

Dmesure : a readability formula for French as a foreign language



Thomas François, Aspirant F.N.R.S.



CENTAL, Université Catholique de Louvain

Presentation at the Interdisciplinary Research on Technology,
Education & Communication

October 20, 2010



Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
- 4 Evaluation
- 5 Perspectives for future research
- 6 References

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Context

- The sector of foreign language teaching is growing and changing :
 - There is a will to optimize the costs of education while improving its quality
 - The number of professionals is insufficient relative to demand.
 - Learners want more flexibility in teaching methods (timetable, place...).

ICALL has been viewed as a solution to these issues through the development of self-learning software.

Contribution of ICALL

- For teachers, ICTs provide access to authentic language and real communication situation :
 - Automatic retrieval of texts on the web
 - Allow interaction with people from other cultures
- ICTs also help relieve teachers of repetitive tasks :
 - Automated design of exercises aimed at the assimilation of specific linguistic forms (such as collocation, grammar notion...) through repetition.

Two kind 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

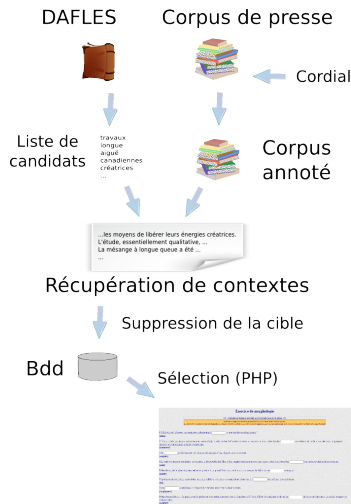
- French : ?
- English : **IR4LL** [Ott, 2009] ; **REAP** [Heilman et al., 2008b], **READ-X** [Miltsakaki and Troutt, 2008]

Generation of exercises : an example

● ALFALEX

[Selva, 2002, Verlinde et al., 2003]

- Automated design of exercises on morphology, gender, collocations...
- Difficulty of the task : 2 levels
- Difficulty of the context is not controlled !
It depends on the level of the corpus used.
- <http://www.kuleuven.be/alfalex/>



An example of this contextual complexity

Exercice de morphologie

*** Complétez les phrases en accordant les mots en italiques en fin de phrase. ***

La forme à compléter est nécessairement différente du mot donné en fin de phrase.

ATTENTION: le nombre de phrases disponibles est limité à 33. Si vous désirez faire des exercices supplémentaires sur la morphologie (avec d'autres exemples), voir FAQ sur la page d'accueil.

1 Il faut choisir la bonne, une musique instrumentale [] et non pas des airs *tapageurs*.
(*doux*)

2 * Autour de la petite poste rénovée sont venus s'adjoindre la mairie, l'office de tourisme, un *secrétariat mutualisé*, l'école [], un médecin et un dentiste, demain une pompe à essence, s'enthousiasme Brigitte Fargevielle.
(*maternel*)

3 Sa [] préfère parler de "l'ambiance incroyable" qui régnait dans le *cabaret*.
(*copain*)

4 La rude vie du petit *séminaire*, les copains, la découverte des filles et les longues discussions avec une jeune *novice* lui ouvrent les [] sur les incertitudes de sa *vocation*.
(*oeil*)

5 Mais le couple le plus attachant est celui qui réunit un grand Black bourré d'humour et une petite Hollandaise [] à croquer.
(*malin*)

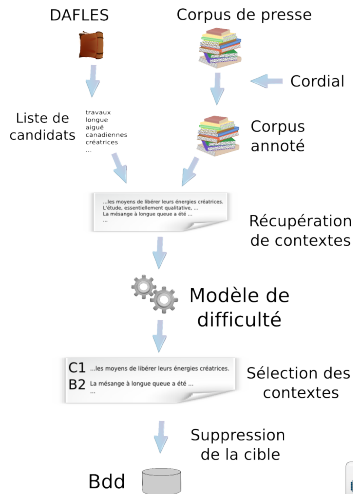
6 Opération de séduction, sans doute, mais qui reflète à l'évidence les aspirations d'une société [] de la *ferule des ayatollahs*.
(*las*)

7 Les [] australiens ont *disputé* la première rencontre de leur tournée.
(*rugbyman*)

8 Mais l'essentiel pour Singapour est de préserver son secteur des services qui représente 70 % du PIB et de continuer à attirer les [] et le savoir-faire dans un certain nombre de secteurs-clés.
(*capital*)

Readability model as a solution

- We can control two aspects :
 - Difficulty of the task : already taken into consideration (2 levels)
 - Contextual difficulty using a difficulty model (see figure)

