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

Plan

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.

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 !

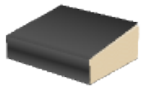
Let’s focus a bit more on this last point !

The corpus issue

Classic approach

Sample of readers

Corpus



Reading comp. tests
Cloze test
Reading speed

Scores per text and per reader

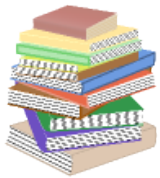
label = mean scores per text

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 (eg. 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.

  $\rightarrow$ 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...
  $\rightarrow$ Correlation between L1 and L2 scores was high ($r = 0.915$)
  $\rightarrow$ 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

Plan

Readability formulas for FFL ?

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 :

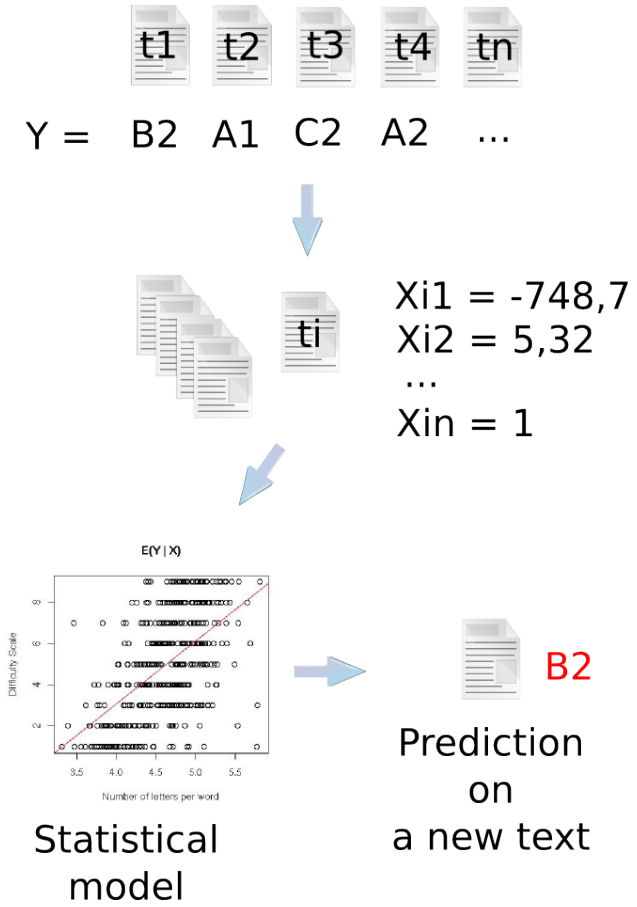
$$\mathrm{FR} = 10 * \mathit{WpS} - \mathit{Cog}$$

WpS : mean number of words per sentence.
Cog : number of cognates per 100 words.

Plan

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 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).

Corpus (2)

Not all FFL textbooks were used :

1. Have to follow the CEFR recommandations (posterior to 2001).
2. Language should be modern (arises from condition 1).
3. Intended audience : young people and adults (not children).
4. General reading : I excluded FSP textbooks.

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

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

TABLE: Number of texts per level, for each textbook series used.

Predictors from the literature

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 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.

Methodology

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]

- **S**upport **V**ector **M**achines [Boser et al., 1992]

Plan

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

	Test6CE			
	r	ρ	$W(p)$	$F(p)$
X75FFFDC	-0.296^{2}	-0.627^{3}	$< 0,001$	0.089
X90FFFC	-0.319^{3}	-0.641^{3}	$< 0,001$	$< 0,001$
PAGoug_2000	0.593^{3}	0.597^{3}	$< 0,001$	0.017
PA_Alterego1a	0.657^{3}	0.652^{3}	$< 0,001$	$< 0,001$
ML3	-0.56^{3}	-0.546^{3}	$< 0,001$	$< 0,001$
meanNGProb.G	0.382^{3}	0.407^{3}	0.011	0.05
NLM	0.479^{3}	0.483^{3}	0.028	0.084
NL90P	0.519^{3}	0.521^{3}	$< 0,001$	0.022
NMP	0.486^{3}	0.618^{3}	$< 0,001$	0.014
PRO.PRE	-0.181^{3}	-0.345^{3}	$< 0,001$	0.226
PPres	0.44^{3}	0.44^{3}	$< 0,001$	0.003
Pres_C	-0.355^{3}	-0.337^{3}	$< 0,001$	$< 0,001$
PP1P2	-0.408^{3}	-0.333^{3}	$< 0,001$	0.008
avLocalLsa_Lem	$0,63^{3}$	$0,63^{3}$	$< 0,001$	0,01
NAColl	/	0.286^{3}	/	/
BINGUI	$0,462^{3}$	$0,462^{3}$	$< 0,001$	0,018

Main results from the bivariate analysis

- Each familly has at least one efficient predictor
  $\rightarrow$ 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

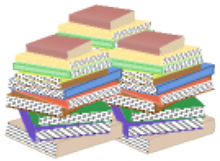
For the modelisation step, various combinations of predictors were attempted :

- Baseline (mimics classic formulas) : NMP + NLM.
- Best predictor/familly (4) : PA_Alterego1a + NMP + avLocalLsa_Lem + BINGUI.
- 2 best predictors/familly (8) : PA_Alterego1a + X90FFFC + NMP + PPres + avLocalLsa_Lem + PP1P2 + BINGUI + NAColl.

