Challenges to specialize readability formulas: a case study on administrative texts

Thomas François¹

¹ CENTAL, IL&C (Université catholique de Louvain)

Seminar für sprachwissenschaft, Universität Tübingen

July, 11th 2014
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
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].
Example of the formula of [Flesch, 1948, 225]:

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

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)
- \(\text{wl}\): number of syllables per 100 words
- \(\text{sl}\): 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.
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.

The statistical model is given by:

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

- \( B_2 \) is the intercept.
- \( A_1, C_2, A_2, \ldots \) are the coefficients for the predictors.

The predictor values are:

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

Prediction on a new text

Statistical model
Some trends in the field

Readability is mostly an Anglo-Saxon field:

- First formulas appeared in the US: they considered only the lexicon. [Lively and Pressey, 1923, Vogel and Washburne, 1928]
- Classic formulae: they are based on linear regression and only 2 predictors (one lexical, one syntactic) [Flesch, 1948, Dale and Chall, 1948]
- The revolution of the cloze test: more complex formulae appeared as well as the first computational efforts. [Smith and Senter, 1967, Bormuth, 1966]
- The cognitive area corresponds to a critique of the classical formulae, unable to take into consideration some more semantic aspects (coherence, cohesion...). [Kintsch and Vipond, 1979, Kemper, 1983]
Recent works: “AI readability”

- This new trend in readability rose with the 21st century [Foltz et al., 1998, 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 any more as a regression one!
- NLP and machine learning processing require a large corpus!
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
Some issues in readability

- Performance are not as good as in other fields (and they depend much on corpus characteristics)
- Few annotated data available and annotations are often questionable
- Lots of features, but not all are that useful
- Mostly generic formulas are designed that do not take into account users’ specificities and context of use
The performance

- Performance remains unsatisfactory for commercial usage in most studies!

<table>
<thead>
<tr>
<th>Étude</th>
<th># cl.</th>
<th>lg.</th>
<th>Acc.</th>
<th>Adj. Acc.</th>
<th>R</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<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>[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>[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>[François, 2011]</td>
<td>6</td>
<td>F. (L2)</td>
<td>49%</td>
<td>80%</td>
<td>0.73</td>
<td>1.23</td>
</tr>
<tr>
<td>[François, 2011]</td>
<td>9</td>
<td>F. (L2)</td>
<td>35%</td>
<td>65%</td>
<td>0.74</td>
<td>1.92</td>
</tr>
<tr>
<td>[Vajjala and Meurers, 2012]</td>
<td>5</td>
<td>E.</td>
<td>93.3%</td>
<td>/</td>
<td>/</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Comparison between various models in [Nelson et al., 2012]:
  - Best model from [Nelson et al., 2012] is SourceRater
    - [Sheehan et al., 2010]
      - $\rho = 0.860$ on Gates-MacGinita corpus
  - REAP achieve lower scores than classic models, such as DRP or Lexile.
The corpus issue

- Very few corpora available: Weekly Reader is mostly used [Schwarm and Ostendorf, 2005, Feng et al., 2010, Vajjala and Meurers, 2012] → risk: high dependence towards one training corpus, as McCall and Crabbs lessons in classic period [Stevens, 1980]

- This dependence has consequences:
  - formulas will be specialized towards this corpus (coefficients)
  - always the same population and type of texts considered

  - SourceRater on smaller ranges: performance decrease drastically
    \[0.21 < \rho < 0.45\] on SAT-9 corpus
  - REAP model achieves \(\rho = 0.543\) on Common Core (informative), but only \(\rho = 0.292\) on narrative

No generic formulas work for all problems
## Quality of annotations in the corpus

<table>
<thead>
<tr>
<th></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>/</td>
<td>/</td>
<td>-746</td>
<td>-763</td>
<td>-766</td>
<td>-787</td>
</tr>
<tr>
<td>-705</td>
<td>-723</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>-749</td>
<td>-757</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-690</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>-758</td>
<td>-766</td>
<td>-777</td>
</tr>
<tr>
<td>-694</td>
<td>/</td>
<td>-746</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-725</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-696</td>
<td>-730</td>
<td>-753</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-731</td>
<td>-742</td>
<td>-733</td>
<td>-766</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>-787</td>
<td>-778</td>
</tr>
<tr>
<td>-664</td>
<td>-712</td>
<td>-756</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-711</td>
<td>-740</td>
<td>-752</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-683</td>
<td>-740</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>-700.09</td>
<td>-732.9</td>
<td>-750.75</td>
<td>-763.52</td>
<td>-771</td>
<td>-779</td>
<td></td>
</tr>
</tbody>
</table>
Other types of judgements

- [van Oosten et al., 2011] had 105 texts assessed by experts (as pairs) and clustered them by similarity of judgements (train one model per cluster).
  → this leads to different models, whose intracluster performance > intercluster.