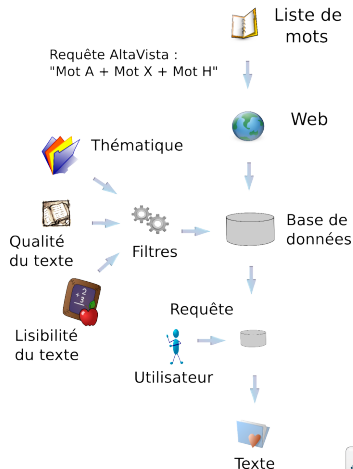


Retrieval of web texts : an example for EFL

● REAP

[Heilman et al., 2008b,
Collins-Thompson and Callan, 2004]

- REAding-specific Practice aims at improving reading comprehension abilities through practice.
- It integrates a SVM thematic classifier
- Difficulty is checked using the readability formulas described in [Collins-Thompson and Callan, 2005, Heilman et al., 2008a]
- <http://reap.cs.cmu.edu/>



Readability : an example

Grammar-based Reading Difficulty Prediction

Grade level predicted: 12.0

Accuracy generally improves with text length. The software will provide estimates for texts of any length, but a minimum length of 30 words is recommended. Also, the system is generally more accurate for grade levels above 2.

Type or paste your text into the box below and press "Submit" to obtain an estimate of the difficulty of your text.

A narrow grave-yard in the heart of a bustling, indifferent city, seen from the windows of a gloomy-looking inn, is at no time an object of enlivening suggestion; and the spectacle is not at its best when the mouldy tombstones and funereal umbrage have received the ineffectual refreshment of a dull, moist snow-fall. If, while the air is thickened by this frosty drizzle, the calendar should happen to indicate that the blessed vernal season is already six weeks old, it will be admitted that no depressing influence is absent from the scene.

Submit

An estimation of the readability of the first lines of *The Europeans* (H.James). It has been assessed by the model of [Heilman et al., 2007].

Url : <http://boston.lti.cs.cmu.edu/demos/readability/index.php>

Retrieval of web texts for FFL

Beyond search engines, there is no tool available for FFL.

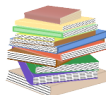
Situation actuelle

Professeur,...
cherche un texte
sur un sujet donné.

**Recherche
Internet**



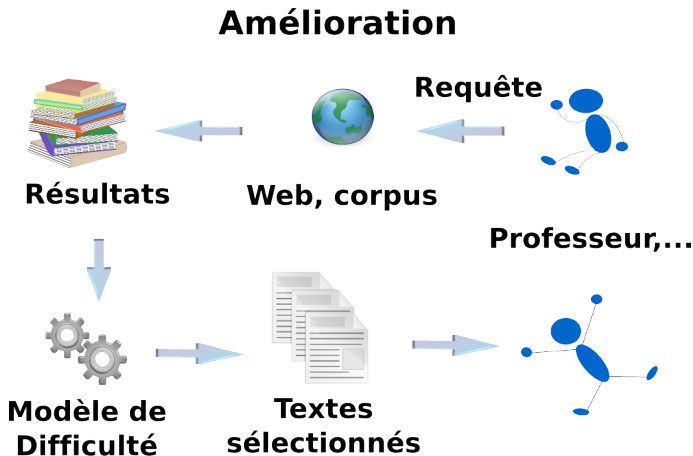
Résultats



Zzzz !



Why a difficulty model would be interesting



Difficulty model : a definition

Strictly speaking, a model for contextual difficulty in FFL :

- ① Amounts to assess the **reading** complexity of a text fragment for a FFL learner
 - What is reading in a foreign language ?
 - What good is it to read when learning a L2 ?
- ② It is an issue better known as readability.
 - What is a readability formula ?
 - What previous work exist ?
 - What should be the characteristics of a readability formula specific to L2 reading ?

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Definitions

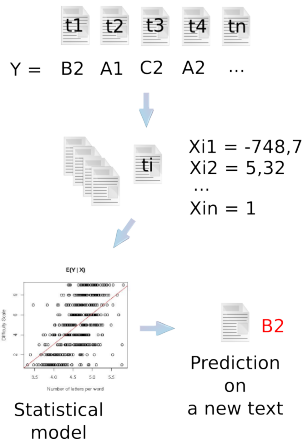
Readability can be defined as :

The sum total (including the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it, read it at a optimal speed, and find it interesting. [Dale and Chall, 1949, 1]

Par lisibilité, nous désignons le degré de difficulté éprouvé par un lecteur essayant de comprendre un texte. [Henry, 1975]

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



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, Bormuth, 1969]
- 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]

Situation for French

There are few studies about French L1 :

- Some formulae for English were used with small adjustments
[Kandel and Moles, 1958, de Landsheere, 1963]
- A few specific formulae for French L1 were coined
[Henry, 1975, Richaudeau, 1979, Mesnager, 1989]

