Computational readability: need for a domain-oriented approach?

Thomas François

B.A.E.F and Fulbright Fellow
University of Pennsylvania

CUNY, September 21, 2012
Plan

1. Brief review of readability and its issues
2. A FFL “AI formula” aiming at specialization
3. Discussion: what level of specificity for readability?
4. References
Plan

1. Brief review of readability and its issues

2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
What is readability?

**Origin**: Readability dates back to the 20s, in the U.S. (only 60s for the French-speaking community).

**Objective**: Aims to assess the difficulty of texts for a given population, without involving direct human judgements.

**Method**: Develop tools, namely readability formulas, which are statistical models able to predict the difficulty of a text given several text characteristics.

Most famous ones are those of [Dale and Chall, 1948] and [Flesch, 1948].
Classic formulas

Example of the formula of [Flesch, 1948, 225] :

\[
\text{Reading Ease} = 206.835 - 0.846 \, \text{wl} - 1.015 \, \text{sl}
\]

where :

- **Reading Ease (RE)** : a score between 0 and 100 (a text for which a 4th grade schoolchild would get 75% of correct answers to a comprehension test)
- **wl** : number of syllables per 100 words
- **sl** : mean number of words per sentence.

- Use of linear regression and **only a few** linguistic **surface** aspects.
- Claim that the formula can be applied to a large variety of situations.
Recent works: “AI readability”

- This new trend in readability rose with the 21st century [Si and Callan, 2001, Collins-Thompson and Callan, 2005].
- It combines NLP-enabled feature extraction with state-of-the-art machine learning algorithms.
- In most cases, readability is considered as a classification problem and not anymore as a regression one!
- NLP and machine learning processing require a large corpus!

Let’s focus a bit more on this last point!
The corpus issue

**Classic approach**

- Sample of readers
- Corpus
- Reading comp. tests
- Cloze test
- Reading speed
- Scores per text and per reader
- label = mean scores per text
- small amount of labelled texts

**AI approach**

- Not possible to use a sample
- A need: a large corpus
- Only left "criterion" = expert judgments
- Eg. Weekly reader (Schwarm and Ostendorf, 2005)
- Reliability ? Coherence ?
Specialization of the formulas

What is specialization?

It means defining a specific population of interest (e.g., children, L2 readers, etc.) AND adapting the model to take into account the specificities of that population.

In other words, it amounts to:

- Use a corpus assessed by the given population to tune the weights of each predictor.
- Adapt some well-known predictors to better fit the specific context.
- Find some new predictors that correspond to specific features of the specific context.
Examples of specialization

- Specialization is not new:
  - Standardized tests readability by [Forbes and Cottle, 1953]
  - 1st-3th grade schoolchildren by [Spache, 1953]
  - Scientific texts by Jacobson (1965) or Shaw (1967)
  - etc.

- More recent works:
  - Scientific texts [Si and Callan, 2001]
  - People with ID [Feng et al., 2009]
  - L2 readers [Heilman et al., 2007, François, 2009a]
Effect of specialization

- The idea is that, for a specific population, a specialized formula should yield better performance than a general model.
  - Spache claimed $R = 0.818$ vs. $R = 0.7$ of Flesch, but no cross-validation!

- Surprisingly, this assumption is not always accepted and has not been thoroughly tested.
Effect of specialization (2)

For the readability of L2:

- Common practice: try to apply a L1 formula to a L2 context [Cornaire, 1988]

- Brown (1998) compared 6 classic formulas on 50 texts (assessed by 2300 students) and got $0.48 < R < 0.55$, while he obtained $R = 0.74$ for his L2 specialized formula.

- BUT Greenfield (1999) had the 32 Bormuth’s excerpts assessed by 200 students and...
  → Correlation between L1 and L2 scores was high ($r = 0.915$)
  → Retrained the 5 formulas on this corpus and get a small gain only.

They both used only two surface features... What about a more complex model?
Plan

1. Brief review of readability and its issues

2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
1. Brief review of readability and its issues

2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
Not much work...