- We had 18 experts annotate 105 administrative texts (with an annotation guide)
  → $0.10 < \alpha < 0.61$ per batch (average = 0.37).

- High agreement seems difficult to reach in readability (SemEval 2012: $\kappa = 0.398$ on the test set).
  "content analysis researchers generally think of $K > .8$ as good reliability, with $.67 < K < .8$ allowing tentative conclusions to be drawn"

- [Krippendorff, 1980, 167]
Some issues : features

- Although theoretically appealing, the effect of semantic and discourse features is questionable

- Review of cohesion measures [Todirascu et al., 2013] :
  - [Bormuth, 1969] tested 10 classes of anaphora (proportion, density, and mean distance between anaphora and antecedent) // two latter features were the best: \( r = 0.523 \) and \( r = -0.392 \) (\( r = -0.605 \) word/sent.)
  - [Kintsch and Vipond, 1979] : the mean number of inferences required in a text is not well correlated
  - [Pitler and Nenkova, 2008] : LSA-based intersentential coherence (\( r = 0.1 \)) and 17 features based discourse entities transition matrix were not significant.
  - [Pitler and Nenkova, 2008] : texts as a bag of discourse relations is a significant variable (\( r = 0.48 \))
An experiment with reference chains features

In [Todirascu et al., 2013], we annotated 20 texts across CEFR levels A2-B2 as regards reference chains.

We computed 41 variables, among which:
- POS-tagged based features (e.g. ratio of pronouns, articles, etc.)
- lexical semantic measures of intersentential coherence, based on tf-idf VSM or LSA
- Entity coherence [Pitler and Nenkova, 2008] : counting the relative frequency of the possible transitions between the four syntactic functions (S, O, C and X)
- Measures of the entity density and length of chains
- New features : Proportion of the various types of expressions included in a reference chain (e.g. indefinite NP, definite NP, personal pronouns, etc.

We show that a few variables based on reference chains are significantly correlated with difficulty, even on a small corpus

<table>
<thead>
<tr>
<th>Variable</th>
<th>Corr. and p-value</th>
<th>Variable</th>
<th>Corr. and p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.PRON</td>
<td>−0.59 (p = 0.005)</td>
<td>3.Pers.Pro. /S</td>
<td>−0.41 (p = 0.07)</td>
</tr>
<tr>
<td>33.Indef NP</td>
<td>−0.50 (p = 0.02)</td>
<td>10.Names /W</td>
<td>−0.4 (p = 0.08)</td>
</tr>
<tr>
<td>18.S → O</td>
<td>0.46 (p = 0.04)</td>
<td>9. nb. def. art. /W</td>
<td>0.38 (p = 0.1)</td>
</tr>
<tr>
<td>22. O → O</td>
<td>−0.44 (p = 0.048)</td>
<td>17. S → S</td>
<td>−0.36 (p = 0.12)</td>
</tr>
</tbody>
</table>
Classical features vs. NLP-based features

Contrasted results

- Several “AI readability” models were reported to outperform classic formulas.
- [Aluisio et al., 2010, François, 2011]: best correlate is a classic feature (av. W/S ; % of W not in a list)
- [François et al., 2014]: best correlate is mean number of words per sentence...

Comparing both types of information

- [François and Miltsakaki, 2012] compared SVM models with the same number of features (20), some are “classical“ and the others NLP-based → ”Classical“: acc. = 38% vs. NLP-based: acc. = 42% (t(9) = 1.5; p = 0.08)!
- When both types are combined within a SVM model, performance rise from acc. = 37.5% to 49%. 

Genericity of formulas

Today, we no longer believe in the universalist approach of classical models (Flesch, etc.)
→ specific population are considered (L2 readers, language-impaired readers, etc.)

However, the type of texts is often neglected

[Nelson et al., 2012] distinguishes between performance on narrative and informative texts
Type of texts: an experiment

We gathered another FFL corpus: simplified readers from A1 to B2
→ Mostly narrative texts, no bias from the task

29 simplified readers collected:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb. of books</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>

We divided the books by chapters and obtained the following training 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>
Typological experiment

"New" Corpus

Corpus readers A1 to B2 !!!