There are even fewer about FFL :

- Previous work : [Cornaire, 1985, Uitdenbogerd, 2005]
- Our first model : [François, 2009a, François, 2009b]

Conclusion

There is indeed a real need for such a model

How to improve the existing

Using NLP and Data Mining techniques

- Provides a better coverage for the formula, since the corpus is bigger
- New statistical algorithm should better model the relations between difficulty and text characteristics
- NLP tools allow to model more complex variables, such as semantic ones, that could not be taken into consideration into previous automated formulae

Building on theoretical contributions from cognitive psychology

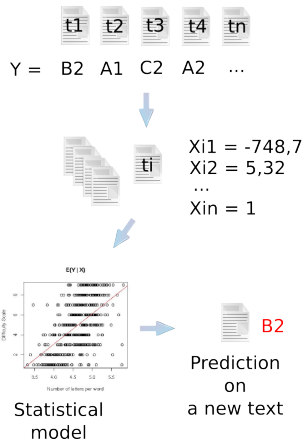
- Studies aiming at defining the reading process may help discovering new difficulty predictors
- Others focusing on the distinction between L1 and L2 reading help developing a model specific to L2 reading
- Reading studies make apparent the limitations of readability formulae

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 **Methodological steps**
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 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

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 **Methodological steps**
 - **The corpus**
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Collecting the corpus

Goal

Collect a corpus labelled by reading-difficulty levels, which implies :

- agreement on the difficulty scale, and
- the use of a criterion to assess the level of each text.

The classic approach

- **The criterion** : tests (comprehension or cloze tests...) are applied on a population.
- **The scale of measurement** : is a quantitative one : percentages corresponding to the average achievement tests. However, they are often transformed into a grade level scale.

My own approach

The needs

- The NLP perspective of readability requires a large number of texts
- The scale of measurement should have a practical interpretation

A solution

- Since 2001, the difficulty level of FFL textbooks has to be expressed in the “Common European Framework of Reference” scale.
- It is then possible to use FFL textbooks as a source of labelled texts.
- The difficulty-level of a text corresponds to the textbook level it comes from...
- This scale is immediately usable by FFL teachers

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, we 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 selecting texts

Obviously, it is not possible to use all textbooks as corpus, nor all texts

Criteria

- Textbooks should be posterior to 2001
- The public target should be adults and young people.
- Textbooks for general purposes (we excluded the FFL for specific purposes textbooks)
- Inside the selected textbooks, only texts associated with a task of reading comprehension were kept

First analyses : the corpus is noisy

Why ?

- In modern textbooks using the actional approach, it happens to meet a complicated text that is associated with a simpler task.
Eg. : RP LM = -731 (at level A1, where the mean = -700)
- On the opposite, it happens to meet a simple text that is associated with a complex task (eg. a song at the B2 level)
- In both case, it seems that our assumption (level of the text = level of the textbook) may not hold.

Some solutions

- Currently, outliers are suppressed.
- We are planning to check manually for such cases
- Creation of a less noisy corpus (using Dmesure)

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps**
 - The corpus
 - Difficulty factors**
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Difficulty predictors

- Lexical predictors (4) :
 - Language models (LM1, LM2, LM3)
 - Mean number of letters per word (NLM)
- Syntactical predictors (12) :
 - Mean number of words per sentence (NMP)
 - 11 binary variables : tense and mood
- Dialogue variables (5) :

They aim to determine whether the text is a dialog or not

Why a language model ?

Theoretical foundation

Various studies in psycholinguistics suggest an association between the difficulty of words and their frequency

[Howes and Solomon, 1951, Brysbaert et al., 2000]

The classic approach

They use a percentage of words absent from a list of the most frequent words of the language (Thorndike's, Dale's or Gougenheim's)

[Dale and Chall, 1948, Henry, 1975]

The language model

A smoothed unigram model may be used advantageously instead of these lists [Collins-Thompson and Callan, 2005]

LM : which is the best token unit ?

Which is the best unit for the unigram model in French ?

- lemma (LM1)
- inflected forms (LM2)
- inflected forms (LM3) disambiguated using TreeTagger [Schmid, 1994]

Results : The correlations with the “difficulty” are similar

Unit	LM1	LM2	LM3
Correlation (r)	-0,58	-0,58	-0,59

Syntactical predictors

- List of 11 binary variables selected :

Conditional	Future	Imperative
Imperfect	Infinitive	Past participle
Present participle	Present	Simple past
Subjunctive present	Subjunctive imperfect	

- Objective** : model the pace of teaching grammar in a context of FLE
- Optimal approach** : automatic recognition of grammatical structures
- Problem** : syntactic parsers for French are still lacking precision and slow. Therefore, we opted for this simplified parameterization.