  $\rightarrow$ Assumption : maximizing the **type** of information in a minimal set.
- Automatic selection of features.
  $\rightarrow$ 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

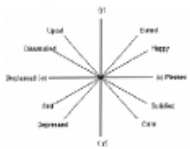
108 texts/level

Development set

40 texts/level

outliers removed

LSA space definition



Meta-parameters :
(grid-search)

C, γ, cp, minsplit

All estimates obtained with
10-fold cross-validation

Test set

68 texts/level

outliers removed

Modelling

Ten-fold
cross-validation

Results: estimation

Evaluation measures

Models were evaluated with these 5 measures :

- Multiple correlation ratio (R).
- Accuracy (acc).
- Adjacent accuracy ($acc - cont$)
  $\rightarrow$ 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

Model	Classifier	Parameters	R	acc	$acc - cont$	$rmse$	mae
Random	/	/	/	16, 6%	44, 4%	/	/
Baseline	SVM	$\gamma = 0,05$; $C = 25$	0, 62	34%	68, 2%	1, 51	1, 06
Model 2009	RLM	/	0, 62	41%	71%	/	/
Expert1	RLM	/	0, 70	39%	74, 2%	1, 34	0, 97
Expert2	SVM	$\gamma = 0,002$; $C = 75$	0, 73	41%	78%	1, 28	0, 94
Auto-OLR	OLR	/	0.71	39.6	76.1	1.33	0.96
Auto	SVM	$\gamma = 0,004$; $C = 5$	0, 73	49%	79, 6%	1, 27	0, 90

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%)

	A1	A2	B1	B2	C1	C2
Adj. acc.	100%	71%	67%	71%	86%	83%

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 :

	Family only		**All except family**	
	Acc.	Adj. acc.	Acc.	Adj. acc.
Lexical	40.5	75.6	41.1	73.5
Syntactic	39.3	69.5	43.2	78.4
Semantic	28.8	61.5	47.8	79.2
FFL	24.9	58.5	47.8	79.6

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.).

Comparison with other studies

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

[Kandel and Moles, 1958] is a general formula for L1 French
$\rightarrow$ 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 ?
  $\rightarrow$ 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 ?

Plan

Another corpus

We gathered manually another FFL corpus : simplified readers

- They are mainly narrative texts (a few are informative)
- No biais 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 :

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

Features analysis

We first ran a bivariate analysis at the readers level :

- Previous best feature **PA_Alterego1b** : $r = 0.280$ vs. 0.657
  $\rightarrow$ Probably due to a **change of topic** (eg. knights)
- Previously interesting BINGUI : *NA* vs. $0,462$
  $\rightarrow$ All the texts contains dialogues, since there are narrative.
- Effect of text length :
  $\rightarrow$ **normTTR_W** : $r = 0.587$ vs. $r = 0.125$
  $\rightarrow$ 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

