WHEN READABILITY MEETS COMPUTATIONAL LINGUISTICS: A NEW PARADIGM IN READABILITY

Thomas François

Publications linguistiques | « Revue française de linguistique appliquée »

2015/2 Vol. XX | pages 79 à 97
ISSN 1386-1204

Article disponible en ligne à l'adresse :


Distribution électronique Cairn.info pour Publications linguistiques.
© Publications linguistiques. Tous droits réservés pour tous pays.

La reproduction ou représentation de cet article, notamment par photocopie, n’est autorisée que dans les limites des conditions générales d’utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.
When readability meets computational linguistics: a new paradigm in readability

Thomas François, Université Catholique de Louvain

Abstract: Readability is an almost century-old field that aims to match readers with texts based on reproducible tools. It has a long history, but it is remarkable that the methodology used to coin readability formulas hardly evolved until recently, when traditional readability studies came into contact with the field of Natural Language Processing (NLP). In this paper, we first briefly retrace the milestones in readability research before the advent of this new paradigm, that we propose to call ‘Artificial Intelligence (AI) readability’ since it results from the conjunction of two AI-connected domains: NLP and machine learning. We then describe in more detail some of the important studies within this new paradigm, before discussing three main issues of the field that need to be resolved in order to make significant advances and present interesting perspectives for future research.

Résumé : La lisibilité est un domaine qui intéresse les chercheurs depuis près d'un siècle et qui vise à associer des lecteurs et des textes à l'aide d'outils reproductibles. Elle possède une longue histoire, dont l'une des caractéristiques le plus remarquables est la stabilité de la méthodologie utilisée pour créer une nouvelle formule. Celle-ci n'a guère évolué jusqu'il y a peu, lorsque l'approche traditionnelle s'est enrichie du contact du domaine du traitement automatique du langage (TAL). Dans cet article, nous retracions brièvement les étapes marquantes du domaine de la lisibilité antérieures à la naissance de ce nouveau paradigme que nous proposons d'appeler la 'lisibilité computationnelle' puisqu'elle est née de la conjonction de deux domaines associés à l'intelligence artificielle : le TAL et l'apprentissage automatisé. Nous décrivons ensuite des études importantes au sein de ce nouveau paradigme avant de discuter trois défis importants qui devraient, à notre sens, être résolus pour que des progrès majeurs soient effectués et qui constituent autant de pistes de recherches.

Keywords: AI readability, assessing reading difficulty of texts, NLP, psycholinguistic

Mots-clés : IA lisibilité, évaluation de la difficulté de lecture des textes, TAL, psycholinguistique

1. Introduction

Since Javal and his Essai sur la physiologie de la lecture (1879), a long tradition of studies has explained the process of reading, either in a first language (L1) or in a second or foreign language (L2). Although significant advances have been made, there are still some dimensions of the reading process that remain unclear. In particular, it is still not plain why one text is more readable than another, a property that is commonly referred to as text readability. Various recent psycholinguistic studies have investigated the impact of given textual characteristics (e.g. frequency of words, abstractness of concepts, type of syntactic structures, etc.) \(^1\) on comprehension, but no unified explanatory model of text readability has

---

\(^1\) See for instance the excellent review by Ferrand (2007) on this topic.

\(^2\) More details on the classic studies can be found in reviews by Chall (1958), Gilliland (1972), Harrisson (1980), Klare (1963, 1974, 1984), DuBay (2004) or François (2011). For another interesting review of
been produced so far. This is why, since the 1920s, parallel attempts were made to assess text readability from a more holistic, but also more pragmatic perspective: the goal was to select texts that fit the reading ability of a given audience. To this aim, several textual characteristics were considered and combined together within a statistical model, called a *readability formula*. Readability formulas arose in the context of the 1929 crisis, as they are tools that can support the teaching of reading – especially through the selection of suitable material – and facilitate access to information when used to control the difficulty level of documents, such as technical manuals or newspaper articles.

In our contemporary society, good reading skills are even more critical than they were in the past, as the amount of written information available has grown exponentially, especially as a result of the expansion of the web. Consequently, readability formulas, which suffered widespread criticism in the 1980s and 1990s, have recently regained some popularity, as witnessed by dedicated journal issues (François & Bernhard 2015), dedicated workshops such as PITR (Predicting and Improving Text Readability for Target Reader Populations), or dozens of recent publications focusing on various languages such as English, French, Portuguese, German, Swedish, Chinese, etc. This renewed interest goes along with methodological changes impelled by the contributions from two other disciplines: Natural Language Processing (NLP) and machine learning. François & Fairon (2012) have suggested calling this new paradigm *Artificial Intelligence (AI) readability*, and we have noticed that it comes along with a field switch, from education to NLP. As a result, there is sometimes a lack of continuity between recent AI readability studies and older readability studies.