Corpus textbooks C1 and C2

NB: sampling is different

Sampling (48/lev.)
Taking out outliers

Textbooks (68/level)

New Model

SVM(38)
$R = 0.845; \text{acc} = 58.2\%$
adj. acc. = 87.7%

SVM(41)
$R = 0.85; \text{acc} = 56\%$
adj. acc. = 88%

SVM(41)
$R = 0.72; \text{acc} = 38\%$
adj. acc. = 81.7%

SVM(41)
$R = 0.72; \text{acc} = 48\%$
adj. acc. = 77.9%

Old Corpus

Old Model

Not available: meanNGProb.G, NCPW, NAColl
Now constant: Infi (1) and med_nbNeighMoreFreq (0)
What have we learned from this?

- Performance slightly increase, but still need to improve before readability reach a large public.
- Experts judgements is mainstream in the field, but reliability of such annotations is questionable.
- Reference corpora allows for better comparability of models, but run the risk of formatting the field.
  - → Penn Treebank “might” be representative of the English language, but Weekly Reader is not representative of all readers and texts.
- No generic readability models account for all problems, but the benefit of specialized formulas (for specific populations and texts) is yet to demonstrate.
- Classic features remains strong predictors of text difficulty, but can be combined with some benefit with NLP-based features
- Specialisation of readability models should be a major concern!
Specializing a formula

What is exactly specialization?

This consists in fitting a model in relation to a specific population of interest (children, L2 readers, etc.), to a specific context of use (type of texts, type of reading, etc.).

Practically, it requires:

- Use a corpus whose difficulty was assessed against this population to tune the model parameters.
- Adapt known predictors to this context (e.g. Alter Ego list)
- Find specific predictors to this population and task (e.g. MWE in [François and Watrin, 2011])
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
Annotate a specialized corpora

A specialized corpus for a specialized formula requires:

- Gathering authentic texts actually used by the target population
- Difficulty measures for the texts, obtained from this population
Annotation methods in readability

Expert judgements  heterogeneity, population not tested, but practical

Comprehension test  population tested, but interaction between questions and texts
  → Davis (1950) : performance differs when questions are asked in a simple or complex vocabulary

Cloze test  population tested, at the word level, but the relation with comprehension is questionable (redundancy ?)

Reading speed  
  [Brown, 1952] compared reading time on difficult texts (306 words/min.) and very hard (235 words/min).
  [Just and Carpenter, 1980] : ocular fixation time of a word corresponds to cognitive processing time.

Non expert judgements  [van Oosten and Hoste, 2011] showed that N (N > 10) non experts can annotated as reliably as experts (binary judgements).
Reading time as criterion: experiments

Reading time is used very little and yet might be the most psychologically reliable criterion.

Methodology

- 28 short texts (100 words), selected from simplified books of levels A1 to B2.
- Presentation of the sentences, one after the other, via a self-presentation software (Linger, MIT)
- The time spent on each sentence is registered; no return back is allowed.
- At the end, one or two comprehension questions check that text was read and understood.
- Results were analysed with a mix-effect model [Baayen et al., 2008] (to suppress the inter-subject variability)
Web interface

We also developed a web interface to administrate the same test on-line (crowdsourcing)
Web interface

Interface showing an example of questions (MCQ)

Question 1:
Quel marché propose Méléagant au roi Arthur ?

- Il veut combattre le chevalier le plus brave d'Arthur et, s'il gagne, prendre la reine Guenièvre avec lui.
- Il veut combattre le roi Arthur et, s'il perd, il rendra les prisonniers qu'il retient dans son royaume.
- Il veut échanger des prisonniers contre la reine.

Check your answers
## Results

<table>
<thead>
<tr>
<th></th>
<th>Linger</th>
<th>DMeasure-Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min-Max RT/W</td>
<td></td>
</tr>
<tr>
<td>Beginners II</td>
<td>717 ms – 78680 ms</td>
<td>562 ms – 45351 ms</td>
</tr>
<tr>
<td>Intermediate I</td>
<td>747 ms – 69250 ms</td>
<td>1296 ms – 61770 ms</td>
</tr>
<tr>
<td>Natives</td>
<td>493 ms – 33050 ms</td>
<td></td>
</tr>
<tr>
<td>nb. subjects</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Corr.</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.579</td>
</tr>
</tbody>
</table>

The method reliability increases as the skill level of readers increase.

When data are normalized at the character level, correlations decreases!
Conclusion

- Various studies tend to show the interest of specialized formulas (population, type of text)
- To investigate this question, it is vital to have a reliable and rapid annotation system for text difficulty
  → Few work in this direction. [van Oosten et al., 2011] suggest crowd-sourcing with non-experts
- We investigated an alternative method: reading time as a criterion
  → The reliability of the method seems good for native readers, but still need to be confirmed for L2 readers.
Perspectives

- Compare more strictly the effect of the type of texts on model performance
- Check that the specialization is useful, but what level of granularity?
- Try other measurement techniques for reading time (at the paragraph level, other task?)
- Adapt the interface in a “serious game” perspective.
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
Administrative texts are known to be difficult to access for a significant proportion of the population.

