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

Thomas François[1]

(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].

Classic formulas

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

$$\text{Reading Ease} = 206,835 - 0,846\, wl - 1,015\, sl$$

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)

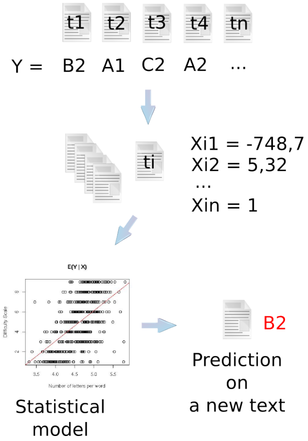
wl : number of syllables per 100 words

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.

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



t1 t2 t3 t4 tn

Y = B2 A1 C2 A2 …

ti

Xi1 = -748,7
Xi2 = 5,32
…
Xin = 1

Statistical model

B2

Prediction on a new text

Some trends in the field

Readability is mostly a 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 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 depends 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 !

Étude	$\sharp$ cl.	lg.	Acc.	Adj. Acc.	R	RMSE
[Collins-Thompson and Callan, 2004]	12	E.	/	/	0.79	/
[Heilman et al., 2008]	12	E.	/	52%	0.77	2.24
[Pitler and Nenkova, 2008]	5	E.	/	/	0.78	/
[Feng et al., 2010]	4	E.	70%	/	/	/
[Kate et al., 2010]	5	E.	/	/	0.82	/
[François, 2011]	6	F. (L2)	49%	80%	0.73	1.23
[François, 2011]	9	F. (L2)	35%	65%	0.74	1.92
[Vajjala and Meurers, 2012]	5	E.	93.3%	/	/	0.15

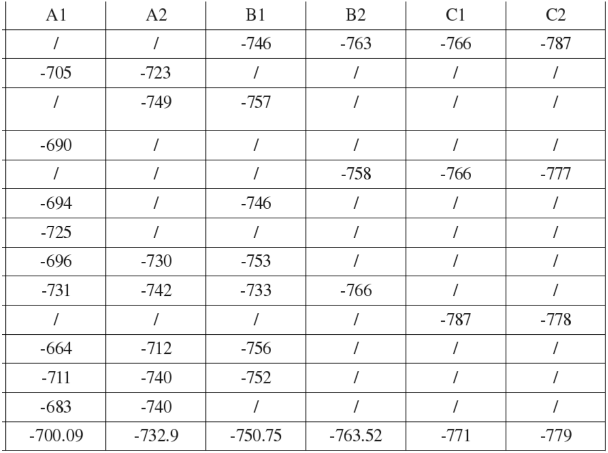
- Comparison between various models in [Nelson et al., 2012] :
  - Best model from [Nelson et al., 2012] is SourceRater [Sheehan et al., 2010]
    $\longrightarrow \rho = 0.860$ on Gates-MacGinite corpus
  - REAP achieve lower scores than classic models, such as DRP or Lexile.

The corpus issue

- Very few corpora available : Weekly Reader is mostly used [Schwarm and Ostendorf, 2005, Feng et al., 2010, Vajjala and Meurers, 2012]
  $\longrightarrow$ 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



A1	A2	B1	B2	C1	C2
/	/	-746	-763	-766	-787
-705	-723	/	/	/	/
/	-749	-757	/	/	/
-690	/	/	/	/	/
/	/	/	-758	-766	-777
-694	/	-746	/	/	/
-725	/	/	/	/	/
-696	-730	-753	/	/	/
-731	-742	-733	-766	/	/
/	/	/	/	-787	-778
-664	-712	-756	/	/	/
-711	-740	-752	/	/	/
-683	-740	/	/	/	/
-700.09	-732.9	-750.75	-763.52	-771	-779

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

Variable	Corr. and p-value	Variable	Corr. and p-value
35.PRON	-0.59 ($p = 0.005$)	3.Pers.Pro./S	$-0.41(p = 0.07)$
33.Indef NP	$-0.50(p = 0.02)$	10.Names/W	$-0.4(p = 0.08)$
18.S $\rightarrow$ O	$0.46(p = 0.04)$	9. nb. def. art./W	$0.38(p = 0.1)$
22. O $\rightarrow$ O	$-0.44(p = 0.048)$	17. S $\rightarrow$ S	$-0.36(p = 0.12)$

