Dmesure and FLELex: two approaches of complexity for French as a foreign language

Thomas François

Seminar in Språkbanken, Göteborgs universitet
February 4th, 2015
Plan

1. Introduction

2. DMesure: a readability model for FFL

3. FLELex: a graded lexicon for FFL
Plan

1. Introduction
   - DMesure : a readability model for FFL
     - Introduction to readability formulas
     - Conception of DMesure
     - Results
   - FLELex : a graded lexicon for FFL
     - Introduction : theoretical background
     - The making of FLELex
The challenge of reading

Reading remains a challenge for a significant part of the population, even in our highly educated societies:

- UE recent report (2009) : 19.6% of 15 year old teenagers are “low achievers” [De Coster et al., 2011, 22]
- [Richard et al., 1993] : On 92 unemployment benefit form filled by people with a low education level, half of the required information was missing.
- [Patel et al., 2002] : Their subjects faced significant problems in understanding the different steps for the proper administration of drugs.
- Besides, reading is also an issue for the large amount of L2 learners faced with written texts (in lecture, administration, web, etc.)
Reading and NLP

Natural language processing can help low readers in various ways:

- Automatic selection of reading materials at their level (readability);
- Automatic generation of reading or language exercises;
- Integration within iCALL software for intelligent feedback, better adaptability or incremental content collection;
- Automatic text simplification (ATS) to improve access to authentic texts;
- Assistive writing tools for writers, etc.
In this talk, we focus on two NLP-enabled resources of CALL, specialized for L2 reading:

**DMesure**: a readability formula for FFL

- DMesure is a readability model that is able to associate a text with a CEFR level (for a reading task).

**FLELex**: a CEFR-graded lexicon for FFL

- FLELex provides a lexicon for which the distribution of each word across CEFR levels is described.
Plan

1. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
Plan

1. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
What is readability?

**Origin**: Readability dates back to the 20s, in the U.S. It is only after 1956 that it spread in the French-speaking community.

**Objective**: Aims to assess the difficulty of texts for a given population, without involving 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].
Example of a formula

Formula of [Dale and Chall, 1948, 18] :

\[ X_1 = 3.6365 + 0.1579 X_2 + 0.0496 X_3 \]

where :

- \( X_1 \) : mean grade level for a schoolchild that would be able to get at least 50% to a comprehension test on this text.
- \( X_2 \) : percentage of words not in the list of Dale (3000 words).
- \( X_3 \) : mean number of word per sentence.

The independant variables \( X_2 \) and \( X_3 \) are the predictors or features.
What are the use for readability formulas?

Readability formula have been used for:

- Selection of materials for textbooks.
- Used in scientific experiments to control the difficulty of textual input data.
- Controlling the difficulty level of publications from various administrations (justice, army, etc..) and newspapers.
- More recently, checking the output of automatic summarization, machine translation, etc. [Antoniadis and Grusson, 1996, Aluisio et al., 2010, Kanungo and Orr, 2009].
What about readability formulas for FFL?

Common approach for foreign language contexts: apply formula designed for natives [Cornaire, 1985]

→ Denial of the specific process of L2 reading.

This approach relies on three suspect assumptions:

- the understanding of readers in the L2 is comparable to that of native speakers.
- the textual features considered in L1 formulas are relevant to L2 reading (and the only relevant ones).
- the weighting of these variables can be the same in a formula for L1 and L2.
An alternative: consider the specificities of the L2 context

Some studies took into account those specificities, described by [Koda, 2005], into readability models:

- [Tharp, 1939] positions himself against the previous approach and offers one of the first specific formulas for FLE, based on cognates.
- [Uitdenbogerd, 2005] suggests a formula that also takes into account cognates:
  \[ FR = 10 \times WpS - Cog \]
  - \( WpS \): mean number of word per sentence.
  - \( Cog \): number of cognates per 100 words.
- [Heilman et al., 2007] compare the efficiency of lexical and syntactic features in L1 and L2 context:
  → grammatical features play a more important role in a L2 model.
Plan

1. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
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.

\[ Y = B_2 A_1 + C_2 A_2 + \ldots \]

