EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD Graduate School

Multidimensional Vector Distance as Measure of Text Level Differences

Xiaobin Chen, Detmar Meurers

Rationale

- Textual complexity assessment criticized for not being based on comprehension or student proficiency.
- Effects of assigning texts of certain complexity level to learners unclear.
- No application of readability research in actual teaching system, i.e., it is unclear how to provide individualized reading input based on learner proficiency levels.
- This study tests whether multidimensional vector distance can be used as a measure of text level difference and its validity to link reader proficiency with text complexity.

A System Prototype

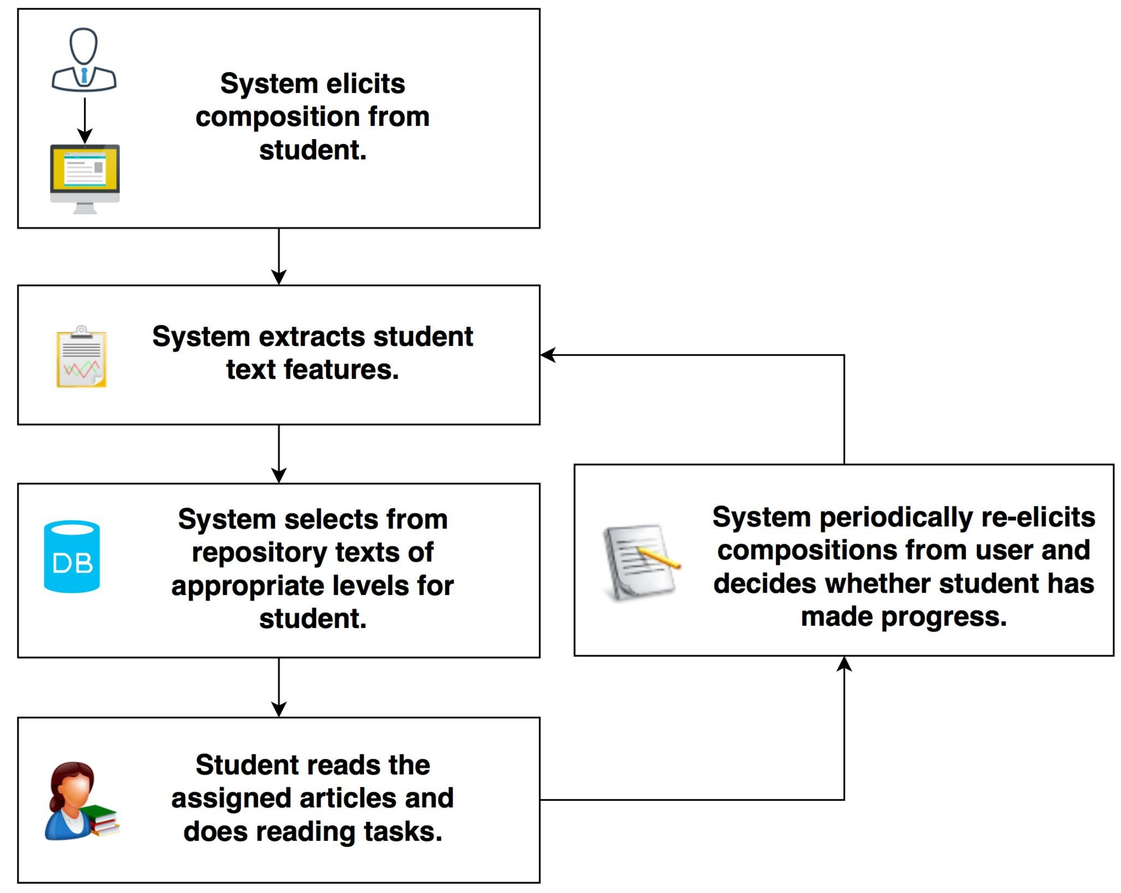

Figure 1. *A reading tutoring system prototype*

Key Question and Proposed Solution

- Question: How to decide what texts are appropriate for students of a given level?
- Solution: A text from the system's article repository that has features closest to those of the composition written by the student.
- Hypothesis: Features (Vajjala and Meurers, 2012) vector distances correlate with level differences, i.e., greater level differences result in greater vector distances.

The Continuation Task Data (Wang & Wang, 2015)

- Number of students: 48
- Task: Continuation writing in English
- Input texts: Chinese/English versions of stories with endings omitted
- Results in the original study: Improved accuracy after reading English input
- Hypothesis for the present study: The distance between composition 1 (Chinese version input) and the English input text should be greater than the distance between composition 2 (English version input) and the input text (Figure 2).

