EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Xiaobin Chen and Detmar Meurers

LEAD Graduate
School & Research
Research Network

LEAD
Graduate School & Research Network

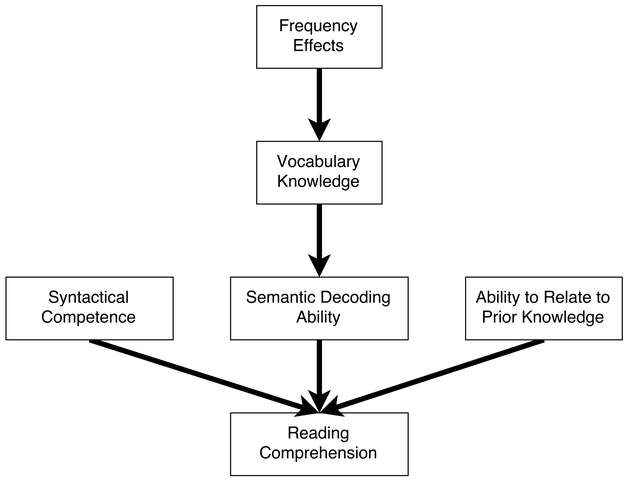
Characterizing Text Difficulty with Word Frequencies

Research Question

How can lexical frequency as a word-level property best be aggregated to characterize readability at the text level?

The Frequency Effect and its impact on comprehension

Word frequency reflects language experience and vocabulary knowledge as important factors of reading comprehension:



Experimental Setup

Use supervised machine learning to test the predictive power of different encodings of lexical frequency at the document level.

- **Frequency Norms** obtained from SUBTLEX-UK and -US (Brysbaert and New, 2009; Van Heuven et al., 2014)
  - Log transformed frequency: Zipf measure
  - Contextual Diversity (CD) measure
- **Machine Learner:** KNN algorithm from R `class` package.
- **Training Corpus:**
  - WeeBit (Vajjala and Meurers, 2012) = Weekly Reader + BBC-Bitesize
  - Five levels, targeting children 7–16 years
  - 616 texts per level
- **Test Corpus:**
  - Common Core State Standard Exemplars (CCSSO, 2010)
  - Five levels, 168 texts
- **Performance measure:** Spearman's correlation coefficient (ρ) between model estimates and actual document level.

Study 1: Adding Standard Deviation

Features: Mean measures with/without SD for Tokens or Types

Results:

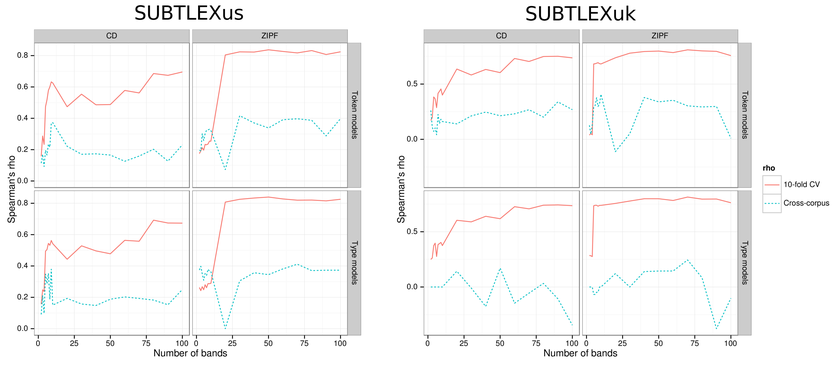
	Token		Type	
	−SD	+SD	−SD	+SD
US-ZIPF	.03	.34***	.33***	.35***
US-CD	-.27***	.28***	.22**	.33***
UK-ZIPF	-.13	.26***	.36***	.38***
UK-CD	.00	.02	.33***	.27***

- +SD models performed better than −SD models.
- Type models uniformly outperformed token models.
- Frequency (Zipf) performed better than contextual diversity (CD) for readability assessment.

Study 2: Mean Frequencies of Words from Language Frequency Bands

Features: frequency means of words from each of the stratified frequency bands of the language

Results:

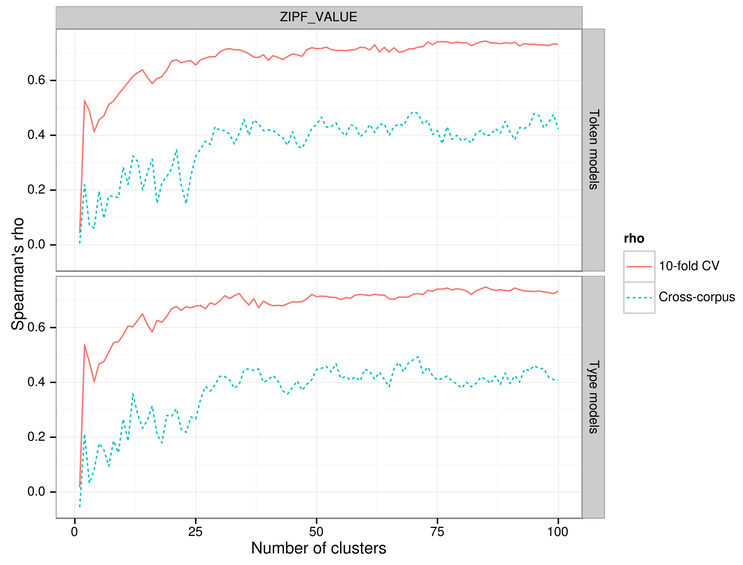


- Similar 10-fold CV performance for type and token models.
- Better generalizability for type models than token models.
- Models with stratified means outperformed those with a grand mean for all words.

Study 3: Frequency Cluster Means

Features: Cluster means of word frequencies

Results:



- Token and type models performed comparably.
- As the number of cluster increased, performance improved.
- Cross-corpus testing performance was stable.

Conclusions

- Adding SD of document word frequencies greatly benefits the readability prediction model.
- Using means of frequency bands in a language or clustering words similar in frequency provides further advantages.
- Method using language band means (Study 2) is more prone to overfitting than the clustering method (Study 3).

References

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

CCSSO (2010). Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects. appendix B: Text exemplars and sample performance tasks. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers. `http://www.corestandards.org/assets/Appendix_B.pdf`.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In Tetreault, J., Burstein, J., and Leacock, C., editors, *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, pages 1–15.