A readability formula for French as a foreign language

Thomas François

B.A.E.F and Fulbright Fellow
CENTAL, Université Catholique de Louvain

CLUNCH, February 09, 2012

Plan

Plan

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.
- Calibration of books for children [Kibby, 1981, Stenner, 1996].
- Used in scientific experiments to control the difficulty of textual input data.
- Controling 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].

Two kinds of applications

Automated design of exercises based on a corpus

- French : **ALEXIA** [Chanier and Selva, 2000] ;
  ALFALEX [Selva, 2002, Verlinde et al., 2003] ;
  MIRTO [Antoniadis and Ponton, 2004, Antoniadis et al., 2005].
- English : **Cloze tests** [Coniam, 1997, Brown et al., 2005] ;
  WERTi [Amaral et al., 2006] ; **VISL** [Bick, 2001]

Web crawlers for the automatic retrieval of web texts on a specific topic and at a specific readability level

- French : **Dmesure** [François and Naets, 2011] (prototype)
- English : **READ-X** [Miltsakaki and Troutt, 2008], **IR4LL** [Ott, 2009] ; **REAP** [Heilman et al., 2008b]

Readability formulas seem to offer various interesting perspectives in iCALL.

What about readability formulas for FFL ?

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

$\rightarrow$ 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 :

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

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 :
  $\rightarrow$ grammatical features play a more important role in a L2 model.

Objectives of this work

First objective

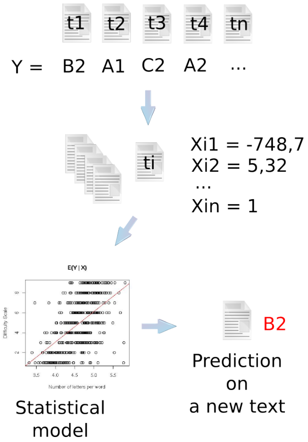
- Design a readability formula (or model) for FFL that may account for the specificities of this context.
- This amounts to three subgoals :
  - Use a corpus assessed for a L2 population to tune the weights for each predictor.
  - Adapt some well-known predictors to better fit the L2 context.
  - Find some predictors that correspond to some specific features of the L2 reading process.

Plan

1. Introduction : readability for FFL
2. Methodology
   - Linguistic predictors of difficulty
   - The corpus
   - The statistical algorithms
3. Results
   - Bivariate analysis
   - Design of the readability model
4. Discussion and conclusions
5. References

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



Plan

Ways for finding good predictors

I considered two distinct lines of research :

1. Adapt features from the L1 French and English literature ;
2. Explore the process of reading in L1 and L2 to discover new features that affects it.

Main types of predictors in readability

4 major periods in readability :

1. **Classic period** : formulas are based on linear regression and mostly use two **indices** (one lexical, one syntactic) [Flesch, 1948, Dale and Chall, 1948]
2. **The cloze test era** : concerns arise about motivated features (= cause of difficulty) [Bormuth, 1969]
3. **Structuro-cognitivist period** : expressed criticism towards the classical formulae, unable to take into consideration some organisational (coherence, cohesion) or cognitive aspects (conceptual density, inference load, etc.)

   [Kintsch and Vipond, 1979, Kemper, 1983]

Main types of predictors in readability (2)

4. **Recent studies** : I gathered them under the term *IA readability*
   → They make use of NLP and machine learning techniques.
   - First IA studies : coherence level as a predictor (estimated through LSA) [Foltz et al., 1998] and the first language model-based approach [Si and Callan, 2001].
   - 2004-2007 : application of NLP techniques to lexical et syntactic levels [Collins-Thompson and Callan, 2005, Schwarm and Ostendorf, 2005, Heilman et al., 2007].
   - After 2007 : Semantic, discourse and cognitive variables are considered [Crossley et al., 2007, Pitler and Nenkova, 2008, Feng et al., 2009].

In our view, *IA readability* aims to bury the hatchet between traditional and structuro-cognitivist paradigms.

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.

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.

Objectives of the work (2)

First objective

Design a readability formula (or model) for FFL that may account for the specificities of this context.

Second objective

Get a better understanding of the IA readability : why does it seem to work better than traditional formulas ?

Plan

The annotation criterion

- Gathering a labeled corpus requires to choose a criterion to assess the reading difficulty of texts.
  $\rightarrow$ After reviewing the literature, I selected **expert judgments**.
- The type of criterion affects the difficulty scale used.
  $\rightarrow$ We extracted 2042 texts from 28 FFL textbooks, following the CEFR scale [Conseil de l'Europe, 2001].