Classical features vs. NLP-based features

Contrasted results

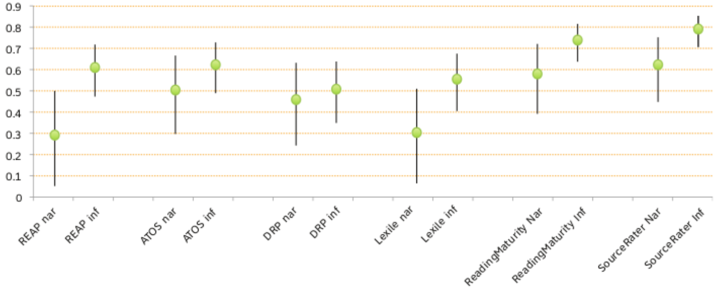
- Several “AI readability” models were reported to outperform classic formulas.
- [Aluisio et al., 2010, François, 2011] : best correlate is a classic feature (av. W/S ; % of W not in a list)
- [François et al., 2014] : best correlate is mean number of words per sentence...

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 $\rightarrow$ ”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

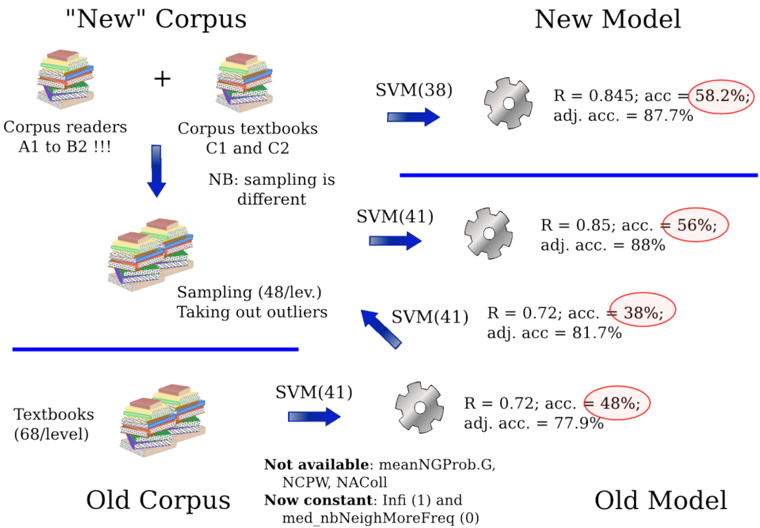
29 simplified readers collected :

	A1	A2	B1	B2
nb. of books	8	9	7	5
nb. of words	41018	71563	73011	59051

We divided the books by chapters and obtained the following training data :

	A1	A2	B1	B2
nb. of obs.	71	114	84	48
nb. of words	41018	71528	73007	59051

Typological experiment



"New" Corpus

Corpus readers A1 to B2 !!!

+

Corpus textbooks C1 and C2

NB: sampling is different

Sampling (48/lev.)
Taking out outliers

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%

Textbooks (68/level)

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 Corpus

Old Model

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

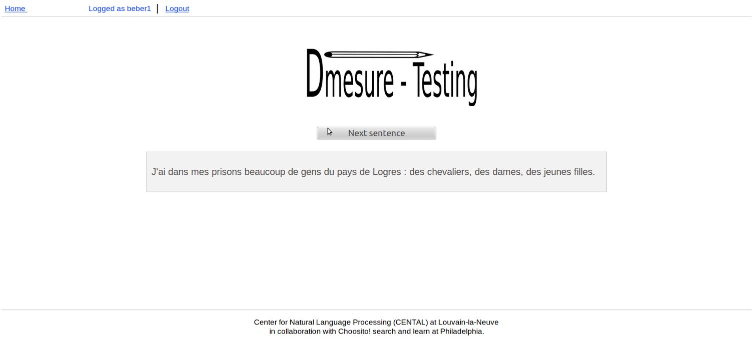
Reading time is used very little and yet might be the most psychologically reliable criterion.

