Readability: a one-hundred-year-old field still in his teens

Thomas François

CENTAL (IL&C), Université Catholique de Louvain

NLG Summer school

July 23, 2015

Plan

Plan

Definition

A common definition of readability is :

> *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]

1. Focuses on text characteristics (reader characteristics are not directly modeled)
2. Readability aims at a group of readers (with homogeneous characteristics), not at an individual.
3. Considers comprehension, reading speed and motivation... in theory !

Readability is not...

Legibility

Legibility is the effect of typographical properties such as font size, font color, the color of the background, the presence of graphics, etc. on the reading process.

Comprehensability

Comprehensability focuses more on a single reader and sees reading as an interactive process including the text, the reader and the situation.

Home-made definition

- Readability aims at assessing the difficulty of texts for a given class of individuals
- Within this class, the characteristics are supposed homogeneous (strong hypothesis)
  $\longrightarrow$ as a consequence, only text characteristics are modeled (we can say that a given word is, in general, more difficult than this other word for the population).
- This means that reading is seen as an interactive process in which the reader and situation are controlled rather than overlooked... in theory !

From *Astérix chez Cléopâtre*.

Readability formulas

- Readability dates back to the 1920s, in the U.S.
- Main goal : develop methods to assess the difficulty of texts for a given population, without involving direct human judgements (and to save efforts).
- These tools = readability formulas.
  $\longrightarrow$ they are statistical models able to predict the difficulty of a text, given several text characteristics.
- Famous ones : [Dale and Chall, 1948], [Flesch, 1948], [Gunning, 1952], [Fry, 1968], or [Kincaid et al., 1975]

Classic formulas : an example

[Flesch, 1948] :

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

where :

Reading Ease (RE) : a score between 0 and 100 (a text for which a 4th grade schoolchild would get 75% of correct answers to a comprehension test)

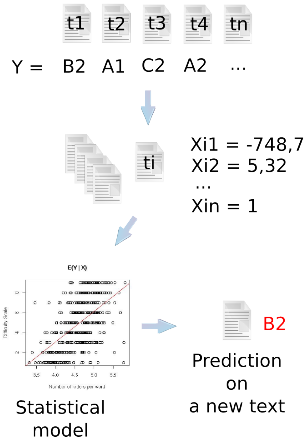
wl **:** number of syllables per 100 words

sl **:** mean number of words per sentence.

- Use of linear regression and **only a few** linguistic **surface** aspects.
- Claim that the formula can be applied to a large variety of situations.

Conception of a formula : methodological steps

1. Collect a corpus of texts whose difficulty has been measured using a criterion such as comprehension tests or cloze tests
2. Define a list of linguistic predictors of the difficulty, such as sentence length or lexical load
3. Design a statistical model (traditionally linear regression) based on the above features and corpus
4. Validate the model



What are the uses 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].
- Assessing automatic text simplification systems [Štajner and Saggion, 2013, Woodsend and Lapata, 2011, Zhu et al., 2010]

Helping writers : an example

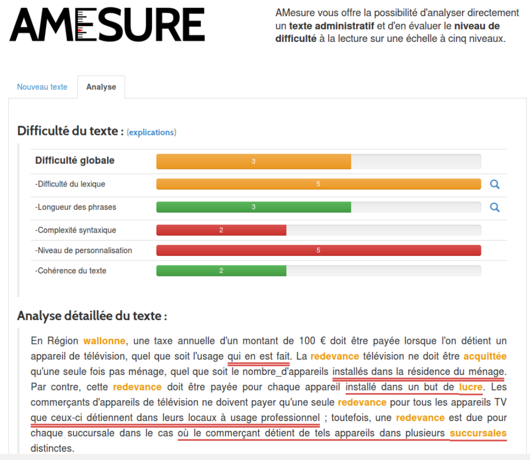


FIGURE : http://cental.uclouvain.be/amesure/

Calibration of books : a commercial example

Lexile Analyzer

- The Lexile framework is an educational tool that matches readers with books, using the Lexile scale [Stenner, 1996].
- Stenner and Malbert Smith III founded MetaMetrics in 1989, that was suported by the National Institute of Health.
- Example of the scale :

Title of work	**Lexile**
Twilight	720L
Harry Potter and the Sorcerer's Stone	880L
The Hobbit	1000L

Checking the output of a NLG system

Can be used to control the difficulty of NLP systems (MT, NLG, ATS)

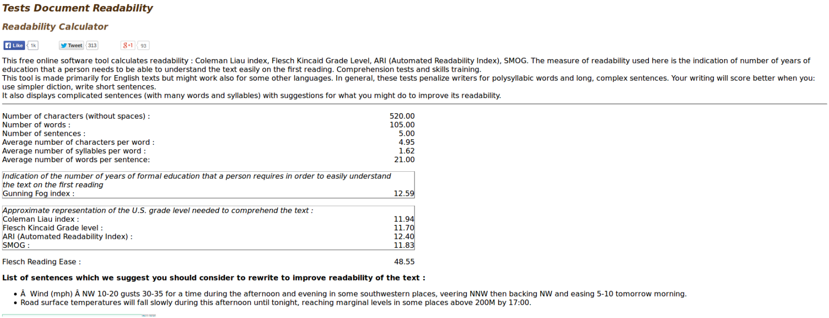
Example from Ehud Reiter's presentation

Overview Road surface temperatures will reach marginal levels on most routes from this evening until tomorrow morning.

Wind (mph) NW 10-20 gusts 30-35 for a time during the afternoon and evening in some southwestern places, veering NNW then backing NW and easing 5-10 tomorrow morning.

Weather Light rain will affect all routes this afternoon, clearing by 17 :00. Fog will affect some central and southern routes after midnight until early morning and light rain will return to all routes. Road surface temperatures will fall slowly during this afternoon until tonight, reaching marginal levels in some places above 200M by 17 :00.

Checking the output of a NLG system


Tests Document Readability

Readability Calculator

Like 1k Tweet 313 +1 93

This free online software tool calculates readability : Coleman Liau index, Flesch Kincaid Grade Level, ARI (Automated Readability Index), SMOG. The measure of readability used here is the indication of number of years of education that a person needs to be able to understand the text easily on the first reading. Comprehension tests and skills training.
This tool is made primarily for English texts but might work also for some other languages. In general, these tests penalize writers for polysyllabic words and long, complex sentences. Your writing will score better when you: use simpler diction, write short sentences.
It also displays complicated sentences (with many words and syllables) with suggestions for what you might do to improve its readability.