- [Tharp, 1939] positions himself against the previous approach and offers one of the first specific formulas for FLE, based on cognates.
- [Cornaire, 1988] investigates the adaptation of the L1 formula for French by [Henry, 1975].
- [Uitdenbogerd, 2005] suggests a formula that also takes into account cognates:

\[
\text{FR} = 10 \times \text{WpS} - \text{Cog}
\]

\text{WpS} : \text{mean number of words per sentence.}
\text{Cog} : \text{number of cognates per 100 words.}
Methodology

Plan

1. Brief review of readability and its issues

2. A FFL "AI formula" aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
Introduction
Methodology

Conception of a formula : methodological steps

1. Collect a corpus of texts whose difficulty has been measured using a criterion such as comprehension tests or cloze tests.

2. Define a list of linguistic predictors of the difficulty, such as sentence length or lexical load.

3. Design a statistical model (traditionally linear regression) based on the above features and corpus.

4. Validate the model.

Y = B2 A1 C2 A2 ...

Xi1 = 748.7
Xi2 = 5.32
...
Xin = 1

Prediction on a new text
The corpus (1)

Criterion = expert judgments = textbooks!
→ The assumption is that the level of a text can be considered the same as the level of the textbook it comes from.

The type of criterion affects the difficulty scale used.
→ We extracted 2042 texts from 28 FFL textbooks, following the CEFR scale [Conseil de l’Europe, 2001].

The CEFR scale
It is the official EU scale for L2 education.
It has 6 levels: A1 (easier), A2, B1, B2, C1, and C2 (higher).
Not all FFL textbooks were used:

1. Have to follow the CEFR recommendations (posterior to 2001).
2. Language should be modern (arises from condition 1).

Another selection was performed at the text level:

1. Only texts related to a reading comprehension task.
2. Instructions were not considered.
## Distribution of the texts per level

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A1+</th>
<th>A2</th>
<th>A2+</th>
<th>B1</th>
<th>B1+</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activités CECR</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>41</td>
<td>39</td>
<td>50</td>
<td>63</td>
<td>8</td>
</tr>
<tr>
<td>Alter Ego</td>
<td>46</td>
<td>44</td>
<td>61</td>
<td>31</td>
<td>74</td>
<td>42</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Comp. écrite</td>
<td>/</td>
<td>/</td>
<td>34</td>
<td>53</td>
<td>39</td>
<td>50</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Connexions</td>
<td>34</td>
<td>26</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Connexions : prep. DELF</td>
<td>/</td>
<td>11</td>
<td>/</td>
<td>12</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Delf/Dalf</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>31</td>
<td>78</td>
<td>19</td>
</tr>
<tr>
<td>Festival</td>
<td>42</td>
<td>34</td>
<td>/</td>
<td>/</td>
<td>28</td>
<td>26</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ici</td>
<td>13</td>
<td>28</td>
<td>25</td>
<td>17</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Panorama</td>
<td>31</td>
<td>27</td>
<td>50</td>
<td>48</td>
<td>56</td>
<td>57</td>
<td>41</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Rond-point</td>
<td>3</td>
<td>19</td>
<td>4</td>
<td>7</td>
<td>21</td>
<td>19</td>
<td>76</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Réussir Dalf</td>
<td>/</td>
<td>17</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>43</td>
<td>22</td>
</tr>
<tr>
<td>Taxi !</td>
<td>27</td>
<td>/</td>
<td>23</td>
<td>21</td>
<td>56</td>
<td>51</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Tout va bien !</td>
<td>/</td>
<td>50</td>
<td>36</td>
<td>56</td>
<td>45</td>
<td>37</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Total</td>
<td>196</td>
<td>256</td>
<td>233</td>
<td>245</td>
<td>360</td>
<td>321</td>
<td>198</td>
<td>184</td>
<td>49</td>
</tr>
</tbody>
</table>

**TABLE:** Number of texts per level, for each textbook series used.
I implemented 406 variables, most of them draw inspiration from previous studies:

- **lexical**: statistics of lexical frequencies; percentage of words not in a reference list; N-gram models; measures of lexical diversity; length of the words;
- **syntactic**: length of the sentences; part-of-speech ratios;
- **semantic**: abstraction and personalisation level; idea density; coherence level measured with LSA;
- **specific to FFL**: detection of dialogue.

Some of them were never experimented in a FFL (or even L2) context.
Contribution of cognitivist studies on the reading process

Psychological description of the reading process provided ideas for new predictors:

- **lexical**: orthographic neighbors; normalized TTR; **number of meanings per words**.
- **syntactic**: verbal moods and tenses;
- **specific to FFL**: characteristics of MWE, **acquisition steps**.