In this paper, we aim to provide an introduction to the field of readability, focusing on the paradigm of AI readability. We start summarizing the broad lines of the readability field since its inception in the 1920s, with the goal of highlighting the roots of AI readability as well as stressing its specificities (see Section 2). In Section 3, we focus on recent studies representative of the AI readability paradigm, while Section 4 discusses some of the main issues of the field as we see it and suggests some perspectives for future research. Finally, Section 5 briefly illustrates AI readability formulas through a concrete example: the 'Amesure' project.

2. **Milestones in the history of readability**

2.1. **The origin of readability**

The aim of readability is to predict the reading difficulty of a text for a specific population, which may be schoolchildren, learners of a foreign language, readers with intellectual disabilities, etc. This objective is achieved through the study of the linguistic characteristics of the text. Reader variables (age, education level, ethnicity, etc.) and the context of reading (type of reading, goal of reading, time limitation, etc.), which have been brought to light by psycho-cognitive studies of the reading processes (see Boyer 1992), are generally considered as homogeneous within the population of interest. This assumption produces models that do not account for the reading process of an individual so well as psychological models such as those of Harm & Seidenberg (2004) or Coltheart & al. (2001). However, their all-encompassing and practical approach to reading, where individual differences are averaged, make them more useful for a classroom context.

---

The first readability formulas appeared in the United States in the 1920s. Various social evolutions, including compulsory education and the reduction of working time in industry led to an increase in the amount of uneducated readers for whom there were not enough readable texts available (Zakaluk & Samuels 1996). To address these needs, some scholars began to develop techniques to match texts and readers. Lively & Pressley (1923) designed what is generally considered to be the first readability formula: it predicted the difficulty of a text considering only its vocabulary, relying on a list of lexical frequencies compiled by Thorndike (1921).

Later, Vogel & Washburne (1928) developed sounder methodological guidelines to develop a readability formula, which remained mostly unchanged until the early 21st century. The first step is to collect a corpus of texts whose difficulty has been previously assessed for a given population using a criterion. A criterion is the measurement technique used to estimate the average understanding of each text by the target population. These techniques can be conventional comprehension tests, reading time, the average judgement of a group of experts, etc. (see François 2014). The validity of the techniques, that is to say the fact that they measure what is intended to be measured, is obviously crucial.

The second step is to identify linguistic characteristics of texts that affect reading comprehension or, at least, are good predictors of it. The efficiency of a single variable is usually measured through its correlation with the criterion representing the difficulty of texts. If one of those factors is sufficiently correlated with the difficulty, it may be retained in the readability formula. This is how Vogel & Washburne (1928) selected two lexical predictors (the number of different words and the number of words missing from Thorndike’s list) and two syntactic ones (the number of prepositions and the number of simple clauses) for their formula.

In a third step, those various predictors \((X_1, X_2, ..., X_n)\) need to be combined within a formula that outputs, for any input text, a single index of its difficulty. Vogel & Washburne (1928) proposed a multiple linear regression equation as follows:

\[
\text{Criterion} = B_0 + B_1 X_1 + B_2 X_2 + ... + B_n X_n
\]

The readability of a text, couched in these mathematical terms, becomes easy to handle. The coefficients \(B_1, [\ldots], B_n\), which represent the relative weight of each linguistic feature, are automatically inferred from the corpus using the least squares method.

This methodology was reused and enhanced in several subsequent studies on English. Dale and Tyler (1934) used multiple choice comprehension tests as a criterion to calibrate the texts in their corpus. McClusky (1934) investigated reading speed as an alternate criterion, while Ojemann (1934) was the first to take into account factors beyond lexical and syntactic levels, namely the density of ideas. Finally, Gray and Leary (1935) discussed more than 289 factors, organized them into four categories (content, style, format, and organization), and devised a formula including only seven predictors.