Our aim: provide a readability formula that classifies administrative texts on a scale, from 1 (very easy) to 5 (very difficult).

Main issue: no training corpus available...

- The popular way in NLP-based readability = take educational texts, already annotated by textbook designers
- No resources of this type for administrative texts!
What type of annotation?

Mixed annotation: reading speed and expert judgements.

- 115 authentic administrative texts (FWB) were scanned (XML) and cut into 220 excerpts.
- The difficulty of the fragments was assessed via the formula by [Kandel and Moles, 1958]
- Sampling of 115 texts across “levels”, to ensure a good representativity of difficulty.
- 10 texts with various scores were selected and tested via AMesure-Testing
- Correlation between ms./word and score KM is good ($r = 0.74$).
# Reading speed data

## Mean reading time per text

<table>
<thead>
<tr>
<th>Text title</th>
<th>KM score</th>
<th>ms. /word</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>La santé de votre enfant</td>
<td>71.3</td>
<td>292.8</td>
<td>1</td>
</tr>
<tr>
<td>Du couple à la famille</td>
<td>86.5</td>
<td>304.9</td>
<td>1</td>
</tr>
<tr>
<td>Des chaussures... Quand les mettre aux pieds ?</td>
<td>81.1</td>
<td>315</td>
<td>2</td>
</tr>
<tr>
<td>A l’école d’une alimentation saine</td>
<td>75.8</td>
<td>324.4</td>
<td>2</td>
</tr>
<tr>
<td>L’enseignement spécialisé</td>
<td>46.2</td>
<td>339.7</td>
<td>3</td>
</tr>
<tr>
<td>Lettre pour la semaine européenne de la vaccination</td>
<td>40.6</td>
<td>340.5</td>
<td>3</td>
</tr>
<tr>
<td>Cumuls de pensions</td>
<td>57.5</td>
<td>372.3</td>
<td>4</td>
</tr>
<tr>
<td>Liquidation des subventions ordinaires 2004</td>
<td>15</td>
<td>376.6</td>
<td>4</td>
</tr>
<tr>
<td>Déclaration de succession</td>
<td>57</td>
<td>379</td>
<td>5</td>
</tr>
<tr>
<td>Tax shelter</td>
<td>36.5</td>
<td>390</td>
<td>5</td>
</tr>
</tbody>
</table>
Annotation by the experts

Second step: 18 experts from FWB

- 7 batches of 15 texts, each was seen by 2.5 judges in average
- Interannotator agreement:
  \[ \text{average } \alpha \text{ de Krippendorf on the batches } = 0.37 \]
- Difficult task: similar task in SemEval has \( \kappa = 0.398 \) [Specia et al., 2012]
- Level of a given text = rounded mean of the judgements

In the end, 115 texts annotated in 5 levels
Predictors

344 variables from [François and Fairon, 2012], 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;
Contribution of cognitive studies on the reading process

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

- **lexical**: orthographic neighbours; normalized TTR.
- **syntactic**: verbal moods and tenses;
## Feature analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable description</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMP</td>
<td>Mean number of words per sentence</td>
<td>0.64</td>
</tr>
<tr>
<td>CON_PRO</td>
<td>nb. of conjunctions on the nb. of pronouns</td>
<td>0.54</td>
</tr>
<tr>
<td>Mean_freqCumNeigh</td>
<td>Mean of the cumulative frequencies of the neighbors</td>
<td>0.50</td>
</tr>
<tr>
<td>MedianFFFDV</td>
<td>Median of the verb frequencies</td>
<td>−0.47</td>
</tr>
<tr>
<td>Ppasse_C</td>
<td>Proportion past participles in the text</td>
<td>0.46</td>
</tr>
<tr>
<td>PAGoug_8000</td>
<td>Proportion of absent words from Gougenheim (8000)</td>
<td>0.44</td>
</tr>
<tr>
<td>PP1P2</td>
<td>Number of S1 and S2 personal pronouns</td>
<td>−0.42</td>
</tr>
<tr>
<td>PM8</td>
<td>Proportion words longer than 8 letters</td>
<td>0.40</td>
</tr>
<tr>
<td>ML3</td>
<td>Smoothed unigram model of inflected forms</td>
<td>−0.32</td>
</tr>
<tr>
<td>TTR_W</td>
<td>Type-Token ratio computed on lemmas</td>
<td>−0.21</td>
</tr>
</tbody>
</table>

