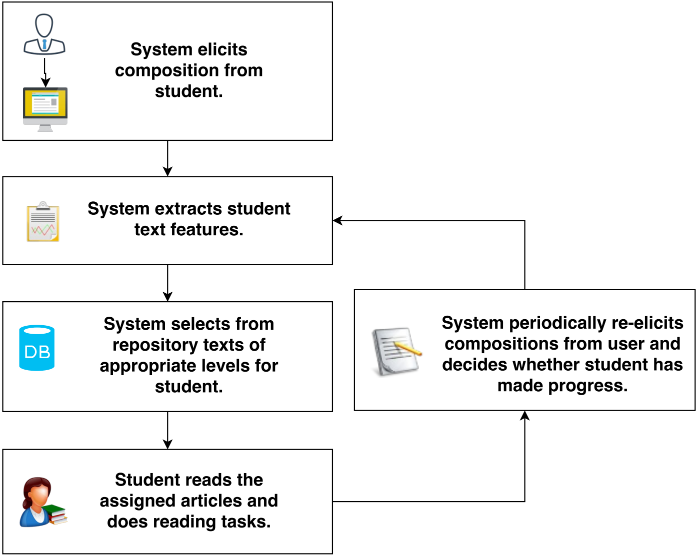
- Readability: the sum of all elements of a text that affects a reader's understanding, reading speed, and level of interest in the text (Dale and Chall, 1949).
- Qualitative and quantitative assessment
- Quantitative assessment: more objective and easier to automatize
- Multiple regression, machine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



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.

Verification of the Vector Distance Method

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 vecotrs 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

Corpus	# Levels	# Texts	Words/Text
Newsela	5	150	763
CW	2	96	641

Test of Hypotheses: Extraction of Feature Vectors

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
- ...

Calculation of Vector Distance

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

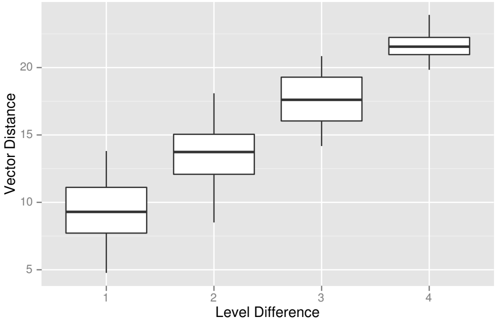


Figure: Feature Vector Euclidean Distance on Text Level Difference

- 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 $p < .001$).

Results: the Continuation Writing Corpus

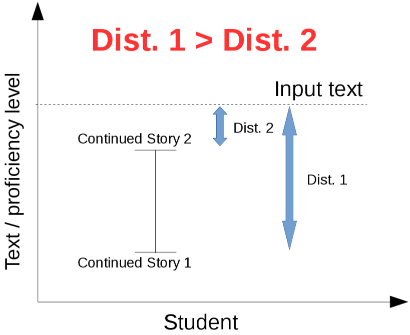


Figure: Illustration of Hypothesis 2 with the CW Corpus

Results: the Continuation Writing Corpus

Table: Results from the Continuation Writing Corpus

	Distance 1	Distance 2
mean	16.66	14.37
sd	3.58	2.84
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

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Crossley, S., Dufty, D., McCarthy, P., and McNamara, D. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the 29th annual conference of the Cognitive Science Society*, pages 197–202, Nashville, Tennessee, USA. Lawrence Erlbaum Associates.

Dale, E. and Chall, J. (1949). The concept of readability. *Elementary English*, 26(3).

Flor, M., Klebanov, B. B., and Sheehan, K. M. (2013). Lexical tightness and text complexity. In *Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 29–38, Atlanta, Georgia. Association for Computational Linguistics.

François, T. and Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for french as a foreign language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 441–447, Hissar, Bulgaria.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics*, Bombay, India. Association for Computational Linguistics.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT 2007*, pages 460–467, Rochester, NY. Association for Computational Linguistics.

Lexile (2007). The Lexile Framework ® for reading: Theoretical Framework and Development. Technical report, MetaMetrics, Inc., Durham, NC.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Lu, X., Gamson, D. a., and Eckert, S. A. (2014). Lexical difficulty and diversity of American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics*, 19(1):94–117.

Milone, M. and Biemiller, A. (2014). The development of ATOS: The Renaissance readability formula. Technical report, Renaissance Learning, Wisconsin Rapids.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Quebec, Canada. Association for Computational Linguistics.

Wang, C. and Wang, M. (2015). Effect of alignment on L2 written production. *Applied Linguistics*, 36(5):503–526.

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