Number of characters (without spaces) :	520.00
Number of words :	105.00
Number of sentences :	5.00
Average number of characters per word :	4.95
Average number of syllables per word :	1.62
Average number of words per sentence:	21.00

Indication of the number of years of formal education that a person requires in order to easily understand the text on the first reading	
Gunning Fog index :	12.59

Approximate representation of the U.S. grade level needed to comprehend the text :	
Coleman Liau index :	11.94
Flesch Kincaid Grade level :	11.70
ARI (Automated Readability Index) :	12.40
SMOG :	11.83

Flesch Reading Ease :	48.55

List of sentences which we suggest you should consider to rewrite to improve readability of the text :

- Â Wind (mph) Â NW 10-20 gusts 30-35 for a time during the afternoon and evening in some southwestern places, veering NNW then backing NW and easing 5-10 tomorrow morning.
- Road surface temperatures will fall slowly during this afternoon until tonight, reaching marginal levels in some places above 200M by 17:00.

FIGURE : http://www.online-utility.org/english/readability_test_and_improve.jsp

Assessing ATS systems

Use in ATS systems :

- [De Belder and Moens, 2010] applied Flesch-Kincaid to the output of their system to characterize it in terms of grade levels.
- [Zhu et al., 2010] computed the Flesch and Lix scores + the perplexity of a trigram model, based on [Schwarm and Ostendorf, 2005].
- [Woodsend and Lapata, 2011] tried Flesch RE and Coleman-Liau, but selected Flesch-Kincaid.
- [Štajner and Saggion, 2013] studied more closely this issue and used three formulas for Spanish (Spaulding's and Anula's)

$\longrightarrow$ Strangely, only "classic" formulas are used !

Main field of application : ICALL

- ICALL (intelligent computer-assisted language learning) use NLP tools within CALL applications
- Examples of use :
  - help the automatic retrieval of authentic texts for teaching purposes
  - assistive tools for non supervised reading or essay writing
- ICALL may also help relieve teachers of repetitive tasks :
  - Automated design of exercises (included adaptative exercises) aimed at the assimilation of specific linguistic forms (such as collocation, grammar notion...).
  - Automated feedback and error detection in learner's production.

Readability formulas can be useful for several of these tasks

Two examples of application

Automated design of exercises based on a corpus

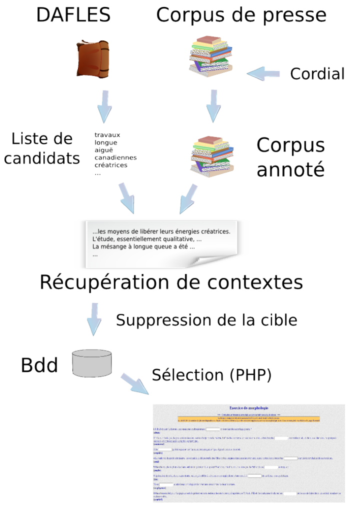
- English : **Cloze tests** [Coniam, 1997, Brown et al., 2005, Lee and Seneff, 2007, Skory and Eskenazi, 2010] ;
  MCQ [Heilman, 2011, Mitkov et al., 2006]
  WERTi [Amaral et al., 2006]
- French : **ALEXIA** [Chanier and Selva, 2000] ;
  ALFALEX [Selva, 2002, Verlinde et al., 2003] ;
  MIRTO [Antoniadis and Ponton, 2004, Antoniadis et al., 2005].

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

- English : **IR4LL** [Ott, 2009] ; **REAP** [Heilman et al., 2008b], **READ-X** [Miltsakaki and Troutt, 2008]
- French : **DMesure** [François and Naets, 2011]
- Portuguese : **REAP** [Marujo et al., 2009]

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 Fargevieille.
{*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*}

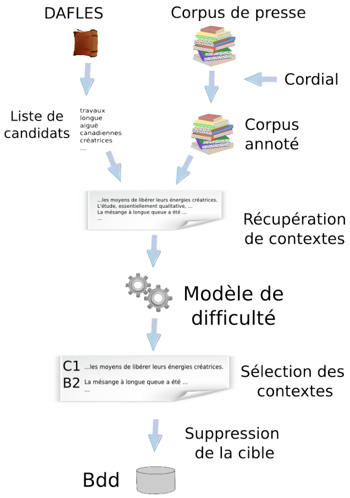
6 Opération de séduction, sans doute, mais qui reflète à l'évidence les aspirations d'une société de la férule des ayatollahs.
{*las*}

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*}

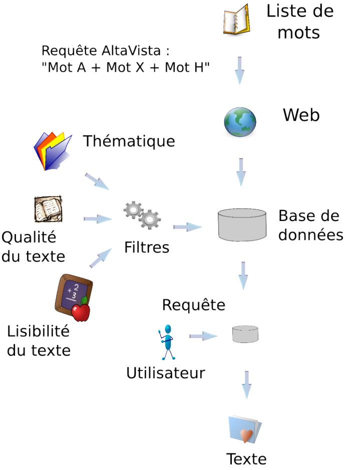
Readability model as a solution

- 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, 2004b]
  - 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

Plan

Main periods in readability

5 major periods in readability :

1. **The origins** : first works in the field. A lot of interesting perspectives, often forgotten in the current studies !
2. **Classic period** : formulas are based on linear regression and mostly use two **indices** (one lexical, one syntactic)
3. **The cloze test era** : concerns arise about motivated features (= cause of difficulty) and difficulty measurement
4. **Structuro-cognitivist period** : takes into account newly discovered textual dimensions (cohesion, structure, inference load, etc.).
   $\longrightarrow$ Period of strong criticisms against the classical formulas
5. **AI readability** : NLP-enabled features are combined with more complex statistical algorithms.

Lively and Pressey (1923)

- [Lively and Pressey, 1923] is generally acknowledge as the first "readability formula"
- The focus only on lexical load, through three indexes :
  1. number of different words
  2. proportion of words absent from [Thorndike, 1921]'s list
  3. a weighted median of the word ranks in the same list (approximation of word frequency).
- They did not combine the indexes. They simply compared the features with a set of 15 textbooks and a newspaper whose difficulty was "known"...

  $\longrightarrow$ median appears to be the best of the three.