Dialogue variables

Assumption

Dialogues are easier to read and understand than narratives, instructional or scientific texts.

⇒ lexicon and syntactical structures are simpler ; topic are closer to everyday life [Dolch, 1948, Flesch, 1948, Gunning, 1952, Henry, 1975]

Selected variables :

- Proportion of personal pronoun of dialogue (1P, 2P)
- Ratio of interjection on the number of words
- Ratio of "!" and "?" on the number of following punctuation marks (!? .)
- Ratio of "!" and "?" on the number of these punctuation marks (!? . ; :)
- Presence of quotation marks for dialogue

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 **Methodological steps**
 - The corpus
 - Difficulty factors
 - **The statistical models**
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Statistical models used

- **Regression models** : they depends on the type of the dependant variable
 - Continuous ⇒ Linear regression
 - Ordinal ⇒ Proportional odds model (PO)
 - Categorical ⇒ Multinomial logistic regression (MLR)
- Models based on **decision trees** :
 - Classification tree (baseline) [Breiman et al., 1984]
 - Boosting [Freund and Schapire, 1996]
 - Bagging [Breiman, 1996]
- **Support Vector Machines** [Boser et al., 1992]

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

Evaluation data

2 scales of measurement = 2 datasets :

- **6-levels model** : an i.i.d sample (Corp6) of 299 texts of the corpus
- **9-levels model** : an i.i.d sample (Corp9) of 449 texts of the corpus

Outliers

Here, an outlier is defined as an observation located more than three standard deviations away from the average of its class.

⇒ Corp6 : 11 outliers (remains 288 texts)

⇒ Corp9 : 12 outliers (remains 437 texts)

Two evaluation procedures

- **The features selection** : we used a stepwise selection based on the Akaike's Information Criterion (AIC) :

$$AIC = -2 * \log\text{-likelihood} + 2k$$

where k = number of parameters in the model

- **Ten-fold cross-validation** : estimation of the model performance on new data were evaluated through 3 measures :
 - **Multiple correlation coefficient** (R)
 - **Accuracy**
 - **Adjacent Accuracy** : proportions of predictions that were within one level of the human-assigned level for the given text [Heilman et al., 2008a]

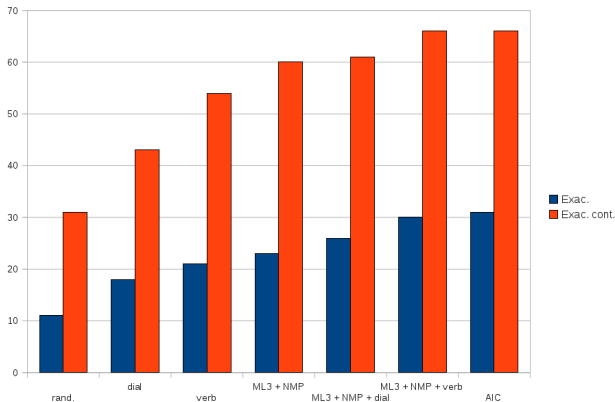
Features selection

- Stepwise selection process is sensitive to variations of the model and training data
- Selected variables for some models :
 - Proportional odds model :
 - **Corp6** : ML1 + ML3 + NMP + PPD + PI + PPEI1 + BINGUI + Futur + Impf + Infi + PPasse + Subp
 - **Corp9** : ML3 + NMP + PPD + PPEI2 + BINGUI + Cond + Futur + Impf + Infi + PPasse + Pres + Subp
 - Multinomial regression model :
 - **Corp6** : ML1 + NLM + NMP + BINGUI + Futur + Impf + Infi + PasseSim
 - **Corp9** : ML2 + ML3 + NMP + PPD + PPEI1 + Cond + Futur + Impf + PPasse + Subi + Subp

Relative importance of the features

However, we often find two lexical variables and NMP : they form the basis of the formula.

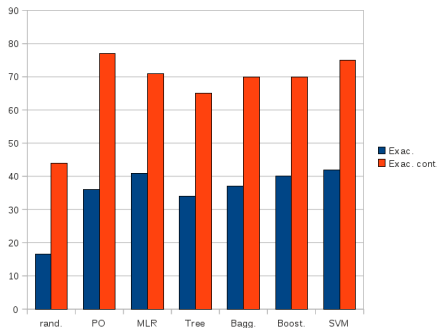
Example of decomposition of the accuracy :



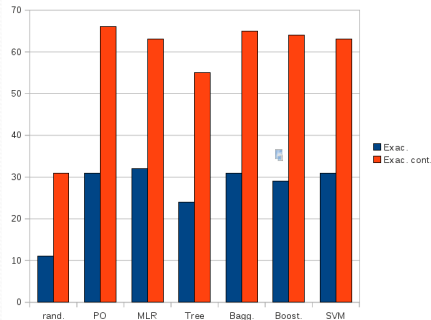
Comparison of the models

Results from the 10-folds cross-validation on both corpus :