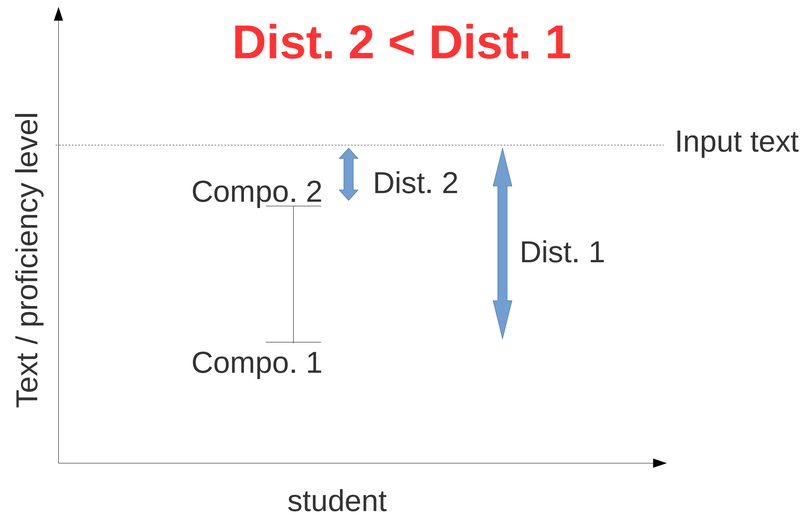

Figure 2. *Illustration of hypothesis*

Results: Continuation Task Data

	Distance between Compo. 1 and input	Distance between Compo. 2 and input
mean	16.66	14.37
sd	3.58	2.84

Paired sample t-test: t = 3.35, df = 47, p < .001

Table 1. *Results of Continuation Data Set*

The Newsela Data Set

- 30 articles * 5 levels/article = 300 texts
- Hypothesis: Greater level gaps (e.g., the distance between levels 1 and 5 compared to that between levels 1 and 2.) are associated with greater vector distances.

Results: Newsela Data

One-way ANOVA: F(3, 296) = 403.1, p <.001 (Figure 3)

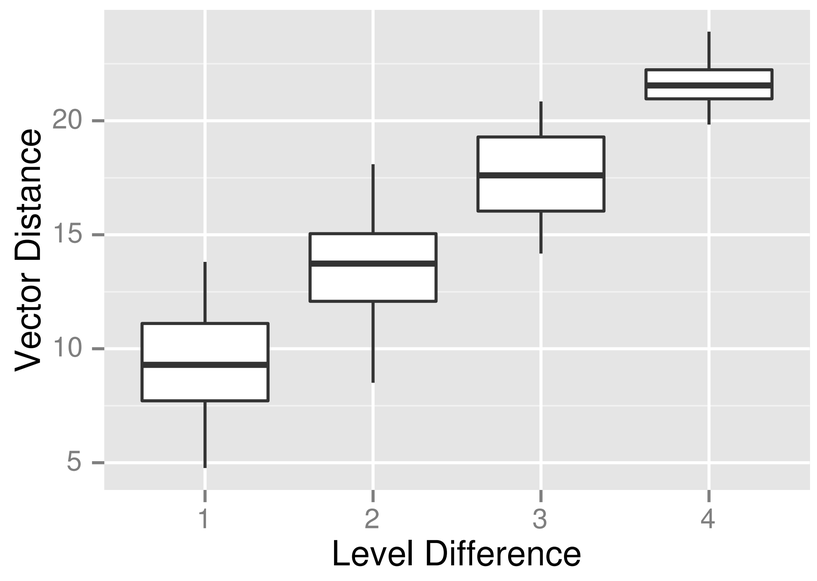

Figure 3. *Vector distance on level difference with Newsela data*

Conclusions

It is valid to use features vector distance

- to measure text readability level difference.
- to link reader proficiency with text readability, potentially for the purpose of individualized reading text assignment.

Outlook

- Reduction of vector dimensions
- What's the optimal distance between input text and student production for language proficiency development? Low, medium, or high?

References

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*.
Wang, C., & Wang, M. (2015). Effect of alignment on L2 written production. *Applied Linguistics*, *36*(5), 503-526.

Universität Tübingen
<LEAD Graduate School>
<Garten Str. 29 · 72074/Tübingen · Germany>
<Telefon +49 7071 2973578> · <xiaobin.chen@lead.uni-tuebingen.de> · <dm@sfs.uni-tuebingen.de>

LEAD Graduate School