- \( X_{i1} = -748.7 \)
- \( X_{i2} = 5.32 \)
- \( \ldots \)
- \( X_{in} = 1 \)

Statistical model

Prediction on a new text

\[ \text{Prediction} = B_2 \]
Conception

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>Series</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>/</td>
<td>/</td>
<td>34</td>
<td>53</td>
<td>39</td>
<td>50</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>/</td>
<td>31</td>
<td>78</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><strong>Total</strong></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**: personnalisation 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.
Conception

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 dependent 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. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
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>Variable</th>
<th>Test6CE $r$</th>
<th>Test6CE $\rho$</th>
<th>$W(p)$</th>
<th>$F(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X75FFFDC</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 (PA_Alterego1a et NMP) and one is NLP-based (avLocalLsa_Lem).
- Surprisingly, some other NLP-based features are poor predictors: N-gram models (where N>1), MWE-based features, etc.
- **Specialization**: the efficiency of 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_Alterego1a + NMP + avLocalLsa_Lem + BINGUI.
- 2 best predictors/family (8) : PA_Alterego1a + 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)

Sampling process

Total corpus: about 2000 texts

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 Model</td>
<td>SVM</td>
<td>$\gamma = 0.05; C = 25$</td>
<td>0.62</td>
<td>34%</td>
<td>68%</td>
<td>1.51</td>
<td>1.06</td>
</tr>
<tr>
<td>Expert1</td>
<td>RLM</td>
<td>/</td>
<td>0.70</td>
<td>39%</td>
<td>74%</td>
<td>1.34</td>
<td>0.97</td>
</tr>
<tr>
<td>Expert2</td>
<td>SVM</td>
<td>$\gamma = 0.002; C = 75$</td>
<td>0.73</td>
<td>41%</td>
<td>78%</td>
<td>1.28</td>
<td>0.94</td>
</tr>
<tr>
<td>Auto-OLR</td>
<td>OLR</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</td>
<td>SVM</td>
<td>$\gamma = 0.004; C = 5$</td>
<td>0.73</td>
<td>49%</td>
<td>79.6%</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%)

<table>
<thead>
<tr>
<th>Adj. acc.</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
<td>71%</td>
<td>67%</td>
<td>71%</td>
<td>86%</td>
<td>83%</td>
</tr>
</tbody>
</table>
**Results**

**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></th>
<th>Family only</th>
<th></th>
<th>All except family</th>
<th></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>

**Results**

- 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.).
First conclusions

- It is the first specific formula for FFL that uses a NLP approach (and one of the few for FFL)
  - The corpus includes a variety of text types, ensuring a wider coverage to the formula
- The criterion used (level of the textbooks according to the CEFR scale) appears questionable: the noise in the corpus can cause a poor learning.
- Our experiments suggest the (slight) superiority of SVM and logistic regression, a technique which is less demanding than the first.
An example: AMesure

AMesure is a free web platform that assess the difficulty of administrative texts:

- Includes a readability formula that classifies texts on a 1-to-5 scale;
- Trained on a small corpus of 115 texts (annotated by FWB experts);
- Selection of 11 variables among 344: model reaches $acc = 50\%$ and $adj - acc = 86\%$;
- Besides the formula, lexical and syntactic diagnosis is provided.
Plan

1. Introduction

2. DMesure : a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex : a graded lexicon for FFL
   - Introduction : theoretical background
   - The making of FLELex
Plan

1. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
The issue of vocabulary

Vocabulary and L2 learning

- Vocabulary knowledge is crucial for L2 learning and a reader must know between 95% to 98% of the words in a text to adequately understand it [Hu and Nation, 2000]

- In readability formulas, the lexical features have been shown to account the most for text difficulty [Chall and Dale, 1995]

- Control the level of vocabulary in a text is therefore valuable for learning...

- It can also be useful for other tasks, such as text simplification.