Modèle à 6 niveaux



Modèle à 9 niveaux



Other similar studies

- **On French L1 :**

⇒ For a 5-classes problem : $R = 0.64$; Acc. and Adj. Acc. are not reported

[Collins-Thompson and Callan, 2005]

- **On English L1 :**

⇒ For a 12-classes problem : $R = 0.64$ (grades 1-6) and 0.79 (grades 7-12) ; Acc. and Adj. Acc. are not reported

[Collins-Thompson and Callan, 2005]

- **On English L2 :**

⇒ For a 12-classes problem : $R = 0.773$ (PO) and 0.582 (MLR) ; Adj. Acc. = 52% (PO) and 45% (MLR)

[Heilman et al., 2008a]

First conclusions

- It is the first specific formula for FFL that uses a NLP approach (and one of the few for FFL)
 - ⇒ The corpus includes a variety of text types, ensuring a wider coverage to the formula
- The criterion used (level of the textbooks according to the CEFR scale) seems questionable : the noise in the corpus can cause a poor learning.
- Our experiments suggest the (slight) superiority of SVM and logistic regression, a technique which is less demanding than the first.
- Optimizing the statistical aspects do not seem very useful for future improvements.

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

3 main lines of research

- **Statistical optimization**

- Multilayer neural networks
- Using reject option

- **New features**

- Experimenting with new variables, or from the literature, either drawing on current work on the reading process

- **Reducing the noise in the corpus**

- Using a manual exploration of the present corpus
- Collecting a new one, whose texts have been validated by teachers and learners

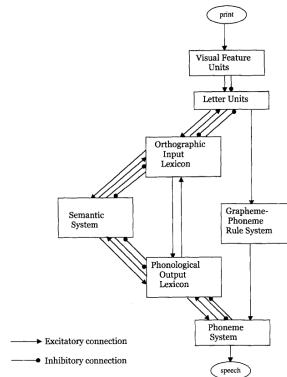
Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

The reading process

Reading is seen as a series of cognitive operations, here summarized as :

- 1 **Visual feature analysis** : the eyes move to capture the words on the page
- 2 **Word recognition** : activity specific to reading that is to recover the meaning of words in memory
- 3 **Comprehension** (share the same brain network for written and oral) : information extraction from the text and integration in long-term memory



Dual-route cascade model

[Coltheart et al., 2001]

Reading process and readability

Hypothesis

The textual elements that slow or hinder these processes can be considered as sources of difficulty, such as :

- Familiarity with words : frequency or age of acquisition (= 1st appearance in FFL textbooks ?)
- Imageability (more or less equivalent to the level of concreteness of words)
- Redundancy rate (TTR failure, N-gram ?)
- Number of different meanings for a word (facilitating effect or not ?)
- Degree of correspondence between the written and phonetic form
- ...

Differences between L1 and L2 reading

An observation

Reading processes in L1 and L2 differ. Therefore, the readability formulas should take into account these differences.

L1 reading : fluency in spoken language pre-exists

- Learn to read = develop a system of correspondences between graphemes and phonemes [Rayner et al., 2001]
- While the importance of decoding is crucial for beginners, it is understanding that prevails for the advanced reader
- Therefore, formulae based on lexical features should be more suitable for beginners, while structural of cognitive factors are better predictors for advanced learners [Chall and Dale, 1995]

Differences between L1 and L2 reading

L2 reading

- Reader must learn the language while they “learn” to read in L2.
- Conversely, he has already more concepts and knowledge about the world [Koda, 2005]
- There are some interferences with existing structures [Bernhardt, 2003]

Consequences on readability

- Before a given threshold, the student is severely handicapped by his lack of language skills in L2 : lexicon and syntax are predominant [Alderson, 1984]
- Beyond that threshold, he may transfer its reading skills from its L1 to its L2. Then, the importance of structural and cognitive factors increases again.
- Interferences with the mother tongue must also be considered, especially through the cognates [Laroche, 1979]

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research**
 - Studies about the reading process
 - Dmesure : a new tool for readability**
- 6 References

Dmesure : a new tool for readability

Dmesure : the alpha version

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) [Connexion](#)



Termes de recherche :

[Recherche avancée](#)

Difficulté du texte :

☐ A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2

[Utiliser 9 niveaux](#)

Rechercher

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@uclouvain.be

Terminé

Dmesure : 2 goals

Dmesure (stands for Difficulté Mesure) aims at the 2 following objectives :