### 2.2. The studies before ‘AI readability’

After these first studies, comes the period usually called ‘the classical period’ (Klare 1963). Studies from the classical period are characterized by a drive for simplicity and efficiency. The lack of computerized procedures for counting variables made the application of complex formulas with numerous features tedious business. In addition, researchers realized that the different predictors were intercorrelated and that two or three of them were usually enough to obtain a reliable formula. This trend was embodied by two of the most famous formulas of the classical period: the Reading Ease (Flesch 1948) formula, intended for adults, and the Dale & Chall’s formula (1948), intended for primary and secondary schoolchildren. Both were based
on two predictors: a syntactic one (the average number of words per sentence) and a lexical one (the average number of syllables per 100 words for Flesch; the number of words not in Dale’s list, which included 3000 lexemes, for the latter). Both studies inspired a crowd of followers, but the methodology remained the same: most formulas included two variables, either lexical or syntactic, and made use of the same corpus for training, the McCall-Crabbs Standard Lessons in Reading.

In the 1960s, the cloze test\(^3\) stimulated a renewal of readability experimentation after Coleman (1965) used it as a new type of criterion. It was considered as a more valid criterion than comprehension tests and indeed led to more efficient formulas. For instance, while the multiple correlation coefficients (R) obtained by the formulas of Flesch or Dale & Chall did not exceed 0.7, Coleman (1965) reported values of between 0.86 and 0.9, and Bormuth (1969) even reached 0.928. The formulas of this period, sometimes called ‘the revolution of the cloze’, included more predictors than classic formulas. This was stimulated by the use of computers, which enabled researchers to automatize the counting of some variables and the training of the statistical models.

Unfortunately, automation did not bring only progress. The automation of readability predictors led researchers to opt for characteristics whose counting rules could be explicitly rendered by a machine (e.g. counting the number of letters), which resulted into a more superficial view of text difficulty. As a result, and also under the influence of advances in cognitive psychology about the reading process, increasingly sharp criticism was directed at readability formulas (Selzer 1981; Duffy 1985; Stevens 1980; Kintsch & Vipond 1979; Kemper 1983). The main criticisms focused on (1) the failure of classic formulas to consider textual aspects such as cohesion, ideas coherence, macro-structural organization, etc.; (2) the fact that readability formulas ignored the interactive dimension of the reading process; or (3) the trend among readability formula users to overgeneralize these tools to populations different from that for which they were designed.

A new approach of readability also originated from this criticism and experimented with new predictors measuring complex psycholinguistic dimensions such as the inference load of a text (Kintsch 1979; Kemper 1983), the conceptual density (Kintsch & Vipond 1979), or organizational aspects (Meyer 1982). However, this approach suffers from two major limitations: complexity and efficiency. As an example, Kintsch and Vipond (1979) devised a formula that achieved excellent results (R = 0.97), but was based on very complex variables, which required a large amount of manual effort during parametrization, as it was not possible to automate it. Kemper (1983) then used a simplified version of the same variables, thus losing some performance (R = 0.76). The second issue was that the results of this much more complex approach were not superior to those of classical formulas, although the latter used much simpler predictors. Kemper (1983) tested whether, at least, it brought a different kind of information, so that combining structuro-cognitivist features with classic variables would be more effective. Unfortunately, such a combination led to minimal improvement in performance (R = 0.78 compared to 0.76).

The most notable formulas of the 1990s relied on the traditional approach: the new Chall & Dale (1995) formula is an update of their previous one; Stenner & al. (1988) and the company Metametrics proposed the Lexile Framework, while researchers from the Renaissance Institute and Touchstone Applied Science Associates designed the Advantage-TASA Open Standard, better known as the ATOS formula (Paul 2003). Although they led to efficient

\(^3\) Developed by Taylor (1953), it is a ‘reading’ exercise that involves in filling in text blanks, which usually appear systematically. The amount of blanks filled by a reader is supposed to be indicative of its understanding of the text.
tools, they did not answer the concerns of the 1980s, keeping shallow features and overlooking higher dimensions of texts.

3. The AI readability

3.1. Foundations

Work on readability had reached the point where torn between the desire to avoid the structuro-cognitivist criticisms and the need to design simple and efficient models, it was no longer the clear, unified field it was in the past. However, the advent of ‘AI readability’ was about to change the situation. In our view, the roots of ‘AI readability’ can be traced back to three pioneering studies. The first is the system of Daoust & al. (1996), called SATO-calibrag, that outputs a difficulty level on a continuum from the first to the fifth secondary school grade. It was maybe the first readability system to rely on a NLP software – called SATO –, to automatically extract 120 linguistic variables. The authors then formulated a multiple linear regression equation including as many as 14 variables, which was able to reach an $R$ of 0.86.