Features in bold have not been implemented so far.
Machine learning algorithms

- **Regression models**: they depend on the type of the dependant variable
  - Continuous $\Rightarrow$ Linear regression
  - Ordinal $\Rightarrow$ Proportional odds model (OLR)
  - Categorical $\Rightarrow$ Multinomial logistic regression (MLR)

- Models based on **decision trees**:
  - Classification tree [Breiman et al., 1984]
  - Boosting [Freund and Schapire, 1996]
  - Bagging [Breiman, 1996]

- **Support Vector Machines** [Boser et al., 1992]
Plan

1. Brief review of readability and its issues

2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
Results in two steps

Our experimentation were conducted in two steps:

1. Evaluation of the predictive ability of variables used alone (= bivariate analysis).

2. Evaluation of the predictive ability of some combinations on variables (= modelisation step).

The goal: limit multicollinearity risks.
## Bivariate analysis: some variables

<table>
<thead>
<tr>
<th></th>
<th>Test6CE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
<td>( \rho )</td>
<td>( W(p) )</td>
<td>( F(p) )</td>
</tr>
<tr>
<td>X75FFFDCC</td>
<td>(-0.296^2)</td>
<td>(-0.627^3)</td>
<td>(&lt; 0,001)</td>
<td>0.089</td>
</tr>
<tr>
<td>X90FFFC</td>
<td>(-0.319^3)</td>
<td>(-0.641^3)</td>
<td>(&lt; 0,001)</td>
<td>(&lt; 0,001)</td>
</tr>
<tr>
<td>PAGoug_2000</td>
<td>0.593^3</td>
<td>0.597^3</td>
<td>(&lt; 0,001)</td>
<td>0.017</td>
</tr>
<tr>
<td>PA_Alterego1a</td>
<td>0.657^3</td>
<td>0.652^3</td>
<td>(&lt; 0,001)</td>
<td>(&lt; 0,001)</td>
</tr>
<tr>
<td>ML3</td>
<td>(-0.56^3)</td>
<td>(-0.546^3)</td>
<td>(&lt; 0,001)</td>
<td>(&lt; 0,001)</td>
</tr>
<tr>
<td>meanNGProb.G</td>
<td>0.382^3</td>
<td>0.407^3</td>
<td>0.011</td>
<td>0.05</td>
</tr>
<tr>
<td>NLM</td>
<td>0.479^3</td>
<td>0.483^3</td>
<td>0.028</td>
<td>0.084</td>
</tr>
<tr>
<td>NL90P</td>
<td>0.519^3</td>
<td>0.521^3</td>
<td>(&lt; 0,001)</td>
<td>0.022</td>
</tr>
<tr>
<td>NMP</td>
<td>0.486^3</td>
<td>0.618^3</td>
<td>(&lt; 0,001)</td>
<td>0.014</td>
</tr>
<tr>
<td>PRO.PRE</td>
<td>(-0.181^3)</td>
<td>(-0.345^3)</td>
<td>(&lt; 0,001)</td>
<td>0.226</td>
</tr>
<tr>
<td>PPRes</td>
<td>0.44^3</td>
<td>0.44^3</td>
<td>(&lt; 0,001)</td>
<td>0.003</td>
</tr>
<tr>
<td>Pres_C</td>
<td>(-0.355^3)</td>
<td>(-0.337^3)</td>
<td>(&lt; 0,001)</td>
<td>(&lt; 0,001)</td>
</tr>
<tr>
<td>PP1P2</td>
<td>(-0.408^3)</td>
<td>(-0.333^3)</td>
<td>(&lt; 0,001)</td>
<td>0.008</td>
</tr>
<tr>
<td>avLocalLsa_Lem</td>
<td>0, 63^3</td>
<td>0, 63^3</td>
<td>(&lt; 0,001)</td>
<td>0, 01</td>
</tr>
<tr>
<td>NAColl</td>
<td>/</td>
<td>0.286^3</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>BINGUI</td>
<td>0, 462^3</td>
<td>0, 462^3</td>
<td>(&lt; 0,001)</td>
<td>0, 018</td>
</tr>
</tbody>
</table>
Main results from the bivariate analysis