- Provide a free tool helping FFL teachers in the use of the web as a corpus for finding teaching materials
- Provide a web 2.0 platform where teachers can participate in assessing the difficulty of texts they have collected through DMeasure and they have used in their teaching

Dmesure : a new tool for readability

Dmesure : the one-text interface

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) | [Connexion](#)



Dmesure vous offre également trois méthodes pour analyser directement un texte et en évaluer le niveau de difficulté à la lecture (sur l'échelle du CECR) pour un apprenant de français langue étrangère :

- Sélectionner la finesse de l'échelle : ☐ Echelle à 6 niveaux ☒ Echelle à 9 niveaux
- Copier-coller le texte dans le champ de saisie ci-dessous :

Étudier, suivre une formation dans un pays de l'Union européenne.

Un étudiant d'un pays de l'Union européenne peut suivre une formation professionnelle ou des études préparant à une profession dans un pays communautaire de son choix.
Il existe de nombreux programmes européens d'échange pour l'apprentissage des langues, les études en général, la formation professionnelle.
Pour avoir des renseignements sur ces programmes on peut s'adresser : aux universités, aux services culturels des ambassades, aux chambres de commerce ou aux associations étudiantes.

- Charger un texte depuis votre ordinateur :
- Indiquer l'URL d'un site internet :

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@uclouvain.be

Terminé

This text comes from the textbook Panorama (A2, p.159)



Dmesure : a new tool for readability

Dmesure : the one-text interface

[Rechercher un texte](#)[Introduire un texte](#)[Aide](#)[Connexion](#)

Dmesure

Difficulté estimée :A2 **Votre texte :**

Étudier, suivre une formation dans un pays de l'Union européenne.

Un étudiant d'un pays de l'Union européenne peut suivre une formation professionnelle ou des études préparant à une profession dans un pays communautaire de son choix.

Il existe de nombreux programmes européens d'échange pour l'apprentissage des langues, les études en général, la formation professionnelle.

Pour avoir des renseignements sur ces programmes on peut s'adresser : aux universités, aux services culturels des ambassades, aux chambres de commerce ou aux organismes européens.

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@ulbouvain.be

Terminé

The model did well on that one !!

Dmesure : a new tool for readability

Dmesure : web search service

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) | [Connexion](#)

Dmesure

Termes de recherche : [Recherche avancée](#)

Difficulté du texte : ☐ A1 ☐ A2 ☐ B1 ☒ B2 ☒ C1 ☒ C2 ☐ [Utiliser 9 niveaux](#)

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@uclouvain.be

Terminé

Dmesure : a new tool for readability

Dmesure : web search service

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) | [Valider un texte](#)

Connecté : beber | [Déconnecter](#)

Dmesure

Termes de recherche :

[Recherche avancée](#)

Difficulté du texte :

☐ A1

☐ A2

☐ B1

☐ B2

☐ C1

☐ C2

☒

[Utiliser 9 niveaux](#)

[Rechercher](#)

Vos résultats Dmesure

Status	Niv. Dmesure	Texte (premières lignes)	URI	Dmesure confidence
202	C2	Le réchauffement climatique, ou comment prendre conscience collectivement du réchauffement planétaire.	http://www.surfinfo.com/articles/rechauffement-climatique.htm	0.9999996818
	C2	A l'origine du réchauffement climatique, les gaz à effet de serre, avec en tête le gaz ... Le réchauffement climatique dilate les océans et fait fondre les calottes ...	http://www.encyclo-ecole.com/65067-253/rechauffement-climatique	0.942403918166
	C2	On appelle aussi réchauffement planétaire ou réchauffement global ... Les conséquences du réchauffement climatique sont l'augmentation de la ...	http://www.lesejourTV.com/lesconsequences-1014-rechauffement-climatique.htm	0.999939587723
	C2	Dossier : Réchauffement climatique. Bien qu'Objectif Liberté nait peu ou pour vocation ... 2009 - Réchauffement climatique : soupçons accablants de fraude ...	http://www.dossierliberte.fr/2009/rechauffement-climatique.html	0.999707783589
	C2	Le réchauffement climatique tant annoncé pourrait bien avoir lieu. Cette question sera bientôt lue lue par les centaines d'experts du monde entier ...	http://www.ardmore.com/le-rechauffement-climatique.html	0.93209366704
	C2	L'archipel menacé de disparaître sous les eaux à cause du réchauffement climatique ...	http://www.courserinternationale.com/2004/08/08/rechauffement-climatique	0.925094429665
	C2	Exposé de Hubert Reeves sur le réchauffement du climat terrestre ... Hubert Reeves - Réchauffement climatique: sikharia 24 vidéos	http://www.youtube.com/watch?v=19905-s1M-38	0.999935964287
	C2	Le réchauffement s'est notamment produit durant deux périodes, de ... Le réchauffement climatique devrait entraîner un abaissement de la mousson d'été en Asie ...	http://www.biodiversitydidac.org/lesconsequences-rechauffement-climatique.htm	0.875226577995
	C2	réchauffement climatique. Toute l'information en français. Recevoir ... Le réchauffement climatique, tu en entends parler tous les jours, partout, dans ...	http://www.lesdixhuit.com/actualites/actualites-climatique	0.9999342139
	C2	Le réchauffement climatique est "la plus grande crise humanitaire de ... Une manière saisissante d'illustrer la problématique du réchauffement climatique ...	http://www.research.fr/le-rechauffement-climatique.htm	0.991905622118