Shortly afterwards, Foltz & al. (1998) investigated a new metric of the lexical coherence of a text based on latent semantic analysis (LSA). Introduced by Landauer & al. (1998), this technique consists of creating a semantic space from large corpora, in which each sentence or paragraph of a text can be located. Sentences situated in the same region in that space are likely to expose related ideas. This property can be used to assess the global coherence of a text as the average cosine similarity of all pairs of adjacent sentences. Foltz & al. (1998) showed that this new measure of lexical coherence was correlated with measures of reading comprehension from poor readers.

A third study was carried out by Si & Callan (2001). The originality of this study involves rephrasing the challenge of predicting text readability as a classification problem, and thus broadening the set of techniques that can be applied. According to Si & Callan (2001, 575), assessing the difficulty of a document $d_i$ amounts to assigning it the class $g$ that maximizes the following probability:

$$P_c = \frac{P_a(g|d_i) + (1 - \lambda)P_b(g|d_i)}{P_a + (1 - \lambda)P_b}$$

$P_c$ is a linear combination of two probabilities: $P_a$ is output by a unigram language model while $P_b$ corresponds to a model of sentence length distribution based on Gaussians. The specificity of their model lies in the language model ($P_a$) that postulates that there are significant differences in the use of a word across levels.

The three aforementioned studies helped to frame a new avenue for research, that we propose to call the ‘AI readability’ paradigm. It does not completely differ from the traditional methodology: training a readability model still requires starting from a labelled corpus, defining a set of variables (now also called features) and then selecting and combining them within a statistical model. However, using NLP helps to design and automatize more linguistically-motivated features, such as those suggested by the structuro-cognitivist paradigm. At the same time, considering readability as a classification problem helps to broaden the set of algorithms that can be used to combine the features. As a result, more sophisticated models can be developed that include far more variables than the classical formulas and use algorithms better able to cope with the linguistic information encoded in those features.

From these first attempts, dozens of subsequent studies have investigated the field of AI readability and it is still difficult to organize this new body of work in significant trends, as
we lack perspective. In this paper, we suggest classifying recent papers into three main approaches, based on the type of features they use. The first, which draws on Si & Callan (2001) and predominates until 2008, focuses on lexical and syntactic features. From 2007, a second trend appears with Crossley & al. (2007), which attempts to combine the first trend with the lessons learned from the structuro-cognitivist period. All dimensions of a text – lexical, syntactic, semantic, discourse, and even pragmatic aspects – are considered in these models. In parallel, a third trend, embodied by Tanaka-Ishii & al. (2010), departs from linguistically-motivated features and tries to look for less-informed models. In the following sections, we discuss in more detail these three trends.

3.2. Focus on lexicon and syntax

The work of Si & Callan (2001) was continued by Collins-Thompson & Callan (2004), who focused on the lexical components of the model and enhanced it to make it capable of analyzing short fragments, in particular web pages. The principle is as follows: based on 550 documents from the web, whose difficulty level varies from 1st grade to 12th grade, they trained 12 language models. Each model represents the multinomial distribution of a vocabulary V within a class Gi, i.e. a grade level. Each word wj in V therefore has a probability P(wj | Gi) and classifying a text amounts to be selecting the class Gi maximizing the following quantity:

$$L(G_i | T) = \sum_{w \in T} C(w) \log P(w | G_i)$$

A specificity of the model is that the probability P(wj | Gi) is smoothed depending upon the frequency of the word wj in adjacent levels. According to the authors, this feature greatly improved the accuracy of the model compared to that of Si & Callan (2001), and also turned out to be more efficient for short passages. They reached an R of 0.79 for English and of 0.64 for French on their test corpus.

Another interesting study from this period is presented by Schwarm and Ostendorf (2005). They brought several innovations to the field of readability. Firstly, they were pioneers in the use of the Weekly Reader as a readability corpus. The Weekly Reader is an educational newspaper with articles tailored for four different grades (from 2nd to 5th grade) that has since been re-used in several studies and is sometimes regarded as a reference corpus. As regards the variables, the authors suggested using a syntactic parser to define parse-based features as measures of the syntactic complexity. They also applied n-grams models, common in NLP, to evaluate the likelihood of sentences. They were also first to combine all their lexico-syntactic features with a support vector model (SVM), an algorithm that has proven very successful in readability since. The resulting model compared favourably with the formulas of Kincaid & al. (1975) and Stenner & al. (1988) on the test corpus. Finally, the authors introduced a new evaluation measure for readability: adjacent accuracy. It corresponds to the percentage of texts correctly classified by the model plus the percentage of texts misclassified only by one level (for an ordinal or discrete scale). Its use was justified by the fact that even humans struggle to agree on the difficulty of a set of texts. However, it should be noted that it is an optimistic measure, especially when the number of classes in the model is small. A surprising finding of Schwarm and Ostendorf (2005) was that syntactic variables contribute little to the overall performance of their model. Heilman & al. (2007) questioned this result, comparing the contribution of syntactic predictors both in a L1 and in a L2 context. They noticed that syntactic variables were more effective for L2 texts, which is consistent with what we know about the reading process in L2.
3.3. Towards all-encompassing models