- Each family has at least one efficient predictor
  → idea: what if I design a formula with those variables?
- Among those, two are traditional ones (\texttt{PA\_Alterego1a} et \texttt{NMP}) and one is NLP-based (\texttt{avLocalLsa\_Lem}).
- Surprisingly, some other NLP-based features are poor predictors: N-gram models (where N>1), MWE-based features, etc.
- \textbf{Specialization}: the efficiency of \texttt{PA\_Alterego1a} provides a rationale for adapting readability models to specific contexts (list for FFL).
Design of the readability model

For the modelisation step, various combinations of predictors were attempted:

- Baseline (mimics classic formulas): NMP + NLM.
- Best predictor/family (4): PA_Altregero1a + NMP + avLocalLsa_Lem + BINGUI.
- 2 best predictors/family (8): PA_Altregero1a + X90FFFC + NMP + PPres + avLocalLsa_Lem + PP1P2 + BINGUI + NAColl.

→ Assumption: maximizing the type of information in a minimal set.

- Automatic selection of features.
  → Assumption: maximizing the quantity of information.

Each set was tested with the 6 statistical algorithms.
Design of the readability model (2)

Total corpus: about 2000 texts

Sampling process

108 texts/level

Development set

40 texts/level

outliers removed

LSA space definition

Meta-parameters: (grid-search) C, γ, cp, msplit

All estimates obtained with 10-fold cross-validation

Test set

68 texts/level

outliers removed

Modelling

Ten-fold cross-validation

Results: estimation
Models were evaluated with these 5 measures:

- Multiple correlation ratio ($R$).
- Accuracy ($acc$).
- Adjacent accuracy ($acc - cont$)
  → proportions of predictions that were within one level of the human-assigned level for the given text [Heilman et al., 2008]
- Root mean square error (RMSE).
- Mean absolute error (MAE).
Main results

<table>
<thead>
<tr>
<th>Model</th>
<th>Classifier</th>
<th>Parameters</th>
<th>$R$</th>
<th>acc</th>
<th>acc – cont</th>
<th>rmse</th>
<th>mae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>16%</td>
<td>44%</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Baseline SVM</td>
<td>γ = 0, 05; C = 25</td>
<td>0, 62</td>
<td>34%</td>
<td>68%</td>
<td>71%</td>
<td>1, 51</td>
<td>1, 06</td>
</tr>
<tr>
<td>Model 2009</td>
<td>/</td>
<td>/</td>
<td>0,62</td>
<td>41%</td>
<td>71%</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Expert1 RLM</td>
<td>/</td>
<td>/</td>
<td>0,70</td>
<td>39%</td>
<td>74, 2%</td>
<td>1, 34</td>
<td>0, 97</td>
</tr>
<tr>
<td>Expert2 SVM</td>
<td>γ = 0, 002; C = 75</td>
<td>0, 73</td>
<td>41%</td>
<td>78%</td>
<td>/</td>
<td>1, 28</td>
<td>0, 94</td>
</tr>
<tr>
<td>Auto-OLR OLR</td>
<td>/</td>
<td>/</td>
<td>0,71</td>
<td>39.6</td>
<td>76.1</td>
<td>1.33</td>
<td>0.96</td>
</tr>
<tr>
<td>Auto SVM</td>
<td>γ = 0, 004; C = 5</td>
<td>0, 73</td>
<td>49%</td>
<td>79, 6%</td>
<td>/</td>
<td>1.27</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Best model

- +32, 4% in comparison with random (acc);
- +8% in comparison with previous 2009 model (acc);
- Adjacent accuracy per level, computed on one of the 10 folds (mean is 79%)
Contribution of the variable families

We compared models either using only one family of predictors, or including all 46 features except those of a given family:

<table>
<thead>
<tr>
<th>Family</th>
<th>Accuracy</th>
<th>Adj. Accuracy</th>
<th>Accuracy</th>
<th>Adj. Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>40.5</td>
<td>75.6</td>
<td>41.1</td>
<td>73.5</td>
</tr>
<tr>
<td>Syntactic</td>
<td>39.3</td>
<td>69.5</td>
<td>43.2</td>
<td>78.4</td>
</tr>
<tr>
<td>Semantic</td>
<td>28.8</td>
<td>61.5</td>
<td>47.8</td>
<td>79.2</td>
</tr>
<tr>
<td>FFL</td>
<td>24.9</td>
<td>58.5</td>
<td>47.8</td>
<td>79.6</td>
</tr>
</tbody>
</table>

Lexical and then syntactic families reach the highest performance and yield the highest loss in accuracy.