Vogel and Washburne (1928)

- [Vogel and Washburne, 1928] are responsible for the design of the classic methodology, still used till today in some papers.
  - They define a list of predictors (textual characteristics) and combine them with a multiple linear regression
  - They stress the importance of the criteria : the dependent variable representing text difficulty.
- Corpus : 152 books assessed according their difficulty and interest by at least 25 children for each of them (part of the *Winnetka Graded Book List*).
- Manual parameterization (with 20 volunteering teachers) of a large amount of linguistic features

  $\longrightarrow$ metrics of the lexical load, of the syntactic structures, ratio of P.O.S, and information about paragraph and book structure.

Vogel and Washburne (1928)

The final formula :

$$X_1 = 17,43 + 0,085\, X_2 + 0,101\, X_3 + 0,604\, X_4 - 0,411\, X_5$$

- X_1 : score to a reading test (*Standford Achievement Test*) ;
- X_2 : number of different word in a 1000 word sample ;
- X_3 : number of prepositions in this sample ;
- X_4 : number of words in the sample that are absent from Thorndike's list ;
- X_5 : number of simple proposition among a 75-sentence sample.

- The multiple correlation coefficient, *R*, reaches 0, 845
- First formula with syntactic features
  $\longrightarrow$ Much more varied features than just the mean number of words per sentence that is framed as classical !

Other interesting works

- [Ojemann, 1934] and [Dale and Tyler, 1934] adapt previous work for adults.
- [Ojemann, 1934] also defines a methodologically stricter criterion : the mean score to a reading comprehension test.
- [McClusky, 1934] investigates the use of reading speed as a criterion.
- [Gray and Leary, 1935] explores as much as 289 features, among which information about idea organization, coherence, etc.

  $\longrightarrow$ among these, they finally implement 44 variables (lexical, syntactic and even number of personal pronoun)

Characteristics of the classic formulas

- Whereas the formulas become more and more complex, integrating more features, [Lorge, 1939] breaks with previous work, seeking more simplicity and efficiency.
  $\longrightarrow$ originates from
  1. detection of multicollinearity between predictors
  2. in the sake of simplicity (still manual work)
- Only lexical and syntactic features are considered
- The most popular criterion is the *Standard Test lessons in Reading* de Mc-Call et Crabbs (1938)

Mc-Call et Crabbs series

Textbook series for children (3rd grade to 8th grade) whose calibration was operated as follows :

> *Each lesson was administered to students along with the Thorndike-McCall Reading Scale (which yields grade scores). Sample sizes generally consisted of several hundred students for each lesson. To determine the grade scores for a lesson, a graph was made with a dot placed at the intersection of each student's raw score and his Thorndike-McCall grade score. A smooth curve was the drawn through the dots and a grade score assigned to each lesson raw score.*
> *[Stevens, 1980]*

This criteria was used by
[Lorge, 1944, Flesch, 1948, Dale and Chall, 1948, Gunning, 1952]

Summary of the most famous classic formulas

- [Flesch, 1948] introduces his Reading Ease (RE) and Human Interest (HI) formulas
  $\longrightarrow$ the latter aims to model the interest of a text, based on "personal" words.
  Issues : formula intended to adults, calibrated on children material + HI is also calibrated on McCall and Crabbs !
- [Dale and Chall, 1948] designed one of the best formula for educative purposes
- [Flesch, 1950] are the first to explore the issue of text abstraction (based on certain grammatical categories)
- [Gunning, 1952] also designed a famous formula, the *Fog index*, more business-oriented, that defines complex words as words with more than 3 syllables.

These work are followed by a step of refining and specializing the formula (1953 to 1965).

Characteristics of the cloze revolution

- The cloze test (= fill-the-blanks) was coined by [Taylor, 1953] as a tool to assess reading comprehension.
- Coleman (1965) is the first to apply it in readability as a new criterion.
- Simultaneously, a second revolution – technological – also contributes to change the field
  $\longrightarrow$ First automated approaches of readability [Smith, 1961]
- With automation, formulas with more variables reappear [Bormuth, 1966]
- More importantly (although it did not had much influence), some researchers designed a set of formulas (for various situations), rather than one universal model.
- Classic approaches (few variables + manual counting) keep on

Smith's work

- [Smith, 1961] coined the *Devereaux index*, intended to children from grade 2 to grade 8.
- Following the simplification trend in the 50's, he argues that letter per word is as efficient as the syllable count or % of simple words.
- This feature is also simpler to count (no linguistic knowledge involved)
- [Danielson and Bryan, 1963] adapted the Smith's formula on an UNIVAC 1105 computer.

Bormuth

Bormuth is one of the most inspiring researcher in the field :

- He address several methodological issues of the field :
  - He shows that the relation between the predictors and the criterion is not linear, rather curvilinear.
  - There is no interaction between features and the level, which means that one unique formula is enough
  - He argues that classic formulas “contain too few variables”
- Based on cloze test, he models readability at text, sentence, and word level !
- He is the first one to use parse tree-based features (showing that are less efficient than number of word per sentence) !
- He stresses the need to report correlation coefficient from a test set and not the training set.
- Work : [Bormuth, 1966, Bormuth, 1969]

Other studies

- [McLaughlin, 1969] : the SMOG formula, with only “one” predictor
- [Kincaid et al., 1975] : adapt three formulas (including Flesch) to the army context
  - Very popular model in current NLP studies...
  - although it was calibrated on soldiers, using fragments from military instruction manual !
- [Coleman and Liau, 1975] argue that converting a text to punched cards is not faster than manually applying a formula
  $\longrightarrow$ used an optical scanner

Characteristics of the period

The rise of constructivism

- Cognitivists and linguists move beyond words and sentences
- Constructivism vision of reading : “people, rather than texts, carry meaning” [Spivey, 1987]
- Mental processes involved in reading are taken into account (memory, understanding, etc.)
- In linguistics, focus on cohesion, coherence and text grammar.

Criticism towards classic readability

- Readability needs to go further sentences and surface variable !
- There is auto-criticism even within the “classic approach” [Harris and Jacobson, 1979]
- Some structuro-cognitivists were very critical
  $\longrightarrow$ e.g. : [Selzer, 1981] : *Readability is a four-letter word*

Some structuro-cognitivist works

- focus on text organisation
  [Armbruster, 1984]
- on discourse cohesion
  [Clark, 1981, Kintsch, 1979]
- on inferential load
  [Kintsch and Vipond, 1979, Kemper, 1983]
- on rhetoric structure
  [Meyer, 1982]
- ...