Our assumption is...

The level of a text can be considered the same as the level of the textbooks it comes from.

The CEFR scale

- It has 6 levels :
  A1 (easier), A2, B1, B2, C1, and C2 (higher)
- Some authors / teachers recommend to refine the scale by dividing certain levels :
  Then, I also used a 9-levels scale : A1 (easier), A1+, A2, A2+, B1, B1+, B2, C1, and C2 (higher)
- This division can better take into account differences in skills for learners of lower levels, where they are more pronounced than in the upper levels.

Criteria for text selection

First, 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.

Problems of this corpus :

Two problems were detected :

1. Low number of texts labeled as C2.
   $\rightarrow$ Preliminary experiments showed that it matters to have balanced classes.
2. Inconsistencies between the annotation from different experts (= textbook publishers).

Two solutions were investigated :

1. Long C2 texts were divided into 2 or 3 fragments $\rightarrow$ 108 texts.
2. I set aside textbooks whose annotations were the most inconsistent.

I thus compared 8 different corpora !

Plan

Statistical models used

- **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.
2. Evaluation of the predictive ability of some combinations on variables (= formulas).

Indeed, there are multicollinearity risks.

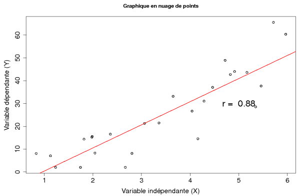
$\rightarrow$ Only 2 out of the 8 corpora were retained.

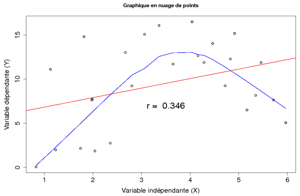
Plan

Evaluation measures

4 measures were calculated for every of the 406 variables, in order to assess their predictive power :

1. Pearson's r : useful for linear associations.
2. Spearman's ρ : useful for the monotonic increasing associations.
3. [Guilford, 1965]'s F test : assess whether the association is linear or not.
4. Shapiro-Wilk's W : assess the normality of the predictor.





Most interesting features

	Test6CE				Test9CE		
	r	ρ	$W(p)$	$F(p)$	r	ρ	$F(p)$
X75FFFDC	-0.296^{2}	-0.627^{3}	$< 0,001$	0.089	-0.367^{3}	-0.623^{3}	0.092
X90FFFC	-0.319^{3}	-0.641^{3}	$< 0,001$	$< 0,001$	-0.246^{3}	-0.628^{3}	$< 0,001$
PAGoug_2000	0.593^{3}	0.597^{3}	$< 0,001$	0.017	0.574^{3}	0.588^{3}	0.313
PA_Alterego1a	0.657^{3}	0.652^{3}	$< 0,001$	$< 0,001$	0.668^{3}	0.672^{3}	0.002
ML3	-0.56^{3}	-0.546^{3}	$< 0,001$	$< 0,001$	-0.556^{3}	-0.552^{3}	0.026
meanNGProb.G	0.382^{3}	0.407^{3}	0.011	0.05	-0.244^{3}	-0.104^{1}	0.417
NLM	0.479^{3}	0.483^{3}	0.028	0.084	0.431^{3}	0.44^{3}	0.027
NL90P	0.519^{3}	0.521^{3}	$< 0,001$	0.022	0.478^{3}	0.485^{3}	0.021
NMP	0.486^{3}	0.618^{3}	$< 0,001$	0.014	0.487^{3}	0.652^{3}	0.031
PRO.PRE	-0.181^{3}	-0.345^{3}	$< 0,001$	0.226	-0.194^{3}	-0.349^{3}	0.021
PPres	0.44^{3}	0.44^{3}	$< 0,001$	0.003	0.463^{3}	0.463^{3}	0.023
Pres_C	-0.355^{3}	-0.337^{3}	$< 0,001$	$< 0,001$	-0.439^{3}	-0.433^{3}	$< 0,001$
PP1P2	-0.408^{3}	-0.333^{3}	$< 0,001$	0.008	-0.405^{3}	-0.346^{3}	$< 0,001$
avLocalLsa_Lem	$0,63^{3}$	$0,63^{3}$	$< 0,001$	0,01	$0,57^{3}$	$0,57^{3}$	0,05
NAColl	/	0.286^{3}	/	/	/	0.253^{3}	/
BINGUI	$0,462^{3}$	$0,462^{3}$	$< 0,001$	0,018	$0,45^{3}$	$0,45^{3}$	0,311

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**.
- The efficiency of **PA_Alterego1a** provides a rationale for adapting readability models to specific contexts (list for FFL).
- Few variables are normally distributed and only part of them are lineary related to our criterion.

