An analysis of a French as a Foreign Language Corpus for Readability Assessment

Thomas François

NLP4CALL, Uppsala

13-11-2014
Readability models have roles to play in iCALL:

1. Find educational material of a given level, on the web
Enhance systems for automatic exercise generation (proposal based on ALFALEX) [Verlinde et al., 2003]
Problematic

- Many formulas “available” for L1, especially for English
  [Collins-Thompson and Callan, 2005, Feng et al., 2010, Vajjala and Meurers, 2012]

- Limited amount of models “available“ for L2
  [Heilman et al., 2007, Schwarm and Ostendorf, 2005]

- AND only two formulas use the CEFR scale (current standard)
  [François and Fairon, 2012, Pilán et al., 2014]

- Problem: Large amount of efforts required to collect the annotated data to train a readability model!
Objectives of the paper

1. Gathering a corpus of texts annotated for difficulty, as large as possible
   ▶ Critical review of 5 common annotation approaches in readability
   ▶ We opt for the extraction of texts from textbook series (with some conditions) and giving the textbook level to all texts extracted from it.
   ▶ We report the collect process of about 2,000 texts for French as a Foreign Language (FFL)

2. Investigate the shortcomings of this criterion (which is mainstream in the field)
   ▶ Produce unbalanced corpus
   ▶ Homogeneity of annotations across textbook series is questionable
2 experiments

Class imbalanced effect
- We compared the results of models either based on a balanced or an unbalanced corpus.
- The ordinal logistic models used include two well-acknowledged features (mean number of words/sentence and of letters/word).
- It confirmed that majority class may deform the prediction space.

Homogeneity of the annotations
- We compared the lexical and syntactic difficulty of textbook series among them (for each level).
- We used ANOVA and MANOVA and noticed significant homogeneity issues in the corpus.
- A relation between the type of pedagogical approach and homogeneity issues also appeared.

Take-home message: Test the homogeneity of annotations when using texts from educational sources (just as we would do for human annotators).
References I