Pro and cons of the structuro-cognitivist approach

- It stresses the importance of considering variables that are likely causes of reading difficulties rather than just proxies.
- [Kintsch, 1979] designed a cognitive model of readability that exhibit a $R = 0.97$, but :
  - mean frequency of words is one of the two best features !
  - [Miller and Kintsch, 1980] confirms that frequency and word length are as important as the number of inferences or reinstatement searches
- [Kemper, 1983] compared a cognitive formule of her own with the Dale and Chall formula and obtained similar results !

$\longrightarrow$ Lexico-syntactic features appears as predictive as structuro-cognitive ones, which are more complex to implement !

The progress of automation

- At first, automation goes with a simplification of linguistic realities :
  - [Coke and Rothkopf, 1970] argue for using the amount of vowels as a count of syllables.
  - The predictors considered becomes more and more surface ones.
- [Daoust et al., 1996] use NLP tools (e.g. P.O.S.-tagger) to parameterize their features
- [Foltz et al., 1998] measure text coherence based on LSA.
- [Si and Callan, 2001] define readability as a classification problem and applies state-of-the-art machine learning methods to it.

Main trends in AI readability

- [Collins-Thompson and Callan, 2005] draw from the language model of Si and Callan (2001), enhance it and include it within a Naïve Bayes classifier.
- [Schwarm and Ostendorf, 2005] implement syntactic variables, based on a syntactic parser and combine all their features within a SVM model.
  → syntactic features do not contribute much to the model !
  → the first to use the Weekly Reader (educative newspaper).
- [Heilman et al., 2007] experiment the contribution of such syntactic features for L2 and show that they are more important.

Main trends in AI readability

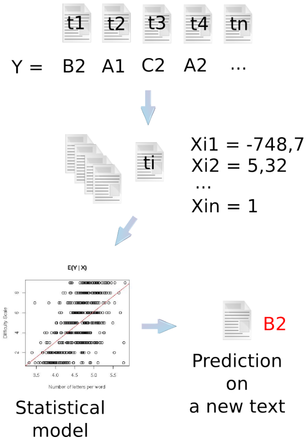
Whereas the first studies focused on lexicon and syntax, then appears work also considering semantic, discourse or cognitive variables.

- [Crossley et al., 2007] design the first NLP-enabled readability formula combining lexical, syntactic and cohesive dimensions, based on Coh-Metrix.
  $\rightarrow$ The cohesive factor is however no significative in the model ($p = 0.062$) !
- [Pitler and Nenkova, 2008] introduce a fully-fledged readability model and confirms the impact of some cognitive factors.
- [Tanaka-Ishii et al., 2010] see readability as a sorting problem : good results.
- [Vajjala and Meurers, 2012] introduce SLA variables in the model and got very high classification accuracy on the Weekly Reader ($93,3\%$).

Plan

The common methodology : a reminder

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 challenge

- Readability assumes that we know which texts are more difficult than other...
  $\longrightarrow$ what means “difficult” ? How can we measured it ?
- It is measured through another variable, easier to measure and correlated with difficulty
  $\longrightarrow$ we call it the criterion !
- Several criteria exists and had been used in readability...
  $\longrightarrow$ none are perfect !

Criteria for readability

Expert judgments : Several experts of a population have to agree on the level of the texts

Texts from textbooks : Variant of expert judgment. Texts are given a level by experts for educative purposes upstream the experiment.

Comprehension test : text comprehension is assessed through questions and the mean of scores for a text = its difficulty.

cloze test : see before

reading speed : reading speed is measured, generally combined with some questions, to check for understanding

recall : proportion of a text that can be recall by a subjects after reading.

Non expert judgements : [van Oosten and Hoste, 2011] show that N (N > 10) non experts can annotated as reliably as experts

...

Expert judgments

Pros and cons

Pros : supposedly reliable, rather convenient (no subjects)
Cons : population is not directly tested

$\longrightarrow$ we model the experts' view of difficulty for the given population

Issue of heterogeneity

- [van Oosten et al., 2011] had 105 texts assessed by experts (as pairs) and clustered them by similarity of judgements (train one model per cluster).
  $\rightarrow$ this leads to different models, whose intracluster performance > intercluster.
- [François et al., 2014a] had 18 experts annotate 105 administrative texts (with an annotation guide)
  $\rightarrow 0.10 < \alpha < 0.61$ per batch (average = 0.37).
- High agreement seems difficult to reach in readability (SemEval 2012 : $\kappa = 0.398$ on the test set).

Using textbooks

Pros and cons

Pros : very convenient (no subjects and no experts !)
$\longrightarrow$ more popular criterion in AI readability, due to the large training corpus needed
Cons : population is not directly tested, heterogeneity

- Very few corpora available : Weekly Reader is mostly used [Schwarm and Ostendorf, 2005, Feng et al., 2010, Vajjala and Meurers, 2012]
$\longrightarrow$ risk : high dependence towards one training corpus, as McCall and Crabbs lessons in classic period [Stevens, 1980]
- This dependence has consequences :
  - formulas will be specialized towards this corpus (coefficients)
  - always the same population and type of texts considered
- Problem of heterogeneity between textbook series

Example of heterogeneity in a corpus

Corpus of L2 textbooks [François and Fairon, 2012]

The textbook corpus

- Criterion = expert judgments = textbooks (level of a text = level of the textbook).
- We used the CEFR scale (official EU scale for L2 education), which has 6 levels [Conseil de l'Europe, 2001]
- Levels are : A1 (easier), A2, B1, B2, C1, and C2 (higher).
- We extracted 2042 texts from 28 FFL textbooks.

Example of heterogeneity in a corpus

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

Other criteria

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

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

Reading speed : population tested, strong theoretical validity, but very expensive !
⟶ self-paces presentation technique might be a cheaper alternative

Recall : population tested, but influence of memory performance + do not correspond to a psychological reality for [Miller and Kintsch, 1980].