In parallel with the aforementioned studies, other researchers, closer to cognitive sciences, extended the notion of readability to include more than the lexico-syntactic properties of a text. Graesser & al. (2004) opened the path in this regard with the software Coh-Metrix, able to automatically parametrize a large panel of new text dimensions among which the age of acquisition of words, the polysemous status of a word, word imageability, and various measures of text cohesion and coherence⁴. Later, Crossley & al. (2007) combined three of these dimensions (word frequency, mean number of words per sentence, and argument overlap, i.e. the number of sentence pairs sharing a noun with the same stem) in a classical readability formula based on linear regression. The three variables explain 91% of the corpus variance in terms of difficulty. However, the addition of the cohesive factor, presented by the authors as a major breakthrough and a way to respond to the criticism of the 1980s, is not convincing. On the one hand, because Chall & Dale's (1995) formula does very well without it, reaching similar performance on the same corpus; on the other hand, because this factor is not significant in the regression model (p = 0.062).

Pitler & Nenkova (2008) further investigated the issue of semantic and discourse features. They selected 30 articles from the Penn Discourse Treebank (Prasad & al. 2008), in which the discourse relations between sentences were annotated, and asked three students to rate the difficulty level of those texts on a scale ranging from 1 to 5. They then implemented 32 variables, including the unigram model of Collins-Thompson & Callan (2004), the tree-based syntactic features of Schwarm & Ostendorf (2005) and several new predictors capturing the discursive relations and the level of cohesion of the texts (e.g. using the lexical chains). The most efficient predictors were the unigram model \((r = 0.4497 \text{ with text difficulty})\), the mean number of verbal phrases per sentence \((r = 0.4213)\) and the log-likelihood of a text with regards to its discursive relations \((r = 0.4835)\). This finding showed that higher level textual dimensions, automatically parametrized, could significantly contribute to the prediction of text readability. However, most of the other discourse and semantic factors were not significantly correlated with difficulty.

Shortly afterwards, Feng & al. (2009) also proposed an integrated readability model, intended for readers with intellectual disabilities, and combining lexical, syntactic and cognitively-motivated features. They implemented similar variables to those of Pitler and Nenkova (2008), but found a much stronger effect for the cognitive predictors. In a later study, Feng & al. (2010) investigated more systematically the various types of features introduced in the ‘AI readability’ paradigm. They distinguished different classes of discourse (including those based on the density of entities, on various characteristics of lexical chains and their entities, and on co-reference inferences), language model-based, syntactic parse tree-based, POS-based and shallow features. They compared the performance of each class of features both with a SVM and logistic regression, using the Weekly Reader corpus as the ‘gold-standard’. Most discourse features, except for entity-density ones, behaved poorly, as was also the case for the syntactic parse tree-based predictors. However, language models and POS-based variables led to higher classification accuracy. The authors conclude: “discourse features do not seem to be very useful in building an accurate readability metric” (Feng & al. 2010, 283). Todirascu & al. (2013) argued that these mixed results might be due to approximations of the NLP systems, since automatically annotating discourse features remains a challenge. They manually annotated lexical chains in 20 texts for French as a foreign language (FFL), whose level ranged from A2 to B2 level on the CEFR scale. The authors correlated various characteristics of lexical chains with the difficulty of these texts

---

⁴ Refer to <http://cohmetrix.com/> for a complete list of implemented features.
and showed that considering the type of entities (e.g. the proportion of pronouns or indefinite NP being an entity in a chain), and not only their syntactic transitions, can be important. However, only four features appeared to be significantly correlated with difficulty, possibly due to the limited size of their corpus.

Besides these investigations, Vajjala & Meurers (2012) experimented with another set of variables, derived from the field of second language acquisition (SLA). Normally intended to capture the development of linguistic complexity in the learners’ productions, these variables proved very useful for the readability prediction task as well. Vajjala & Meurers, using the Weekly Reader corpus, which they expanded with materials of higher grade level, were able to reach a very high classification accuracy (93.3%). SLA features, such as the lexical density or measures of lexical variation (i.e. the ratio of a given POS category on all lexical words in a document), proved just as efficient (accuracy of 82.3%) as other lexical variables previously used in ‘AI readability’ (82.4%).