Page 1 sur 6

Enregistrements 1 - 10 sur 95

Terminé

Dmesure : a new tool for readability

Dmesure : the teacher interface

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) | [Valider un texte](#)
[Connecte : bebert](#) | [Déconnecte](#)


Merci de valider les textes que vous avez lus ou testés en classe. Cela permettra d'améliorer les performances de Dmesure.
[en savoir plus : [Rechercher un texte](#) ; [Valider un texte](#) ; [Annuler une validation précédente](#)]

Les derniers textes que vous avez consultés							
Valide	Date	Niv. dmesure	Texte (premières lignes)	Lit	Votre niv.	Validation	L3 étudiants
	2010-10-19	A1	Bonjour, mon ami. Comment vas-tu ?				
	2010-10-19	C2	Bonjour. Comment ça va ?		A1	lu	
	2010-10-19	A1	Bonjour. Comment ça va ?				
	2010-10-19	A1	Bonjour. Comment ça va ?				
	2010-08-12	C1	Kaori Muraji est considérée comme l'une des guitaristes classiques les plus en prometteuse d'aujourd'hui, ses sets et exécutés sont toujours vus avec intérêt par la communauté de musique classique. Née à Tokyo au Japon, Kaori Muraji a reçu ses premières leçons de guitare par son père				
	2010-06-03	C2	Stendhal : Un article de Wikipedia, l'encyclopédie libre. Aller à : Navigation, rechercher Question book-4.org Des informations de cet article ou section devraient être mises en ligne sur son	http://fr.wikipedia.org/wiki/Stendhal		lu	dans
	2010-07-30	C1	Les tentatives de paix pendant la Première Guerre mondiale déclenchent les différents appels lancés par le pape ou le président américain Woodrow Wilson, ainsi que les propositions formulées à partir de 1915, principalement par les deux principaux membres de la Triple Alliance, l'Allemagne et l'Autriche-Hongrie.		A2	testé	
	2010-07-29	C2	A small but important difference between "is null" and "is not" is the following: "is null" tests if an "expression" (not a "variable") is null, while "is not" tests if a "variable" has null value or is undefined. The difference is mentioned in the two following experiments:				
	2010-07-28	C1	Napoléon Bonaparte[] Né le 15 août 1769 à Ajaccio, en Corse ; mort le 5 mai 1821 sur l'île Sainte-Hélène. Il fut général, premier consul, puis empereur des Français. Il fut un conquérant de l'Europe continentale. Objet des son vivant d'une légende dorée comme d'une légende noire.		A1+		
	2010-07-28	C2	PhD Thesis Specialized Education Publications Prof. experience Teaching Contact (IMAGES) Thomas François PhD student, Assistant PhD Computational Linguistics Thomas François (thomasfrancois.be) http://central.fr.ac.be/thomasfrancois/				

Page 1 sur 2

Enregistrements 1 - 10 sur 14

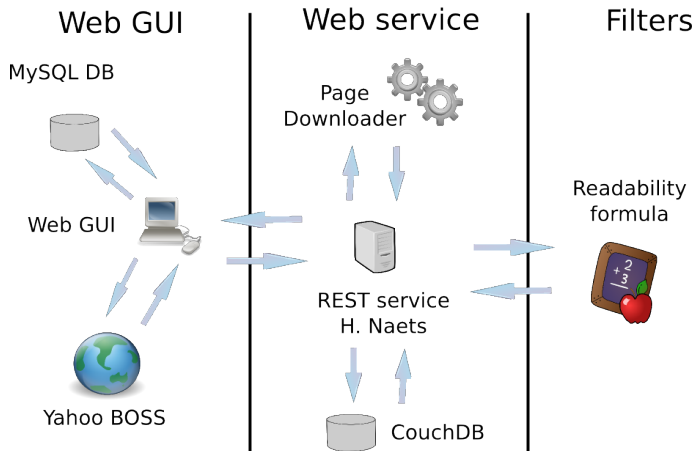
Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@ulouvain.be

Terminé

Dmesure : a new tool for readability

Dmesure : the architecture

Architecture of Dmesure



Dmesure : Some preliminary remarks

- While still needing to be debugged, the architecture seems suited to the task
- The one-text interface already gives good results, but the web search tool gives very poor predictions
- This is explained by the nature of text data found on the web. More work must be done on :
 - Cleaning the boilerplate
 - Checking the language correctness
 - Adapting the readability model to this specific task (some predictors are better suited to the web environment)

Dmesure : a new tool for readability

The end

Difficulté estimée :

A2 ?