Conclusion about criterion

- No optimal criterion !
- Best seems to be experts judgements, provided there is a controlled annotation process (and good experts)
- Most promising, reading speed, but not enough validating studies
- Criterion is probably the factor that impact the most readability formulas performance (difficult to compare all work)

Predictors in readability

Characteristics of a good predictor

- Should have a high correlation with the criteria
  Beware ! [Carrell, 1987] better separated corpus leads to better correlation... and performance !
- Should have a low correlation with other predictors
- Predictors should be measured in reliable and reproducible way (not always possible)
- Today, most of the features are psycholinguistically motivated [François, 2011]

Main types of predictors in readability

Classes of predictors

Predictors are generally classified according the text dimension they model :

- **Lexical features**
- **Syntactic features**
- **Semantic features**
- **Discourse features**
- **Other features** : specialized predictors

Lexical predictors

- frequency or log(freq) of words [Howes and Solomon, 1951]
- percentage of words not in a reference list of simple words [Dale and Chall, 1948]
- N-gram models [Si and Callan, 2001, Pitler and Nenkova, 2008, François, 2009, Kate et al., 2010]
  $\longrightarrow$ needs to be normalized (e.g. n-root)
- measure of the lexical familiarity (not implemented)
- measure of the lexical diversity (e.g. Type-token ratio) [Lively and Pressey, 1923]
- age of acquisition [Vajjala and Meurers, 2014b]
- orthographical neighbors [François and Fairon, 2012]
- word length (in letter, syllables, affixes, etc.) [Gray and Leary, 1935]

Lexical predictors generally stand out as the best category [Chall and Dale, 1995]

Syntactic predictors

- sentence length [Vogel and Washburne, 1928]
- proxies for the syntactic complexity :
    - % of simple sentence [Vogel and Washburne, 1928]
    - type of phrases or clauses (adjectival, prepositional, etc.)
    - length of dependency links [Dell'Orletta et al., 2014b]
- difficulty of actual syntactic structures [Bormuth, 1969, Heilman et al., 2007]
- tree-based features (word depth of Yngve (1960)), depth of tree, etc. [Bormuth, 1969, Schwarm and Ostendorf, 2005]
- P.O.S.-tag ratio [Vogel and Washburne, 1928, Bormuth, 1966]
- complexity of the verbal tenses and moods [Heilman et al., 2007, François, 2009]

Semantic predictors

- proportion of abstract words [Lorge, 1939, Henry, 1975, Graesser et al., 2004, Sheehan et al., 2013]
- imageability [Graesser et al., 2004, Sheehan et al., 2013]
- personnalisation level of the text [Dale and Tyler, 1934]
- conceptual density [McClusky, 1934, Kemper, 1983]
- polysemy : the impact of the number of senses [Beinborn et al., 2012]
- compositional semantics [Beinborn et al., 2012]
  $\longrightarrow$ sentences are represented by semantic networks consisting of conceptual nodes linked by semantic relations (nb. of nodes and relations).

Discourse predictors

- inference load [Kintsch and Vipond, 1979]
- coherence level measured with LSA [Pitler and Nenkova, 2008]
- likelihood of texts as a bag of discourse relations [Pitler and Nenkova, 2008]
- probabilities of transition between syntactic functions of entities [Pitler and Nenkova, 2008]
- other characteristics of lexical chains [Feng et al., 2009, Todirascu et al., 2013]
- lexical tighness [Flor and Klebanov, 2014]
- detection of dialogue [Henry, 1975]
- interactive/conversational style [Sheehan et al., 2013]

Other predictors

- characteristics of MWE [François and Watrin, 2011]
- SLA-based features [Vajjala and Meurers, 2012]
- Using only words [Tanaka-Ishii et al., 2010]
- ...

The modelling

- Annotated corpus + features $\longrightarrow$ training of your favorite ML algorithm
  $\rightarrow$ Most popular today = SVM, but also regression (linear or logistic), etc.
- Typical ML training process (X-folds cross-validation)
- Evaluation metrics differs :
  - 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).

Example of the performance

- Performance remains unsatisfactory for commercial usage in most studies !

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

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

Readability for other languages

English is dominant in the field, but there are work for other languages :

- French : [Henry, 1975, François and Fairon, 2012, Dascalu, 2014]
- Spanish : [Spaulding, 1956, Anula, 2007]
- Japanese : [Tanaka-Ishii et al., 2010]
- Swedish : [Pilán et al., 2014]
- Italian : [Dell'Orletta et al., 2011]
- German : [Vor der Brück and Hartrumpf, 2007, Hancke et al., 2012]
- Chinese : [Sung et al., 2014]
- Arabic : [Al-Khalifa and Al-Ajlan, 2010]

Conclusion

- Readability is an old lady, that did not evolved much methodologically.
- Lately, NLP-ebabled features and ML revitalized the field
$\rightarrow$ However, we give up some validity in the criterion to get more data !
- Some textual dimensions are still to be explored (semantics, macrostructure, pragmatics)
- Performance are OK, but seems unsatisfactory for a large commercial usage
$\rightarrow$ we still do not know exactly what is difficulty !
- Readability and text simplification are getting closer to each other.

Plan

Some issues in readability

1. Corpus issues (availability, validity, heterogeneity)
2. Specialization of the formula (genre, public)
3. Lots of features available, but are they all similarly useful ?
4. Modeling smaller textual fragments

Corpus issues

Already discussed before (lack, heterogeneity)...

- Current methods requires large annotated corpora, but very few are available :
  - Weekly Reader (seems possible to get it)
  - Wikipedia - Vikidia (used as a two-level corpus)
- There is a need for reference corpus, freely available !
- Other issue : scale depends on the population...
  $\rightarrow$ which scale to favour ?
- Same need in each different language

Corpus issues

Crowdsourcing as a solution ?

- Crowdsourcing can be a way to collect a large amount of difficulty labels for texts [De Clercq et al., 2014]
- Integrate it within a reading plateforme that stimulates readers to produce data !

Specialization of the formulas

What is specialization ?

It first meant defining a specific population of interest (eg. children, L2 readers, etc.) AND adapting the model to take into account the specificities of that population.
NOW, we also consider specializing formulas for text genre.

In other words, it amounts to :

- Use a corpus of the target type of texts, 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
(e.g. MWE for L2 readers [François and Watrin, 2011])

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, 2011]
  - informative and literary texts [Dell'Orletta et al., 2014a]

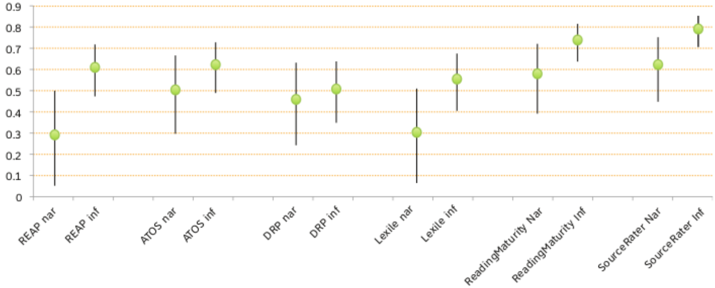
Rationales for population adaptation

- Common practice : try to apply a L1 formula to a L2 context
- 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 cloze scores was high ($r = 0.915$)
  $\rightarrow$ Retrained the 6 formulas on this corpus and get a small gain only.