3.4. Non expert models

In the long tradition of readability, one of the stable aspects of the methodology is the habit of defining variables that capture linguistic characteristics of texts which are somehow related to text difficulty. Such an operation not only requires some expertise (since linguistic knowledge is used during parametrization), but also amounts to a kind of informational filter applied to the raw material of the text. This filter has sometimes been very simplistic, especially when text difficulty was approached only with shallow features, such as the number of letters in words or the number of words in sentences. This is why replacing feature engineering by algorithms meant being able to start from words in the text and automatically extract features offered a promising avenue for research in the future. The only study, to the best of our knowledge, to investigate such an approach is Tanaka-Ishii & al. (2010). They only used word frequencies as input features for SVM models, but suggested a very interesting idea, namely to train a binary model to act as a sorting function. This way the model is able to sort any collection of texts by difficulty with a high level of accuracy. Moreover, the corpus needed to train the model is much simpler to collect. The main shortcoming of their approach is that it is hardly able to capture more than lexical difficulty. However, we believe that, with the advent of neural networks and deep learning in the NLP community, these methods could be used to capture a larger set of textual dimensions automatically.

4. Challenges and perspectives in readability

In this last section of the paper, we discuss certain issues or perspectives that we deem interesting or crucial for the future of readability. The first one is the question of the corpus and its annotation. We will go on to argue in favour of specialization of the formulas, before concluding this section with attempts to move from text readability to sentence and even word readability.

4.1. The issue of measuring text difficulty

One of the shortcomings of the AI readability approach, compared to the classical one, is that it requires much more labelled data to train the models, since more complex statistical models mean more parameters to train. Whereas it was previously possible to calibrate a limited amount of texts on a sample of the population (for instance using a cloze test), this criterion has now been abandoned in favour of a more convenient one: collecting texts that have already been assigned a grade level by experts. Researchers thus used the Weekly Reader
corpus (Schwarm & Ostendorf 2005; Feng & al. 2010; Vajjala & al. 2012, etc.), a collection of texts from textbooks (François 2009; Volodina & al. 2013), or two sets of documents whose difficulty is highly contrasted (Dell’Orletta & al. 2011).

However, there are serious concerns about the validity of such corpora, as they result from the judgement of a small set of experts. François (2014) recently published a homogeneity analysis of a textbook-based readability corpus carried out with ANOVA tests and concluded that texts assigned to the same level by experts may display significant variation in lexical and syntactic difficulty. This finding is consistent with experiments carried out by van Osten & al. (2011) with human experts. These authors created pairs of texts from a range of 105 documents and asked a panel of experts to select the most difficult text in each pair. In a second step, experts were automatically clustered according to similarity of their judgements and several corpora were thus created, each corresponding to a cluster of experts. Finally, a readability model was trained on each of these corpora and their performance was assessed both on the texts from the same corpus (intracluster) and texts from the other clusters (intercluster). It appeared that the intracluster performance was higher than the intercluster one, a situation which is likely to happen when one generalizes a readability formula based on expert judgements on new texts.

To further investigate this issue, we carried out some additional experiments with two corpora of FFL texts, supposedly similar. We started with François & Fairon’s (2012) corpus, whose difficulty ranges from A1 to C2 levels. We randomly selected 68 texts per level and trained a SVM model with 41 of the 46 variables used in François & Fairon’s (2012) model5. It reached 48% of accuracy and 78% of adjacent accuracy, which is very similar to the results reported in the paper. We then collected another corpus of FFL texts, based on simplified readers. As the simplified readers were commercialized by the same publishers as the textbooks used in the first corpus, we were expecting the two corpora to be similar. However, when we applied the model to this new corpus, we witnessed a significant and unexpected drop in accuracy, down to 38%, although adjacent accuracy slightly increased (81.7%). This means that, when we apply our readability formula to new texts supposedly similar to those in the training corpus, as we would do in a real applicative context of this model, the predictions of the model becomes fuzzier, although it does not make more critical mistakes. One of the explanations of this phenomenon is that most texts in the second corpus were narrative, whereas various types of texts were represented in the first (see Section 4.2 for the influence of genre).