Lexical features are the only ones to reduce the amount of critical mistakes (adj. acc.).
## Comparison with other studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[Kandel and Moles, 1958]</td>
<td>(rég.)</td>
<td>F.</td>
<td>33%</td>
<td>/</td>
<td>0.55</td>
<td>/</td>
</tr>
<tr>
<td>[Si and Callan, 2001]</td>
<td>3</td>
<td>E.</td>
<td>75, 4%</td>
<td>/</td>
<td>/</td>
<td>0.64</td>
</tr>
<tr>
<td>[Collins-Thompson and Callan, 2004]</td>
<td>6</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>0, 64</td>
<td>/</td>
</tr>
<tr>
<td>[Collins-Thompson and Callan, 2004]</td>
<td>12</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>0, 79</td>
<td>/</td>
</tr>
<tr>
<td>[Collins-Thompson and Callan, 2004]</td>
<td>5</td>
<td>F.</td>
<td>/</td>
<td>/</td>
<td>0, 64</td>
<td>/</td>
</tr>
<tr>
<td>[Schwarm and Ostendorf, 2005]</td>
<td>4</td>
<td>E.</td>
<td>/</td>
<td>79% à 94, 5%</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>[Heilman et al., 2007]</td>
<td>12</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>0, 72</td>
<td>2, 17</td>
</tr>
<tr>
<td>[Heilman et al., 2007]</td>
<td>4</td>
<td>E. (L2)</td>
<td>/</td>
<td>/</td>
<td>0, 81</td>
<td>0, 66</td>
</tr>
<tr>
<td>[Heilman et al., 2008]</td>
<td>12</td>
<td>E.</td>
<td>/</td>
<td>45%</td>
<td>0, 58</td>
<td>2, 94</td>
</tr>
<tr>
<td>[Heilman et al., 2008]</td>
<td>12</td>
<td>E.</td>
<td>/</td>
<td>52%</td>
<td>0, 77</td>
<td>2, 24</td>
</tr>
<tr>
<td>[Pitler and Nenkova, 2008]</td>
<td>5</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>0, 78</td>
<td>/</td>
</tr>
<tr>
<td>[François, 2009b]</td>
<td>6</td>
<td>F.  (L2)</td>
<td>41%</td>
<td>71%</td>
<td>0, 62</td>
<td>/</td>
</tr>
<tr>
<td>[François, 2009b]</td>
<td>9</td>
<td>F.  (L2)</td>
<td>32%</td>
<td>63%</td>
<td>0, 72</td>
<td>2, 24</td>
</tr>
<tr>
<td>Feng et al., 2009</td>
<td>4</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>−0, 34</td>
<td>0, 57</td>
</tr>
<tr>
<td>Feng et al., 2010</td>
<td>4</td>
<td>E.</td>
<td>70%</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Kate et al., 2010</td>
<td>5</td>
<td>E.</td>
<td>/</td>
<td>/</td>
<td>0, 82</td>
<td>/</td>
</tr>
<tr>
<td>6-classes model</td>
<td>6</td>
<td>F.  (L2)</td>
<td>49%</td>
<td>80%</td>
<td>0, 73</td>
<td>1, 23</td>
</tr>
</tbody>
</table>

[Kandel and Moles, 1958] is a general formula for L1 French → on our test data, its accuracy = 33%!

---

---
Where does this improvement come from?

There are mainly three reasons:

- Better features due to NLP-enabled extraction (see [François and Miltsakaki, 2012])
- Better training algorithms (see [François and Miltsakaki, 2012])
- Effect of specialization?
  → BUT our baseline (trained on specialized corpus) reaches 34% vs. 33%!

A 4th reason?