We need more tests on real readers, with modern formulas !

Rationales for genre adaptation

- [Nelson et al., 2012] distinguishes between performance of various famous models on narrative and informative texts



Rationales for genre adaptation

- [Sheehan et al., 2013] analyzed differences between literary and informative texts :
  - Literary texts includes more core vocabulary of the language [Lee, 2001]
  - "Content area texts often received inflated readability scores since key concepts that are rare are often repeated, which increases vocabulary load" [Hiebert and Mesmer, 2013].

  $\longrightarrow$ Readability formulas tends to overestimated informative text difficulty and underestimate it for literary texts !
- [Sheehan et al., 2013] developed an unbiaised model for each type of texts.
- [Dell'Orletta et al., 2014a] confirmed that a readability model can only correctly assigned labels to the same genre of texts it was trained on.

Type of texts : an experiment

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

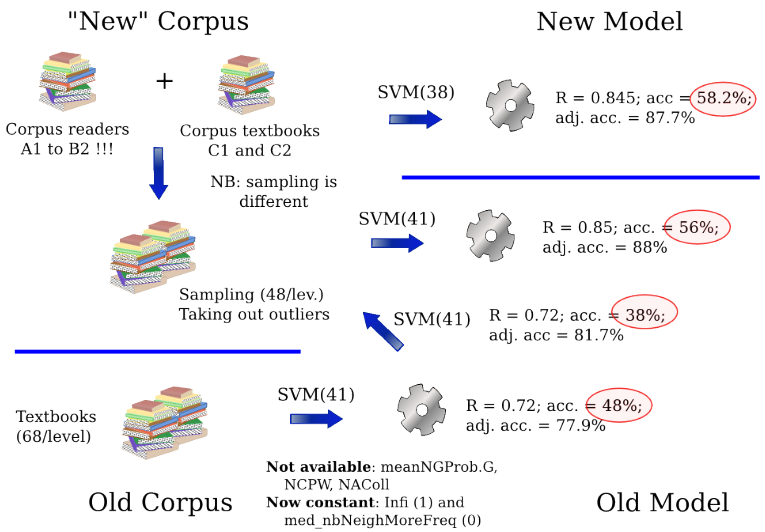
29 simplified readers collected :

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

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

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

Even mixed models seems to have trouble !



Contribution of the variable families

Based on [François and Fairon, 2012], 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.).

The semantic/discourse features

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

An experiment with reference chains features

- In [Todirascu et al., 2013], we annotated 20 texts across CEFR levels A2-B2 as regards reference chains.
- We computed 41 variables, among which :
  - POS-tagged based features (e.g. ratio of pronouns, articles, etc.)
  - lexical semantic measures of intersentential coherence, based on tf-idf VSM or LSA
  - Entity coherence [Pitler and Nenkova, 2008] : counting the relative frequency of the possible transitions between the four syntactic functions (S, O, C and X)
  - Measures of the entity density and length of chains
  - New features : Proportion of the various types of expressions included in a reference chain (e.g. indefinite NP, definite NP, personal pronouns, etc.
- We show that a few variables based on reference chains are significantly correlated with difficulty, even on a small corpus

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

Classical features vs. NLP-based features

Contrasted results

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

Comparing both types of information

- [François and Miltsakaki, 2012] compared SVM models with the same number of features (20), some are “classical“ and the others NLP-based $\rightarrow$ ”Classical“ : $acc. = 38\%$ vs. NLP-based : $acc. = 42\%$ ($t(9) = 1.5; p = 0.08$) !
- When both types are combined within a SVM model, performance rise from $acc. = 37,5\%$ to 49%.

What have we learned from this ?

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

Moving below texts

- Traditionnally, readability aimed to assess text difficulty
  $\longrightarrow$ several samples of at least 100 words !
- Apply to shorter fragments, they usually fails
  $\longrightarrow$ due to the limited amount of material and statistical approach
- However, for web use [Collins-Thompson and Callan, 2005] or exercise generation [Pilán et al., 2014], we need model able to perform well on short context !
- Extreme approach : measure word difficulty with readability methods.

Sentence readability

- First to investigate is probably [Bormuth, 1966] (using cloze test) ! $\longrightarrow$ model with 6 variables obtains $R = 0.665$ against $R = 0.934$ for text level !
- [Fry, 1990] : classic formula, adapted for short passages :

$$\text{Readability} = \frac{\text{Word Difficulty} + \text{Sentence Difficulty}}{2} \tag{1}$$

  - the analyst selects at least three essential content words and look their grade level up in the *Living Word Vocabulary* [Dale and O'Rourke, 1981]
  - In each sentence, count words, then transform the score into a grade level using a table.

Sentence readability : a renewal

- [Collins-Thompson and Callan, 2004a] : Web-oriented model
  - Use a smoothed Unigramm model
  - Hypothesis : has a finer-grained model of word usage, so better able to assess short texts
    $\longrightarrow$ // with idea of [Fry, 1990]
- [Dell'Orletta et al., 2011] combines lexical and syntactic features within a SVM
  $\longrightarrow$ accurracy at document level = 98% ; at sentence level = 78%
- [Pilán et al., 2014] : similar approach, but add semantic features (polysemy, idea density, etc.)
  $\longrightarrow$ accurracy at sentence level = 71% (also binary)
- [Vajjala and Meurers, 2014a] : add SLA features for 66%.

Word “readability”

- First to investigate word difficulty in context (e.g. word depth) is again [Bormuth, 1969] !
  $\longrightarrow$ model with 5 variables obtains $R = 0.505$ against $R = 0.934$!
- [Shardlow, 2013] wants to assess word difficulty in the context of ATS (for substitution)
  $\longrightarrow$ They use Wikipedia edit history.
- [Gala et al., 2013] learns a SVM model based on a lexicon with three difficulty level [Lété et al., 2004] and 49 lexical variables (freq., morphemes, nb. letters, polysemy, etc.)
  $\longrightarrow$ Beat the frequency baseline only by 2% !

Word “readability”

Another approach is to learn graded lexicon from corpus

- [Brooke et al., 2012] learns to discriminate between pairs of words
- Create 4500 pairs from words in three differents levels and then crowdsourced the pair relation (first learned word)
- They combine document readability, simple and co-occurence features.