As a second experiment, we re-trained another SVM model (with only 38 variables) on the second corpus and assessed its performance with a 10-fold cross-validation process, as was done for the first model. This new formula was able to reach much better accuracy (58.2%) and adjacent accuracy (87.7%) levels than the first model. Consequently, it seems that it is difficult to identify the specific characteristics of texts and the population that are modelled by a readability formula; this may lead to the risk of applying the formula to texts or for readers with different characteristics. The literature review also highlighted two other issues of current readability corpora: (1) expert judgement (or use of a text previously classified by experts) predominates as a criterion in ‘AI readability’ although no experimental data confirm that it is the best proxy for text difficulty; (2) on the contrary, it seems difficult to ensure the validity of such annotation as even human experts have trouble agreeing on text readability.

5 The missing variables were either not available for the new corpus or appeared to be consistent with it (e.g. the variable ‘presence of an infinitive verb’ always is 1 in the second corpus).
4.2. Specializing the formulas

Another issue related to the previous one is whether or not we should specialize the formulas according to a specific public. Traditional formulas were generally seen as universal tools that could be applied to any context. For instance, Flesch claimed his formula was suitable for all adults, although he trained it on the McCall & Crabb's lessons, intended for schoolchildren. However, some researchers began fairly early on to design specialized formulas, either for specific populations, such as young children (Spache 1953; Harris & Jacobson 1973), second language learners (Tharp 1939), soldiers (Kincaid & al. 1975), or for specific types of texts, such as standardized tests (Forbes & Cottle 1953), scientific texts (Jacobson 1965), or technical documents (Hull 1979, cited by Klare 1984).

Despite all these specialized formulas, very few experiments have tried to determine whether it is best to use a specialized model instead of a more generic one. In the context of English as a L2, Brown (1998) administered cloze test versions of 50 texts to 2,300 Japanese learners of English and correlated the results of the test with six classical formulas. The correlations being low (between 0.48 and 0.55), he tended to conclude against the validity of these classical formulas for L2, especially as he was able to obtain an R of 0.74 with a specialized formula he retrained on his corpus. These results were later contradicted by Greenfield (2004) who also compared the results of 5 classical formulas with cloze scores of Japanese learners of English and got much higher values on a different corpus. Moreover, because he used 30 of the Bormuth's passages as his corpus, for which the cloze scores of native readers were available, he was able to correlate native and L2 readers' scores and obtained a correlation of 0.915. He therefore concluded that the L1 formulas may be appropriate for L2 readers too, but replicating such studies in different contexts is necessary to obtain any conclusive findings.

As regards text adaptation, more conclusive data have been gathered. Nelson & al. (2012) carried out an extensive evaluation of several readability formulas on narrative and informative texts. They observed a large difference of performance for most models on both genres of texts, the best correlation being obtained on informative texts. Sheehan & al. (2013) analysed in further detail the differences between literary and informative texts and noticed that readability formulas tend to overestimate informative text difficulty and underestimate it for literary texts. This is due to the fact that literary texts include more core vocabulary of the language, while informative documents contain more content area, which “often received inflated readability scores since key concepts that are rare are often repeated, which increases vocabulary load”. Other studies (Dell'Orelta & al. 2014; Vajjala & Meurers 2014) confirmed the effect of genre on readability prediction performance. We therefore believe that there is a need for specialized formulas, both for specific text genres and for specific reader populations. The main reason preventing more specialized models from being built is, in our view, the lack of suitable corpora. There is clearly a need to develop a method able to collect difficulty annotations that are both valid and inexpensive for a target population. De Clerq & al. (2014) suggested using crowdsourcing to this aim, arguing that a large amount of non-expert annotators can provide labels as reliably as experts, but more work is needed to confirm the value of such criterion for readability.

4.3. From text to words

Traditionally, readability aims to assess difficulty at the text level. It is common knowledge in the field that formulas should not be applied to fragments smaller than 100 words. However, there are several contexts in which this issue needs to be overcome such as web page readability (Collins-Thompson & Callan 2004), automatic exercise generation (Pilán & al.
Readability and computational linguistics: a new paradigm in readability