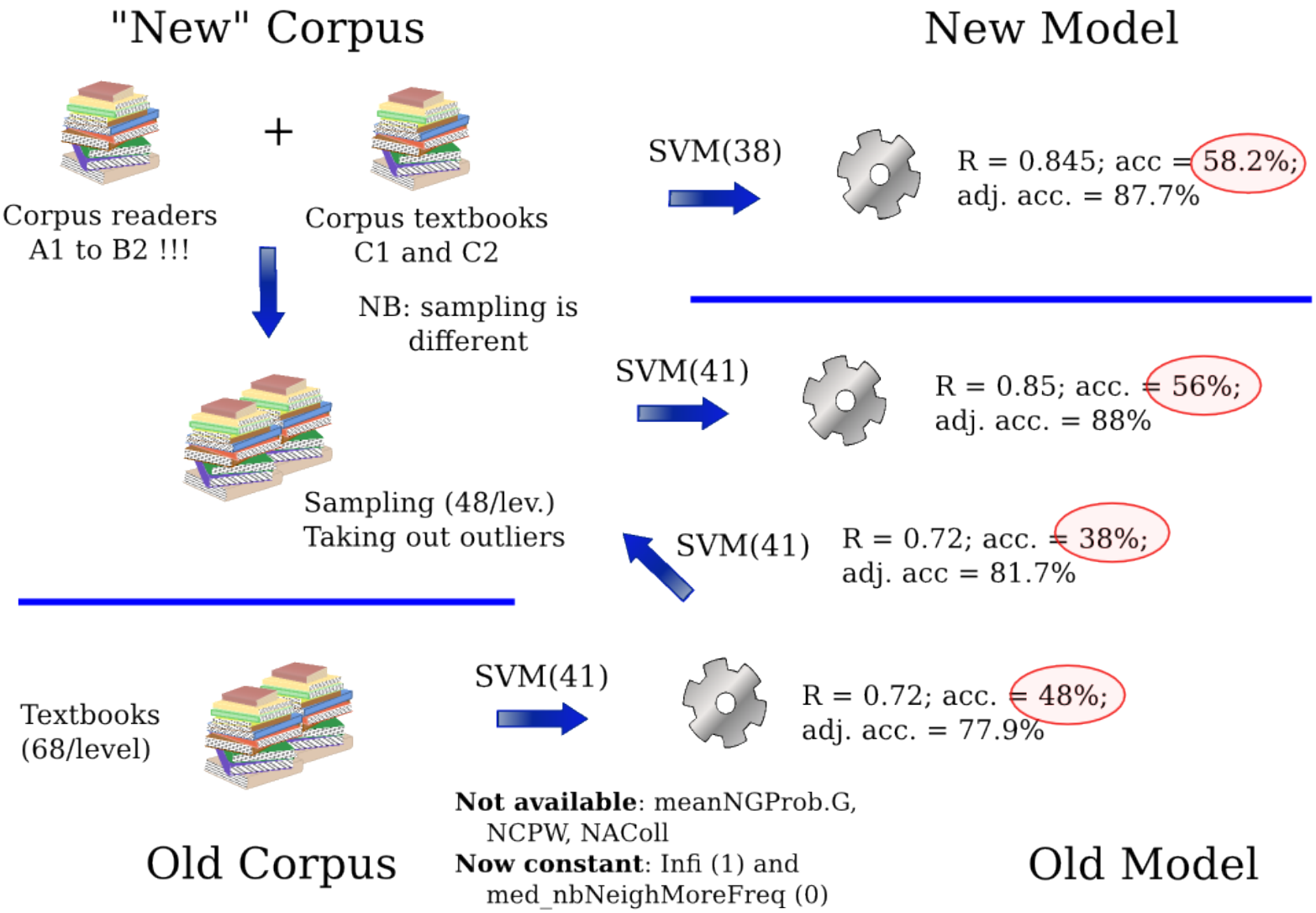
29 readers are not enough to train a readability model !

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

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

Correlations decrease but remain mostly coherent with previous figure.

Various experiments



Conclusions

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

- Loss of accuracy between both models **for the same population** !
  $\rightarrow$ 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
  $\rightarrow$ Some features are even constant !
- Obvious benefit of specializing the model :
  $\rightarrow$ Just retraining the same model on the new corpus : $+8\%$ (better coefficients)
  $\rightarrow$ Retraining + features selection : $+10\%$
- This also suggests that best path for improvement of readability models might be related to the training corpus.
  $\rightarrow$ due to higher homogeneity or just specialization ? ? ?

Perspectives : how to get specialized data ?

- 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.

End

Thank you for your attention.

Questions and comments are welcomed

Plan

References I


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.


Breiman, L. (1996).
Bagging predictors.
Machine learning, 24(2) :123–140.


Breiman, L., Friedman, H., Olsen, R., and Stone, J. (1984).
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.

References II


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.


Conseil de l'Europe (2001).
Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.
Hatier, Paris.


Cornaire, C. (1988).
La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère.
Canadian Modern Language Review, 44(2) :261–273.


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

References III

Feng, L., Elhadad, N., and Huenerfauth, M. (2009).
Cognitively motivated features for readability assessment.
In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237.

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.

Forbes, F. and Cottle, W. (1953).
A new method for determining readability of standardized tests.
Journal of Applied Psychology, 37(3) :185–190.

References IV


François, T. (2009a).
Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL.
In *Proceedings of the 12th Conference of the EACL : Student Research Workshop*, pages 19–27.


François, T. (2009b).
Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE.
In *11eme Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*.


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


Freund, Y. and Schapire, R. (1996).
Experiments with a new boosting algorithm.
In *Machine Learning : Proceedings of the Thirteenth International Conference*, pages 148–156.


Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007).
Combining lexical and grammatical features to improve readability measures for first and second language texts.
In *Proceedings of NAACL HLT*, pages 460–467.


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.


Henry, G. (1975).
Comment mesurer la lisibilité.
Labor, Bruxelles.

References VI


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


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.


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.


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.

References VII


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.


Spache, G. (1953).
A new readability formula for primary-grade reading materials.
The Elementary School Journal, 53(7) :410–413.


Tharp, J. (1939).
The Measurement of Vocabulary Difficulty.
Modern Language Journal, pages 169–178.


Uitdenbogerd, S. (2005).
Readability of French as a foreign language and its uses.
In *Proceedings of the Australian Document Computing Symposium*, pages 19–25.