- FLELex [François et al., 2014b]

FLELex

- **Goal :** build a lexical resource describing the distribution of French words accross the 6 CEFR levels.
- **Method** : Estimate the probability from a corpus of annotated texts for FFL (above corpora).
  - Texts were tagged with TreeTagger and a CFR-tagger able to detect MWE [Constant and Sigogne, 2011]
  - Learner's knowledge of MWE lags far behind their general vocabulary knowledge [Bahns and Eldaw, 1993]
  - We used the dispersion index [Carroll et al., 1971] to normalize frequencies
- FLELex-TT has 14,236 entries (no MWEs, but manually cleaned)
- FLELex-CRF includes 17,871 entries (MWEs, nut not cleaned yet)

Example of entries

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

The resource is freely available at
`http://cental.uclouvain.be/flelex/`

Other languages in progress (Swedish, Spanish,...)

General Conclusion

- Readability is an old lady... falling back to its teens
  $\longrightarrow$ Contribution of NLP revived the field and there is plenty to do
- Issues of corpora (no reference, performance varies, annotation validity)
- The unit is the token (sometimes MWE), but must be the sense !
- Specialisation IS an issue... there is a need for adaptive and personalized formulas
- Porting the model to sentence level and get good results remains a challenge
- Score or diagnosis ? Depends on the application.

Introductory materials

State-of-the-art papers/books

- KLARE, G. (1963). The Measurement of Readability. Iowa State University Press, Ames, IA.
- CHALL, J. and DALE, E. (1995). Readability Revisited : The New Dale-Chall Readability Formula. Brookline Books, Cambridge.
- COLLINS-THOMPSON, K. (2014). Computational Assessment of Text Readability : A survey of current and future research. In François, T. and Delphine B. (eds.), *Recent Advances in Automatic Readability Assessment and Text Simplification*. Special issue of International Journal of Applied Linguistics 165 :2 (2014). 243 pp. (pp. 97–135).
- FRANÇOIS, T. (2011). La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? *International Journal of Applied Linguistics (ITL)*, 160.

Bibliographies on the web

- https ://sites.google.com/site/readabilitybib/bibliography
- http ://www.sfs.uni-tuebingen.de/ svajjala/research/readability-bibliography.html

The end

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

Plan

References I

Al-Khalifa, S. and Al-Ajlan, A. (2010).
Automatic readability measurements of the arabic text : An exploratory study.
35(2C).

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.

References II

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

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

Anula, A. (2007).
Tipos de textos, complejidad lingüıstica y facilicitación lectora.
In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

Armbruster, B. (1984).
The problem of "Inconsiderate text".
In Duffey, G., editor, *Compehension instruction : Perspectives and suggestions*, pages 202–217. Longman, New York.

Bahns, J. and Eldaw, M. (1993).
Should We Teach EFL Students Collocations ?
System, 21(1) :101–14.

References III

Beinborn, L., Zesch, T., and Gurevych, I. (2012).
Towards fine-grained readability measures for self-directed language learning.
In *Electronic Conference Proceedings*, volume 80, pages 11–19.

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

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

Brooke, J., Tsang, V., Jacob, D., Shein, F., and Hirst, G. (2012).
Building readability lexicons with unannotated corpora.
In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 33–39. Association for Computational Linguistics.

References IV

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.

Carrell, P. (1987).
Readability in ESL.
Reading in a Foreign Language, 4(1) :21–40.

Carroll, J., Davies, P., and Richman, B. (1971).
The American Heritage word frequency book.
Houghton Mifflin Boston.

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.

References V

Clark, C. (1981).
Assessing Comprehensibility : The PHAN System.
The Reading Teacher, 34(6) :670–675.

Coke, E. and Rothkopf, E. (1970).
Note on a simple algorithm for a computer-produced reading ease score.
Journal of Applied Psychology, 54(3) :208–210.

Coleman, M. and Liau, T. (1975).
A computer readability formula designed for machine scoring.
Journal of Applied Psychology, 60(2) :283–284.

Collins-Thompson, K. and Callan, J. (2004a).
A language modeling approach to predicting reading difficulty.
In *Proceedings of HLT/NAACL 2004*, pages 193–200, Boston, USA.

Collins-Thompson, K. and Callan, J. (2004b).
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*, pages 545–546.

References VI

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.

Constant, M. and Sigogne, A. (2011).
Mwu-aware part-of-speech tagging with a crf model and lexical resources.
In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 49–56.

References VII

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.

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

Dale, E. and O'Rourke, J. (1981).
The living word vocabulary : A national vocabulary inventory.
World Book-Childcraft International, Chicago.

References VIII

Dale, E. and Tyler, R. (1934).
A study of the factors influencing the difficulty of reading materials for adults of limited reading ability.
The Library Quarterly, 4 :384–412.

Danielson, W. and Bryan, S. (1963).
Computer automation of two readability formulas.
Journalism Quarterly, 40(2) :201–205.

Daoust, F., Laroche, L., and Ouellet, L. (1996).
SATO-CALIBRAGE : Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement.
Revue québécoise de linguistique, 25(1) :205–234.

Dascalu, M. (2014).
Readerbench (2)-individual assessment through reading strategies and textual complexity.
In *Analyzing Discourse and Text Complexity for Learning and Collaborating*, pages 161–188. Springer.

References IX

De Belder, J. and Moens, M.-F. (2010).
Text simplification for children.
In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., and Macken, L. (2014).
Using the crowd for readability prediction.
Natural Language Engineering, 20(3) :293–325.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011).
Read-it : Assessing readability of italian texts with a view to text simplification.
In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2014a).
Assessing document and sentence readability in less resourced languages and across textual genres.
International Journal of Applied Linguistics, 165(2) :163–193.

References X

Dell'Orletta, F., Wieling, M., Cimino, A., Venturi, G., and Montemagni, S. (2014b).
Assessing the readability of sentences : Which corpora and features ?
Proceedings of the 9th BEA Workshop, pages 163–173.

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.

Flesch, R. (1950).
Measuring the level of abstraction.
Journal of Applied Psychology, 34(6) :384–390.

References XI

Flor, M. and Klebanov, B. B. (2014).
Associative lexical cohesion as a factor in text complexity.
International Journal of Applied Linguistics, 165(2) :223–258.

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

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

François, T. (2009).
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.

References XII

François, T. (2011).
Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.
PhD thesis, Université Catholique de Louvain.
Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

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

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

François, T., Gala, N., Watrin, P., and Fairon, C. (2014b).
FLELex : a graded lexical resource for French foreign learners.
In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*.

References XIII

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)*.

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*.