What about the contribution of NLP ?

- The LSA-based features is among the best (with ML3). This seems to confirm the value of NLP for readability...
- However, a lot of NLP variables are poor predictors : N-gram models (where N>1), MWE-based features, etc.

Plan

Comparison of several feature sets

In the second step, various combinations of predictors were attempted :

- Baseline (that 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.
- Automatic selection of features.
  $\rightarrow$ Assumption : maximizing the **quantity** of information.

Each set was tested with the 6 statistical algorithms, for our 2 scales (6 and 9 levels).

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., 2008a]
- Root mean square error (RMSE).
- Mean absolute error (MAE).

Main results

Model	Classifieur	Paramètres	R	acc	$acc - cont$	$rmse$	mae
Corpus with 6 classes							
Random	/	/	/	$16,6\%$	$44,4\%$	/	/
Baseline	SVM	$\gamma = 0,05$; $C = 25$	$0,62$	34%	$68,2\%$	$1,51$	$1,06$
Expert1	RLM	/	$0,70$	39%	$74,2\%$	$1,34$	$0,97$
Expert2	SVM	$\gamma = 0,002$; $C = 75$	$0,73$	41%	78%	$1,28$	$0,94$
Model 2009	RLM	/	$0,62$	41%	71%	/	/
Auto	SVM	$\gamma = 0,004$; $C = 5$	$0,73$	49%	$79,6\%$	$1,27$	$0,90$
Corpus with 9 classes							
Random	/	/	/	$11,1\%$	$30,8\%$	/	/
Baseline	SVM	$\gamma = 0,01$; $C = 40$	$0,68$	$26,5\%$	$54,5\%$	$2,27$	$1,29$
Expert1	RLM	/	$0,74$	$27,5\%$	$58,1\%$	$1,95$	$1,20$
Expert2	SVM	$\gamma = 0,006$; $C = 20$	$0,75$	31%	$62,3\%$	$1,90$	$1,17$
Model 2009	RLM	/	$0,72$	32%	63%	/	/
Auto	SVM	$\gamma = 0,004$; $C = 15$	$0,74$	35%	$65,4\%$	$1,92$	$1,15$

Best models

- $+32,4\%$ (6 classes) and $+23,9\%$ (9 classes) in comparison with random (acc) ;
- $+8\%$ (6) and $+3\%$ (9) in comparison with previous 2009 model (acc) ;

Comparison with other studies

Étude	♯ cl.	lg.	Acc.	Cont. Acc.	R	RMSE
[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., 2008a]	12	E.	/	45%	0, 58	2, 94
[Heilman et al., 2008a]	12	E.	/	52%	0, 77	2, 24
[Pitler and Nenkova, 2008]	5	E.	/	/	0, 78	/
[François, 2009]	6	F. (L2)	41%	71%	0, 62	/
[François, 2009]	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
9-classes model	9	F. (L2)	35%	65%	0, 74	1, 92

[Schwarm and Ostendorf, 2005] : gain from random for *acc* − *cont* is +24, 5% to +29%, while it is a mean of +36% for our model.

Plan

What about our 2 goals ?

1. Design a new readability formula better tuned for L2 (FFL) contexts.

- New readability formula using SVM and 46 variables, offering state-of-the-art performances.
- 1st FFL formula using NLP and machine learning techniques.
- What about the 3 levels of tuning the formula for L2 context ? :
  - I used a corpus assessed for L2 learners, but did not assess its specific contribution.
  - Adaptation of classic predictors to the L2 context appeared highly successful (PA_Alterego1a).
  - The new features specific to the L2 context were mostly poor predictors.

Perspectives at this level :

- Assess the contribution of tuning the corpus to performances.
- Expand the number of specific L2 predictors (e.g. influence of the L1).

What about our 2 goals ? (2)

2. Contributions of NLP and machine learning to readability

- Independently, several "NLP variables" appeared to be good predictors (LSA, unigram, POS ratio, etc.).
- However, when combined with classic features, their contribution drop (LSA is even not retained).
  → It appears that some variables (MWE-based) are suffering from errors and approximations inherent to NLP programms.
- Most of the gain from classic formulas might be due to a combinaison of better training algorithms, able to use efficiently more variables.

Perspectives at this level :

- Run experiments to clear out the contribution of the new features and the machine learning algorithms.
- Replicate them for another context (L1 English).