In the second section of this talk, we aim at assessing the difficulty of the lexicon
Psycholinguistic investigates the complexity of words through various dimensions:

- Word frequency effect: correlation between frequency of words and difficulty [Brysbaert et al., 2000]
- The age-of-acquisition seems to play a role in decoding, independently of the word frequency [Gerhand and Barry, 1999]
- The number of orthographic neighbours [Andrews, 1997]
- Concreteness and imageability of words [Schwanenflugel et al., 1988]
- The familiarity of readers with words (and morphemes) also helps recognition [Gernsbacher, 1984]
- The number of (known) senses [Millis and Button, 1989]
Approaches in L2 learning and teaching

- There is also a bunch of studies in vocabulary learning that correlates words characteristics with ease of learning.

- [Laufer, 1997] focused on factors such as familiarity of phonemes, regularity in pronunciation, fixed stress, consistency of the sound-script relationship, derivational regularity, morphological transparency, number of meanings, etc.

- Another approach is to defined graduated lexicon lists on which the learning process and materials selection can be based.
  → Question: how are these lists obtained?
Frequency lists

One of the first lists was collected by [Thorndike, 1921]: list of 10,000 words with frequencies computed from a corpus of 4,500,000 words.

[Henmon, 1924]: *French Word Book*
→ These lists were defined from frequencies (based on the word frequency effect) in the general language.

Several issues are inherent to this approach:

- Frequency estimation is not always robust ([Thorndike, 1921]: second half of the list less robust)
- [Michéa, 1953] highlighted that some common words in language (available words) are not well estimated.
- Not obvious how to transform frequencies into educational levels.

Frequency lists are not really educationally-graded resources!
Graded lists

- Graded list for L1 French is Manulex [Lètè et al., 2004] :
  - About 23,900 lemmas whose distributions have been estimated on primary schoolbooks.
  - The corpus includes 54 textbooks from CP (6 years) to CM2 (11 years)
  - Three levels were defined : CP is 1 ; CE1 is 2 and 3 spans from CE2 to CM2.

<table>
<thead>
<tr>
<th>Word</th>
<th>Pos</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pomme</td>
<td>N</td>
<td>724</td>
<td>306</td>
<td>224</td>
</tr>
<tr>
<td>vieillard</td>
<td>N</td>
<td>-</td>
<td>13</td>
<td>68</td>
</tr>
<tr>
<td>patriarche</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>cambrioleur</td>
<td>N</td>
<td>2</td>
<td>-</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>31%</td>
<td>21%</td>
<td>48%</td>
</tr>
</tbody>
</table>
A current reference for L2 learning is the CEFR referentials [Beacco and Porquier, 2007]

They give more precisions than the CEFR about the specific lexical skills to learn, but...

No distinctions are made between words within a level

The format is not suitable for NLP approaches

Concerns has been raised as regards the validity of these referentials (e.g. KELLY, VALILEX)
What did we learn?

- It is acknowledged that it is possible to relate a word difficulty with some of its characteristics
- Current approaches generally focus on one or a few characteristics → ReSyf
- No graded resource (such as Manulex) for L2 context → FLELex

Collaborators

Nuria Gala, Cédrick Fairon, Patrick Watrin, Anaïs Tack
Plan

1. Introduction

2. DMesure: a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results

3. FLELex: a graded lexicon for FFL
   - Introduction: theoretical background
   - The making of FLELex
Objectives of the FLELex project

- Offer a lexical resource describing the distribution of French words in FFL textbooks.
  → Textbooks using the CEFR scale, we get a distribution of words across the 6 levels of the CEFR.

- This distribution is learned from a corpus and the frequencies are adapted for a better estimation.

- Possible uses:
  - Targetted vocabulary learning (which word to learn at which level)
  - Comparing the frequency of usage of synonyms
  - Using it within a language model for various iCALL tasks (readability, etc.)
  - Apply it for automatic text simplification (ATS)
Methodology

1. Collect a corpus of texts from FFL textbooks
2. Tag the corpus to desambiguate forms as regards part-of-speeches
3. Compute normalized frequencies, with an adequate estimator
4. Exploring the resource
The training corpus

We collected 28 textbooks and 29 simplified books, amounting to a total of 2,071 texts and 777,000 words.