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

Fry, E. (1968).
A readability formula that saves time.
Journal of reading, 11(7) :513–578.

References XIV

Fry, E. (1990).
A readability formula for short passages.
Journal of Reading, 33(8) :594–597.

Gala, N., François, T., and Fairon, C. (2013).
Towards a french lexicon with difficulty measures : Nlp helping to bridge the gap between traditional dictionaries and specialized lexicons.
In *Electronic lexicography in the 21st century : thinking outside the paper (eLex2013)*, pages 132–151.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004).
Coh-Metrix : Analysis of text on cohesion and language.
Behavior Research Methods, Instruments, & Computers, 36(2) :193–202.

Gray, W. and Leary, B. (1935).
What makes a book readable.
University of Chicago Press, Chicago : Illinois.

References XV

Gunning, R. (1952).
The technique of clear writing.
McGraw-Hill, New York.

Hancke, J., Vajjala, S., and Meurers, D. (2012).
Readability classification for german using lexical, syntactic, and morphological features.
In *Proceedings of COLING*, pages 1063–1080.

Harris, A. and Jacobson, M. (1979).
A framework for readability research : Moving beyond Herbert Spencer.
Journal of Reading, 22(5) :390–398.

Heilman, M. (2011).
Automatic factual question generation from text.
PhD thesis, Carnegie Mellon University.

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.

References XVI

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.

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

Hiebert, E. H. and Mesmer, H. A. (2013).
Upping the ante of text complexity in the common core state standards examining its potential impact on young readers.
Educational Researcher, 42(1) :44–51.

References XVII

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

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 XVIII

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

Kincaid, J., Fishburne, R., Rodgers, R., and Chissom, B. (1975).
Derivation of new readability formulas for navy enlisted personnel.
Technical report, number 8-75, Research Branch Report.

Kintsch, W. (1979).
On modeling comprehension.
Educational Psychologist, 14(1) :3–14.

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.

References XIX

Lee, D. Y. (2001).
Defining core vocabulary and tracking its distribution across spoken and written genres.
Journal of English Linguistics, 29(3) :250–278.

Lee, J. and Seneff, S. (2007).
Automatic generation of cloze items for prepositions.
In *INTERSPEECH*, pages 2173–2176.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004).
Manulex : A grade-level lexical database from French elementary-school readers.
Behavior Research Methods, Instruments and Computers, 36 :156–166.

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

Lorge, I. (1939).
Predicting reading difficulty of selections for children.
Elementary English Review, 16(6) :229–33.

References XX

Lorge, I. (1944).
Predicting readability.
the Teachers College Record, 45(6) :404–419.

Marujo, L., Lopes, J., Mamede, N. J., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J., and Viana, C. (2009).
Porting reap to european portuguese.
In *SLaTE*, pages 69–72. Citeseer.

McClusky, H. (1934).
A Quantitative Analysis of the Difficulty of Reading Materials.
The Journal of Educational Research, 28 :276–282.

McLaughlin, G. (1969).
SMOG grading : A new readability formula.
Journal of reading, 12(8) :639–646.

Meyer, B. (1982).
Reading research and the composition teacher : The importance of plans.
College composition and communication, 33(1) :37–49.

References XXI

Miller, J. and Kintsch, W. (1980).
Readability and recall of short prose passages : A theoretical analysis.
Journal of Experimental Psychology : Human Learning and Memory, 6(4) :335–354.

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.

Mitkov, R., An Ha, L., and Karamanis, N. (2006).
A computer-aided environment for generating multiple-choice test items.
Natural Language Engineering, 12(02) :177–194.

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

References XXII

Ojemann, R. (1934).
The reading ability of parents and factors associated with the reading difficulty of parent education materials.
University of Iowa Studies in Child Welfare, 8 :11–32.

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.

Pilán, I., Volodina, E., and Johansson, R. (2014).
Rule-based and machine learning approaches for second language sentence-level readability.
In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184.

References XXIII

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.

Selzer, J. (1981).
Readability is a four-letter word.
Journal of business communication, 18(4) :23–34.

References XXIV

Shardlow, M. (2013).
The cw corpus : A new resource for evaluating the identification of complex words.
ACL 2013, pages 69–77.

Sheehan, K. M., Flor, M., and Napolitano, D. (2013).
A two-stage approach for generating unbiased estimates of text complexity.
In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58.

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

Si, L. and Callan, J. (2001).
A statistical model for scientific readability.
In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 574–576. ACM New York, NY, USA.

References XXV

Skory, A. and Eskenazi, M. (2010).
Predicting cloze task quality for vocabulary training.
In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–56.

Smith, E. (1961).
Devereaux readability index.
The Journal of Educational Research, 54(8) :289–303.

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

Spaulding, S. (1956).
A Spanish readability formula.
Modern Language Journal, 40(8) :433–441.

Spivey, N. (1987).
Construing constructivism : : Reading research in the United States.
Poetics, 16(2) :169–192.

References XXVI

Štajner, S. and Saggion, H. (2013).
Readability indices for automatic evaluation of text simplification systems : A feasibility study for spanish.
In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 374–382, Nagoya, Japan.

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

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

Sung, Y.-T., Chen, J.-L., Cha, J.-H., Tseng, H.-C., Chang, T.-H., and Chang, K.-E. (2014).
Constructing and validating readability models : the method of integrating multilevel linguistic features with machine learning.
Behavior research methods, 47(2) :340–354.

References XXVII

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010).
Sorting texts by readability.
Computational Linguistics, 36(2) :203–227.

Taylor, W. (1953).
Cloze procedure : A new tool for measuring readability.
Journalism quarterly, 30(4) :415–433.

Thorndike, E. (1921).
Word knowledge in the elementary school.
The Teachers College Record, 22(4) :334–370.

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

References XXVIII

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

Vajjala, S. and Meurers, D. (2014a).
Assessing the relative reading level of sentence pairs for text simplification.
EACL 2014, pages 288–297.

Vajjala, S. and Meurers, D. (2014b).
Exploring measures of “readability” for spoken language : Analyzing linguistic features of subtitles to identify age-specific tv programs.
In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 21–29.

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

References XXIX

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

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.

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.

Vor der Brück, T. and Hartrumpf, S. (2007).
A semantically oriented readability checker for german.
In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274.

References XXX

Woodsend, K. and Lapata, M. (2011).
Learning to simplify sentences with quasi-synchronous grammar and integer programming.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420. Association for Computational Linguistics.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010).
A monolingual tree-based translation model for sentence simplification.
In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.