Votre texte :

Merci pour votre attention.

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

Plan

- 1 Introduction : readability for ICALL
- 2 What is readability ?
- 3 Methodological steps
 - The corpus
 - Difficulty factors
 - The statistical models
- 4 Evaluation
- 5 Perspectives for future research
 - Studies about the reading process
 - Dmesure : a new tool for readability
- 6 References

References I



Alderson, J. (1984).

Reading in a foreign language : a reading problem or a language problem ?

In Alderson, J. and Urquhart, A., editors, *Reading in a Foreign Language*, pages 1–24. Longman, New York.



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 Ponton, C. (2004).

MIRTO : un système au service de l'enseignement des langues.

In *Proc. of UNTELE 2004*, Compiègne, France.

References II



Bernhardt, E. (2003).

Challenges to reading research from a multilingual world.

Reading Research Quarterly, 37(1) :112–117.



Bick, E. (2001).

The VISL system : research and applicative aspects of IT-based learning.

In *Proceedings of NoDaLiDa*, Uppsala.



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 III



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.



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.

References IV



Brysbaert, M., Lange, M., and Van Wijnendaele, I. (2000).

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

European Journal of Cognitive Psychology, 12(1) :65–85.



Chall, J. and Dale, E. (1995).

Readability Revisited : The New Dale-Chall Readability Formula.

Brookline Books, Cambridge.



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).

Information retrieval for language tutoring : An overview of the REAP project.

In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, page 545.

References V



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.



Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001).

DRC : A dual route cascaded model of visual word recognition and reading aloud.

Psychological Review, 108(1) :204–256.



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.



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 VI



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



Dale, E. and Chall, J. (1949).
The concept of readability.
Elementary English, 26(1) :19–26.



de Landsheere, G. (1963).
Pour une application des tests de lisibilité de Flesch à la langue française.
Le Travail Humain, 26 :141–154.



Dolch, E. (1948).
Problems in reading.
The Garrard Press, Champaign : Illinois.



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

References VII



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.

http://www-lipn.univ-paris13.fr/taln09/pdf/RECITAL4_.pdf.



Freund, Y. and Schapire, R. (1996).

Experiments with a new boosting algorithm.

In Machine Learning : Proceedings of the Thirteenth International Conference, pages 148–156.

References VIII



Gunning, R. (1952).
The technique of clear writing.
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.



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.

References IX



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



Howes, D. and Solomon, R. (1951).
Visual duration threshold as a function of word probability.
Journal of Experimental Psychology, 41(40) :1–4.



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



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.
Perspectives on Memory Research, pages 329–366.

References X



Koda, K. (2005).

Insights into second language reading : A cross-linguistic approach.
Cambridge University Press, Cambridge.



Laroche, J. (1979).

Readability measurement for foreign-language materials.
System, 7(2) :131–135.



Lively, B. and Pressey, S. (1923).

A method for measuring the “vocabulary burden” of textbooks.
Educational Administration and Supervision, 9 :389–398.



Mesnager, J. (1989).

Lisibilité des textes pour enfants : un nouvel outil ?
Communication et Langages, 79 :18–38.



Miltsakaki, E. and Trout, 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 XI



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>.



Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., and Seidenberg, M. (2001).

How psychological science informs the teaching of reading.

Psychological Science in the Public Interest, 2(2) :31–74.



Richaudeau, F. (1979).

Une nouvelle formule de lisibilité.

Communication et Langages, 44 :5–26.



Schmid, H. (1994).

Probabilistic part-of-speech tagging using decision trees.

In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

References XII



Selva, T. (2002).

Génération automatique d'exercices contextuels de vocabulaire.

In *Actes de TALN 2002*, pages 185–194.



Smith, E. and Senter, R. (1967).

Automated Readability Index.

Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.



Uitdenbogerd, S. (2005).

Readability of French as a foreign language and its uses.

In *Proceedings of the Australian Document Computing Symposium*, pages 19–25.



Verlinde, S., Selva, T., and Binon, J. (2003).

Alfalex : un environnement d'apprentissage du vocabulaire français en ligne, interactif et automatisé.

Romanesque, 28(1) :42–62.

References XIII



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.