<table>
<thead>
<tr>
<th>Genre</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue</td>
<td>153 (23,276)</td>
<td>72 (17,990)</td>
<td>39 (11,140)</td>
<td>5 (1,698)</td>
</tr>
<tr>
<td>E-mail, mail</td>
<td>41 (4,547)</td>
<td>24 (2,868)</td>
<td>44 (11,193)</td>
<td>18 (4,193)</td>
</tr>
<tr>
<td>Sentences</td>
<td>56 (7,072)</td>
<td>21 (4,130)</td>
<td>12 (1,913)</td>
<td>5 (928)</td>
</tr>
<tr>
<td>Varias</td>
<td>31 (3,990)</td>
<td>36 (4,439)</td>
<td>23 (5,124)</td>
<td>14 (1,868)</td>
</tr>
<tr>
<td>Text</td>
<td>171 (23,707)</td>
<td>325 (65,690)</td>
<td>563 (147,603)</td>
<td>156 (63,014)</td>
</tr>
<tr>
<td>Readers</td>
<td>8 (41,018)</td>
<td>9 (71,563)</td>
<td>7 (73,011)</td>
<td>5 (59,051)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>460 (103,610)</td>
<td>487 (166,680)</td>
<td>688 (249,984)</td>
<td>203 (130,752)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genre</th>
<th>C1</th>
<th>C2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue</td>
<td>/</td>
<td>/</td>
<td>269 (54,104)</td>
</tr>
<tr>
<td>E-mail, mail</td>
<td>8 (2,144)</td>
<td>1 (398)</td>
<td>136 (25,343)</td>
</tr>
<tr>
<td>Sentences</td>
<td>/</td>
<td>/</td>
<td>94 (14,043)</td>
</tr>
<tr>
<td>Varias</td>
<td>1 (272)</td>
<td>/</td>
<td>105 (15,693)</td>
</tr>
<tr>
<td>Text</td>
<td>175 (89,911)</td>
<td>48 (34,084)</td>
<td>1,438 (424,009)</td>
</tr>
<tr>
<td>Readers</td>
<td>/</td>
<td>/</td>
<td>29 (244,643)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>184 (92,327)</td>
<td>49 (34,482)</td>
<td>2,071 (777,835)</td>
</tr>
</tbody>
</table>
The tagging process

**Goal**: obtain the lemma of every form observed in the corpus and disambiguate homographic forms with different P.O.S.

→ Using inflecting forms would imply splitting frequency density across several forms.

→ It would also imply that we consider learners unable to relate inflected forms.

**Problem**: The tagger precision matters, otherwise we can get:

- entries with wrong part-of-speech tag (e.g. *adoptez* PREP or *tu* ADV);
- entries with a non attested lemma (e.g. *faire partir* instead of *faire partie*);
- likely tags that but are erroneous in the specific context of the word.
One well-known limitation of taggers is their ability to extract multi-word expression (MWE)!

MWEs include a set of heterogeneous linguistic objects (collocations, compound words, idioms, etc.)

Learner’s knowledge of MWE lags far behind their general vocabulary knowledge [Bahns and Eldaw, 1993] → Therefore, including such linguistic forms in a graded-lexicon for FFL purposes appears as crucial!
The selected taggers

We selected two taggers and compared their performance:

**TreeTagger**
- TreeTagger [Schmid, 1994] is widely used and acknowledged
- Easy to use (wrappers exist for various programming languages)
- Not anymore state-of-the-art performance and cannot detect MWEs

**a CRF-based tagger**
- CRF-taggers are state-of-the-art and can be trained to detect MWEs
- We used one drawing from the work of [Constant and Sigogne, 2011] and developed by EarlyTracks.
Computing the distributions

- We used the dispersion index [Carroll et al., 1971]

\[
D_{w,K} = \frac{\log(\sum p_i) - \sum p_i \log(p_i)}{\sum p_i} / \log(I) \tag{1}
\]

\(K = \) CEFR level; \(I = \) number of textbooks in level \(K\);
\(p_i = \) word probability in textbook \(i\).