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)

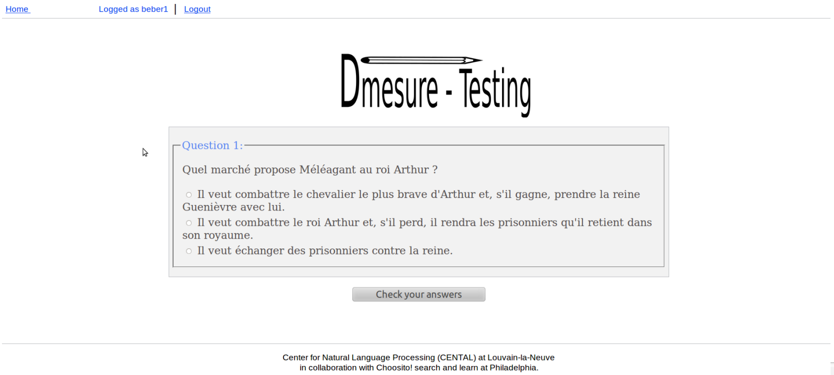


J'ai dans mes prisons beaucoup de gens du pays de Logres : des chevaliers, des dames, des jeunes filles.

Center for Natural Language Processing (CENTAL) at Louvain-la-Neuve
in collaboration with Choosito! search and learn at Philadelphia.

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.

Center for Natural Language Processing (CENTAL) at Louvain-la-Neuve
in collaboration with Choosito! search and learn at Philadelphia.

Results

Linger			
	Min-Max RT/W	nb. subjects	Corr.
Beginners II	717ms – 78680ms	9	$0,33$
Intermediate I	747ms – 69250ms	4	$0,32$

DMesure-Testing			
	Min-Max RT/W	nb. subjects	Corr.
Beginners II	562ms – 45351ms	9	$0,07$
Intermediate I	1296ms – 61770ms	4	$0,29$
Natives	493ms – 33050ms	6	$0,579$

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

Context

- 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

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

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 :
  $\longrightarrow$ average α 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

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

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

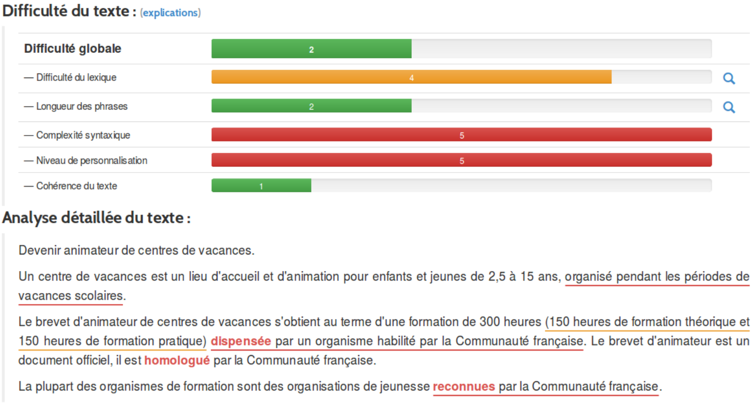
  ⟶ global diagnosis on text difficulty is less crucial

Local diagnosis appears more important (very few work on this) !

Content of the platform

- Global estimation of text difficulty via the above-mentionned 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)

Difficulté globale	2
— Difficulté du lexique	4
— Longueur des phrases	2
— Complexité syntaxique	5
— Niveau de personnalisation	5
— Cohérence du texte	1

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)
  $\longrightarrow$ Sentences from Wikipedia and Vikidia
- Implementation of 19 rules from the typology within a simplification system (ATS)
  $\longrightarrow$ 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 I

- Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010).
  Readability assessment for text simplification.
  In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles.

- Baayen, R. H., Davidson, D. J., and Bates, D. (2008).
  Mixed-effects modeling with crossed random effects for subjects and items.
  Journal of memory and language, 59(4) :390–412.

- Bormuth, J. (1966).
  Readability : A new approach.
  Reading research quarterly, 1(3) :79–132.

- Bormuth, J. (1969).
  Development of Readability Analysis.
  Technical report, Projet n°7-0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

References II

Boser, B., Guyon, I., and Vapnik, V. (1992).
A training algorithm for optimal margin classifiers.
In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2014).
Syntactic sentence simplification for french.
In *Proceedings of the 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.

Brown, J. (1952).
The Flesch Formula 'Through the Looking Glass'.
College English, 13(7) :393–394.

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.

References III

Collins-Thompson, K. and Callan, J. (2005).
Predicting reading difficulty with statistical language models.
Journal of the American Society for Information Science and Technology, 56(13) :1448–1462.

