

# Readability-Based Search

Alexandra L. Uitdenbogerd

School of Computer Science and IT, RMIT

# Some Types of Text Retrieval

- Documents “about” some topic
- Known document
- Find me something interesting
- Examples of word or phrase usage (concordancer)
- Find me something at my reading level in language X
- Other linguistic study

# The Importance of Language Acquisition

- Many people need to regularly communicate in a second/foreign language.
- More people speak English as a foreign/second language (~1100 million) than as a first language (~400 million).
- Over 200 million speak Spanish, Hindi and Russian as a second/foreign language.

# Readability-Based Searcher

- Web Text recommender system for foreign language learners
- Recommends based on appropriate level of difficulty - readability

# Measuring Readability

- Readability measures are based on estimates of vocabulary and grammar difficulty
- several readability measures based on the following text attributes:
  - word frequency
  - word length (syllables, characters)
  - sentence length

- Example:

$$\text{Fog index} = \frac{\text{no. of words}}{\text{no. of sentences}} + \frac{\text{no. of 3-syllable words}}{\text{no. of words}} \times 100 \times 0.4$$

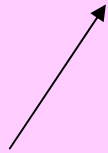
# General Research Questions

- Can we better estimate readability with on-line text using today's computer capabilities? (Early work on readability mostly focused on manual methods)
- How does readability differ for different languages?
- Do different readability measures apply for foreign language readers than for native readers?

# Example of Cognates and Readability

La potion des pythons

Immobilise les dragons



Rare words  
that are  
cognates

Que de bruit

Qu'on produit



Common  
words (non-  
cognate)

# Current Results

- Sentence length in words seems to be very good compared to more sophisticated measures of readability for French-English bilingualism
- Incorporating cognates yields slight improvement.
- Web collections contain documents with a wide range of readability (as measured by readability measures).
- Using Web collections poses significant problems for readability measurement due to its structure and diversity.

# Future Work

- Develop automatic cognate identification techniques
- Test and compare lisibilité formulae.
- Developing web-capable readability measurement techniques.