Best feature is NMP (classic variable)!
Training the model

- Selection of features on 2 criteria:
  - Best features based on the correlation analysis
  - Best feature within its subfamily (e.g. language model, TTR, etc.)
- Statistical algorithm is SVM [Boser et al., 1992] with linear kernel and L2 norm
- Performance estimation (10-fold CV):
  - Accuracy = 58%
  - Adjacent accuracy = 91%
- [François and Fairon, 2012]: Accuracy = 50% and Adjacent accuracy = 80%
Conclusion

- A specialized formula for administrative texts in French
  → not the first specialized formula, but...
- Mixed annotation based on crowdsourcing and expert judgements
  → reading times seems more reliable
- Good performance with a small amount of texts, but... (level 1 and 5 !)
- The formula is available on the web
  (http://cental.uclouvain.be/amesure/)

For this context, just a formula do not seem useful enough
Perspectives

- Ask the experts to annotate the 10 texts with reading time
- Use only reading time to annotate texts (in a crowdsourcing setting)
- Compare reading times measured with AMesure-testing with eye-tracking data
- Provide a more precise diagnosis to writers from the administration!
Plan

1. Brief introduction of readability
2. Some issues with readability models
3. How to get annotated data?
4. AMesure: a readability model for administrative texts
5. AMesure: towards a readability platform
A readability platform

- In educative contexts, readability usually aims to gather textual resources for teaching or self-practice (REAP, DMesure, Choosito!, etc.)
- For administrative texts, the goal is to optimize the transmission of information
  → global diagnosis on text difficulty is less crucial

Local diagnosis appears more important (very few work on this)!
Content of the platform

- Global estimation of text difficulty via the above-mentioned model
- Global readability indicators on a specific textual dimension:
  - Number of difficult words (PAGoug_8000)
  - Mean length of sentences (NMP)
  - Syntactic complexity (CON_PRO)
  - Personalisation rate (PP1P2)
  - Intersentential cohesion based on LSA space
- Local difficulties: rare words and syntactic structures
The AMesure platform

Difficulté du texte : (explications)

<table>
<thead>
<tr>
<th>Difficulté globale</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulté du lexique</td>
<td>4</td>
</tr>
<tr>
<td>Longueur des phrases</td>
<td>2</td>
</tr>
<tr>
<td>Complexité syntaxique</td>
<td>5</td>
</tr>
<tr>
<td>Niveau de personnalisation</td>
<td>5</td>
</tr>
<tr>
<td>Cohérence du texte</td>
<td>1</td>
</tr>
</tbody>
</table>

Analyse détaillée du texte :

Devenir animateur de centres de vacances.

Un centre de vacances est un lieu d'accueil et d'animation pour enfants et jeunes de 2,5 à 15 ans, organisé pendant les périodes de vacances scolaires.

Le brevet d'animateur de centres de vacances s'obtient au terme d'une formation de 300 heures (150 heures de formation théorique et 150 heures de formation pratique) dispensée par un organisme habilité par la Communauté française. Le brevet d'animateur est un document officiel, il est homologué par la Communauté française.

La plupart des organismes de formation sont des organisations de jeunesse reconnues par la Communauté française.
How to define local complexity

**Lexical complexity**
- Based on lexical frequencies from Lexique 3 [New et al., 2004]
- We used a fixed threshold, but a slider might be preferred

**Syntactic complexity**
- Based on a typology of simplifications from [Brouwers et al., 2014]
- Typology was obtained from the manual analysis of a corpus of parallel sentences (original and simplified versions)
  → Sentences from Wikipedia and Vikidia
- Implementation of 19 rules from the typology within a simplification system (ATS)
  → parsing, detection of structures with Tregex and reordering with Tsurgeon (not performed here)
- Currently, detects passive, subordinate clauses and parenthesis.
Thanks!

**Original**: I would like to express our warmest thanks to the sincere attention you showed during my presentation. I urge you to ask questions if you have some.

**Simplified**: Thanks! Questions are welcome.


References II

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

  Syntactic sentence simplification for french.
  In *Proceedings of the 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.

  The Flesch Formula ’Through the Looking Glass’.
  *College English, 13*(7):393–394.

  A language modeling approach to predicting reading difficulty.
  In *Proceedings of HLT/NAACL 2004*, pages 193–200, Boston, USA.


*Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.*
PhD thesis, Université Catholique de Louvain.
Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

AMesure : une formule de lisibilité pour les textes administratifs.
In *Actes de la 21e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2014).*

An ”AI readability” formula for French as a foreign language.

Do NLP and machine learning improve traditional readability formulas ?
In *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012).*