Dale, E. and Chall, J. (1948).
A formula for predicting readability.
Educational research bulletin, 27(1) :11–28.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010).
A Comparison of Features for Automatic Readability Assessment.
In *COLING 2010 : Poster Volume*, pages 276–284.

Flesch, R. (1948).
A new readability yardstick.
Journal of Applied Psychology, 32(3) :221–233.

Foltz, P., Kintsch, W., and Landauer, T. (1998).
The measurement of textual coherence with latent semantic analysis.
Discourse processes, 25(2) :285–307.

References IV

François, T. (2011).
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.

François, T., Brouwers, L., Naets, H., and Fairon, C. (2014).
AMesure : une formule de lisibilité pour les textes administratifs.
In *Actes de la 21e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2014).*

François, T. and Fairon, C. (2012).
An "AI readability" formula for French as a foreign language.
In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 466–477.

François, T. and Miltsakaki, E. (2012).
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).*

References V

François, T. and Watrin, P. (2011).
On the contribution of MWE-based features to a readability formula for French as a foreign language.
In *Proceedings of the International Conference RANLP 2011*.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008).
An analysis of statistical models and features for reading difficulty prediction.
In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–8.

Just, M. and Carpenter, P. (1980).
A theory of reading : From eye fixations to comprehension.
Psychological review, 87(4) :329–354.

Kandel, L. and Moles, A. (1958).
Application de l’indice de Flesch à la langue française.
Cahiers Études de Radio-Télévision, 19 :253–274.

References VI

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010).
Learning to predict readability using diverse linguistic features.
In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Kemper, S. (1983).
Measuring the inference load of a text.
Journal of Educational Psychology, 75(3) :391–401.

Kintsch, W. and Vipond, D. (1979).
Reading comprehension and readability in educational practice and psychological theory.
In Nilsson, L., editor, *Perspectives on Memory Research*, pages 329–365.
Lawrence Erlbaum, Hillsdale, NJ.

Krippendorff, K. (1980).
Content analysis : An introduction to its methodology.
Sage, Beverly Hills, CA.

References VII

Lively, B. and Pressey, S. (1923).
A method for measuring the “vocabulary burden” of textbooks.
Educational Administration and Supervision, 9 :389–398.

Nelson, J., Perfetti, C., Liben, D., and Liben, M. (2012).
Measures of text difficulty : Testing their predictive value for grade levels and student performance.
Student Achievement Partners.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004).
Lexique 2 : A new French lexical database.
Behavior Research Methods, Instruments, & Computers, 36(3) :516.

Pitler, E. and Nenkova, A. (2008).
Revisiting readability : A unified framework for predicting text quality.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

References VIII

Schwarm, S. and Ostendorf, M. (2005).
Reading level assessment using support vector machines and statistical language models.
Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 523–530.

Sheehan, K. M., Kostin, I., Futagi, Y., and Flor, M. (2010).
Generating automated text complexity classifications that are aligned with targeted text complexity standards.
Technical report, Educational Testing Service, RR-10-28.

Si, L. and Callan, J. (2001).
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.

Smith, E. and Senter, R. (1967).
Automated Readability Index.
Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

References IX

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012).
Semeval-2012 task 1 : English lexical simplification.
In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355.

Stevens, K. (1980).
Readability formulae and McCall-Crabbs standard test lessons in reading.
The Reading Teacher, 33(4) :413–415.

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L., and Bernhard, D. (2013).
Coherence and cohesion for the assessment of text readability.
Natural Language Processing and Cognitive Science, pages 11–19.

Vajjala, S. and Meurers, D. (2012).
On improving the accuracy of readability classification using insights from second language acquisition.
In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173.

References X

van Oosten, P. and Hoste, V. (2011).
Readability Annotation : Replacing the Expert by the Crowd.
In *Sixth Workshop on Innovative Use of NLP for Building Educational Applications*.

van Oosten, P., Hoste, V., and Tanghe, D. (2011).
A posteriori agreement as a quality measure for readability prediction systems.
In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 424–435. Springer, Berlin / Heidelberg.

Vogel, M. and Washburne, C. (1928).
An objective method of determining grade placement of children's reading material.
The Elementary School Journal, 28(5) :373–381.