- Then, raw frequencies are normalized as follows:

\[
U = \left( \frac{1\,000\,000}{N_k} \right) [RFL \times D + (1 - D) \times f_{min}] \tag{2}
\]

where \(N_k = \) number of tokens at level \(k\);
\(f_{min} = \frac{1}{N} \sum f_i s_i\) with \(f_i = \) word frequency in textbook \(i\) and \(s_i = \) number of words in textbook \(i\).
The two FLELex

We got two different versions of FLELex

**FLELex-TT**
- Includes 14,236 entries, but no MWEs!
- It is based on Treetagger and is easy to use for NLP purposes
- It has been manually checked

**FLELex-CRF**
- Includes 17,871 entries, among which several thousands of MWEs
- Better performance means better estimations of frequency distributions, but segmentation errors yields to a few odd entries
- Not manually cleaned (so far)
### Example of entries

<table>
<thead>
<tr>
<th>lemma</th>
<th>tag</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiture (1)</td>
<td>NOM</td>
<td>633.3</td>
<td>598.5</td>
<td>482.7</td>
<td>202.7</td>
<td>271.9</td>
<td>25.9</td>
<td>461.5</td>
</tr>
<tr>
<td>abandonner (2)</td>
<td>VER</td>
<td>35.5</td>
<td>62.3</td>
<td>104.8</td>
<td>79.8</td>
<td>73.6</td>
<td>28.5</td>
<td>78.2</td>
</tr>
<tr>
<td>justice (3)</td>
<td>NOM</td>
<td>3.9</td>
<td>17.3</td>
<td>79.1</td>
<td>13.2</td>
<td>106.3</td>
<td>72.9</td>
<td>48.1</td>
</tr>
<tr>
<td>kilo (4)</td>
<td>NOM</td>
<td>40.3</td>
<td>29.9</td>
<td>10.2</td>
<td>0</td>
<td>1.6</td>
<td>0</td>
<td>19.8</td>
</tr>
<tr>
<td>logique (5)</td>
<td>NOM</td>
<td>0</td>
<td>0</td>
<td>6.8</td>
<td>18.6</td>
<td>36.3</td>
<td>9.6</td>
<td>9.9</td>
</tr>
<tr>
<td>en bas (6)</td>
<td>ADV</td>
<td>34.9</td>
<td>28.5</td>
<td>13</td>
<td>32.8</td>
<td>1.6</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>en clair (7)</td>
<td>ADV</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8.2</td>
<td>19.5</td>
<td>1.2</td>
</tr>
<tr>
<td>sous réserve de (8)</td>
<td>PREP</td>
<td>0</td>
<td>0</td>
<td>0.361</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The resource is freely available at

http://cental.uclouvain.be/flelex/
A few figures about FLELex

- A majority of the words are nouns in both lists (respectively 51% and 55%)
- TT-version includes 33% of hapaxes while only 26% of the entries have 10 occurrences or more.
- CRF-version includes 20% of hapaxes while 31% of the entries have 10 occurrences or more.
- We compared FLELex-TT with another lexicon : Lexique 3 [New et al., 2004]
  - Only 622 entries of FLELex-TT were missing from Lexique 3
- Correlation between total frequencies in FLELex-TT and Lexique3 is high : 0.84
Démonstration
FLELex

Perspectives

- Manually clean the CRF version
- Add a tab to the web site that would allow to directly analyze a text
- Use FLELex to predict the known/unknown vocabulary of a given reader
- Offer “FLELex” versions for other languages (currently on-going work for Swedish and perspectives for Spanish)
Difficulté estimée : A2

Votre texte :

Merci pour votre attention.

Sachez que les questions et les commentaires sont les bienvenus :-)

The end
References I


References II

*Niveau A1 pour le français : utilisateur-apprenant élémentaire.*
Didier.

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.

The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language.
References III


References IV


References V


References VI


References VII


References IX

Le traitement de documents administratifs par des populations de bas niveau de formation.

Probabilistic part-of-speech tagging using decision trees.
In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

Context availability and lexical decisions for abstract and concrete words* 1.

Measuring reading comprehension with the lexile framework.
In *Fourth North American Conference on Adolescent/Adult Literacy*. 

---

**CENTAL**
References X