Additional assumption : multidimensionnality

Assumption = getting the best performance using different textual informations

- 4 dimensions (OLR) : acc : $36,8\%$ and $acc - cont$: $77,8\%$ vs. automatic selection of 4 var. (lexico-syntactics) : acc : 40% et $acc - cont$: $76,1\%$!
  $\rightarrow$ The assumption does not seem to stand !
- Moreover, LSA-based features sometimes suffers from multicollinearity with other lexico-syntactic variables
  $\rightarrow$ Are semantic and discourse features really bringing new information to lower level predictors in a L2 context ?

Perspectives at this level :

- Replicate this experimentation with other semantic and discourse features.
- Check if this result would stand for L1 formulas. L2 readers probably encounter more problems at lexico-syntactic levels than natives.

The end

Difficulté estimée :	A2
Votre texte :	Merci pour votre attention. Sachez que les questions et les commentaires sont les bienvenus :-)

Bibliography link :

→ https ://sites.google.com/site/readabilitybib/bibliography

Plan

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.

Amaral, L., Metcalf, V., and Meurers, D. (2006).
Language awareness through re-use of NLP technology.
In *Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO*, University of Hawaii.

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., and Ponton, C. (2005).
Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO.
Apprentissage des langues et systèmes d'information et de communication (ALSIC), 8(1) :65–79.

Antoniadis, G. and Grusson, Y. (1996).
Modélisation et génération automatique de la lisibilité de textes.
In *ILN 96 : Informatique et Langue Naturelle*.

References II

Antoniadis, G. and Ponton, C. (2004).
MIRTO : un système au service de l'enseignement des langues.
In *Proc. of UNTELE 2004*, Compiègne, France.

Bick, E. (2001).
The VISL system : research and applicative aspects of IT-based learning.
In *Proceedings of NoDaLiDa*, Uppsala.

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.

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.

References III

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.

Brown, J., Frishkoff, G., and Eskenazi, M. (2005).
Automatic question generation for vocabulary assessment.
In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Vancouver, Canada.

Chanier, T. and Selva, T. (2000).
Génération automatique d'activités lexicales dans le système ALEXIA.
Sciences et Techniques Educatives, 7(2) :385–412.

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 IV

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.

Coniam, D. (1997).
A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests.
Calico Journal, 14 :15–34.

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

Cornaire, C. (1985).
La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère.
PhD thesis, Université de Montréal, Montréal.

References V

Crossley, S., Dufty, D., McCarthy, P., and McNamara, D. (2007).
Toward a new readability : A mixed model approach.
In *Proceedings of the 29th annual conference of the Cognitive Science Society*, pages 197–202.

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

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.

References VI

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.

François, T. (2009).
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 Naets, H. (2011).
Dmesure : a readability platform for French as a foreign language.
In *Computational Linguistics in the Netherlands (CLIN21), University College Ghent, 11 February*.

References VII

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

Guilford, J. (1965).
Fundamental statistics in psychology and education.
McGraw-Hill, New-York.

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. (2008a).
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.

References VIII

Heilman, M., Zhao, L., Pino, J., and Eskenazi, M. (2008b).
Retrieval of reading materials for vocabulary and reading practice.
In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88.

Kanungo, T. and Orr, D. (2009).
Predicting the readability of short web summaries.
In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211.

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.

References IX

Kibby, M. (1981).
Test Review : The Degrees of Reading Power.
Journal of Reading, 24(5) :416–427.

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.

Koda, K. (2005).
Insights into second language reading : A cross-linguistic approach.
Cambridge University Press, Cambridge.

Miltsakaki, E. and Troutt, A. (2008).
Real-time web text classification and analysis of reading difficulty.
In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–97.

References X

Ott, N. (2009).
Information Retrieval for Language Learning : An Exploration of Text Difficulty Measures.
Master's thesis, University of Tübingen, Seminar für Sprachwissenschaft.
http ://drni.de/zap/ma-thesis.

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.

Selva, T. (2002).
Génération automatique d'exercices contextuels de vocabulaire.
In *Actes de TALN 2002*, pages 185–194.

References XI

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.

Stenner, A. (1996).
Measuring reading comprehension with the lexile framework.
In *Fourth North American Conference on Adolescent/Adult Literacy*.

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.

References XII

Verlinde, S., Selva, T., and Binon, J. (2003).
Alfalex : un environnement d'apprentisage du vocabulaire français en ligne, interactif et automatisé.
Romaneske, 28(1) :42–62.