What if we test this model on a different FFL corpus?
1. Brief review of readability and its issues

2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results

3. Discussion: what level of specificity for readability?

4. References
Another corpus

We gathered manually another FFL corpus: simplified readers

- They are mainly narrative texts (a few are informative)
- No bias from the task on the text difficulty as in textbooks
- In France, readers might not so much have been “written to the formula” like in the U.S.
- Unfortunately, only series available from A1 to B2!
- We gathered 29 readers:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb. of readers</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>nb. of words</td>
<td>41018</td>
<td>71563</td>
<td>73011</td>
<td>59051</td>
</tr>
</tbody>
</table>
Features analysis

We first ran a bivariate analysis at the readers level:

- Previous best feature $\text{PA}_\text{Alterego}1b : r = 0.280$ vs. $0.657$
  → Probably due to a change of topic (eg. knights)
- Previously interesting BINGUI : $NA$ vs. 0, 462
  → All the texts contain dialogues, since there are narrative.
- Effect of text length:
  → $\text{normTTR}_W : r = 0.587$ vs. $r = 0.125$
  → Discrete tense-based features were not efficient anymore.
- High efficiency of continuous tense-based features:
  - Proportion of conditional tenses in all tenses : $r = 0.626$
  - Proportion of imperfect tenses in all tenses : $r = 0.760$
  - Proportion of present in all tenses : $r = -0.839$
Splitting the data

Problem
29 readers are not enough to train a readability model!

We split the books by chapters and got the following data:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb. of obs.</td>
<td>71</td>
<td>114</td>
<td>84</td>
<td>48</td>
</tr>
<tr>
<td>nb. of words</td>
<td>41018</td>
<td>71528</td>
<td>73007</td>
<td>59051</td>
</tr>
</tbody>
</table>

Correlations decrease but remain mostly coherent with previous figure.
Various experiments

"New" Corpus

Corpus readers A1 to B2 !!!

Corpus textbooks C1 and C2

NB: sampling is different

Sampling (48/lev.)
Taking out outliers

Old Corpus

Textbooks (68/level)

New Model

SVM(38)  R = 0.845; acc = 58.2%;
adj. acc. = 87.7%

SVM(41)  R = 0.85; acc. = 56%;
adj. acc. = 88%

SVM(41)  R = 0.72; acc. = 38%;
adj. acc. = 81.7%

SVM(41)  R = 0.72; acc. = 48%;
adj. acc. = 77.9%

Not available: meanNGProb.G,
NCPW, NAColl

Now constant: Infi (1) and
med_nbNeighMoreFreq (0)

Old Model
Conclusions

Some findings appeal for specialization of formulas depending on the type of texts:

- Loss of accuracy between both models for the same population! → However, if accuracy drops (−10%), adj. acc. remains more stable (+4%)  
- Lot of variations in the predictor power, which are related to the specific characteristics of the texts → Some features are even constant!  
- Obvious benefit of specializing the model: → Just retraining the same model on the new corpus: +8% (better coefficients) → Retraining + features selection: +10%  
- This also suggests that best path for improvement of readability models might be related to the training corpus. → due to higher homogeneity or just specialization?
We are currently studying a user-oriented way of getting labelled data (close to crowd-sourcing)

http://www.choosito.com/dmesure/index.php (Demonstration)

This should allow to get a lot of reliable data, but there is clearly a motivational problem involved.
Thank you for your attention.

Questions and comments are welcomed
Plan

1. Brief review of readability and its issues
2. A FFL “AI formula” aiming at specialization
   - Introduction
   - Methodology
   - Results
3. Discussion: what level of specificity for readability?
4. References
References I

A training algorithm for optimal margin classifiers.
In *Proceedings of the fifth annual workshop on Computational learning theory*,
pages 144–152.

Bagging predictors.
*Machine learning, 24*(2) :123–140.

*Classification and regression trees.*
Chapman & Hall, New York.

Collins-Thompson, K. and Callan, J. (2004).
A language modeling approach to predicting reading difficulty.
In *Proceedings of HLT/NAACL 2004*, pages 193–200, Boston, USA.
Predicting reading difficulty with statistical language models.

*Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.*
Hatier, Paris.

La lisibilité : essai d’application de la formule courte d’Henry au français langue étrangère.

A formula for predicting readability.
*Educational research bulletin*, 27(1) :11–28.
References III


References V

Experiments with a new boosting algorithm.

Combining lexical and grammatical features to improve readability measures for first and second language texts.

An analysis of statistical models and features for reading difficulty prediction.

*Comment mesurer la lisibilité.*
Labor, Bruxelles.


References VII

A statistical model for scientific readability.
In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 574–576. ACM New York, NY, USA.

Spache, G. (1953).
A new readability formula for primary-grade reading materials.

Tharp, J. (1939).
The Measurement of Vocabulary Difficulty.

Readability of French as a foreign language and its uses.