2014), or the use of readability models to evaluate or inform a text simplification system (Vajjala & Meurers 2014). The first to investigate sentence and word readability was Bormuth (1966), taking advantage of the properties of cloze tests. He designed a sentence model that reached an R of 0.665, while the word-level formula only obtained 0.505 (vs 0.934 at document level). Later, Fry (1990) designed a traditional formula adapted for short passages, giving more weight to the lexical variable, a strategy that was reused by Collins-Thompson & Callan (2004). More all-encompassing approaches of sentence readability were undertaken by Dell'Orletta (2011) that combined lexical and syntactic features in a SVM. Similarly to Bormuth, they obtained a better accuracy at document level (98% on a binary dataset) than at sentence level (78%). That assessing sentence readability is even more challenging than at document level was confirmed by Pilán & al. (2014), who found an accuracy of 71% on a binary task. Vajjala & Meurers (2014), who used SLA features at the sentence level, also got a low level of accuracy (66%), which they believe was due to the heterogeneity of their training corpus. Although we do not know of any experiment comparing the validity of difficulty annotation at the text and sentence level, this seems to be an interesting issue to investigate as an explanation for the superiority of text-level readability models.

At the word level, the situation is even more problematic, as shown by Bormuth's results. Gala & al. (2013) obtained a similar picture. Using two lists of simple words ‘classified’ respectively into three levels (Lété & al. 2004) and six levels (François & al. 2014), they trained a SVM model with 49 features, such as the number of letters or syllables, word frequency, polysemic status of words, and including some original features based morphemic information. However, they only obtained a small gain (+2%) compared to a simple frequency baseline. Baeza-Yates & al. (2015) tried a similar approach on Spanish, using a gold-standard list of simple and difficulty words to train a sequential minimal optimization (SMO) algorithm. They reached 72% accuracy on this binary task and observed that their two best predictors were the number of orthographical neighbours (d1) and word length.

5. ‘AMesure’: illustrating AI readability formulas

Having discussed the major studies of AI readability and some of the challenges raised by this new paradigm, we conclude this paper with a brief illustration of the creation and use of an AI readability formula, drawn from our own research. This readability model is called ‘AMesure’ as it aims to assess the reading difficulty of administrative texts. Administrative texts are acknowledged to be a challenge for a significant amount of the population, which is all the more preoccupying since they are related to essential pragmatic issues (e.g. getting unemployment benefit, paying taxes, creating a company, etc.). ‘AMesure’ is therefore in line with the numerous initiatives taken by administrations around the world to simplify the documents they publish. Compared to most readability models reported in the literature, it has the advantage of being freely accessible on a web platform6, making it a good candidate for our illustration. This section will first detail the design of the readability model available on the ‘AMesure’ platform, and then will describe the current and future functionalities of the platform itself.

In the previous sections, we mentioned that most current AI readability formulas have been trained on a corpus of graded texts extracted from different educative sources. Unfortunately, such a corpus does not exist for administrative texts and, in accordance with our above considerations about the need to specialize the formulas, we did not want to use educational materials for training. This means that we had both to define a difficulty scale

6 <http://cental.uclouvain.be/amasure/>. The ‘AMesure’ platform was developed with the support of the Fédération Wallonie-Bruxelles.
and then annotate texts according to this scale. To this aim, we applied an annotation protocol in two steps.

First, we collected 115 authentic administrative texts from the Fédération Wallonie-Bruxelles administration, dealing with eight different themes (e.g. sport, culture, education, etc.). We then produced a rough estimate of the reading difficulty of the 115 texts, using a classic readability formula for French by Kandel and Moles (1958), and sampled ten texts with varied difficulty in order to have them read by ten subjects on a self-paced reading interface. The interface recorded the subjects’ reading time per sentence. After the subjects had finished reading, they were given two multiple-choice questions to make sure that they actually read and understood the text. Reading times were processed with a mixed-effect model (Baayen & al. 2008) in order to discard the subject-specific variability in reading times and to obtain a ms./word average for each texts. The reading times obtained for the ten texts are displayed in Table 1, along with their Kandel and Moles scores. It should be mentioned that the Pearson correlation between both readability measures was quite high ($r = 0.74$).

Based on the reading times, we defined five different difficulty levels (see column ‘Level’ in Table 1) and distributed two texts in each level. The selected scale ranged from 1, for ‘very easy’ texts, to 5 for ‘very complex’ texts (cf. appendix).

Table 1. Ranking of the 10 tested texts according to reading times.

<table>
<thead>
<tr>
<th>Title of the text</th>
<th>KM Score</th>
<th>ms./word</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>La santé de votre enfant</td>
<td>71.3</td>
<td>292.8</td>
<td>1</td>
</tr>
<tr>
<td>Du couple à la famille (se séparer …)</td>
<td>86.5</td>
<td>304.9</td>
<td>1</td>
</tr>
<tr>
<td>Des chaussures… Quand les mettre aux pieds ?</td>
<td>81.1</td>
<td>315</td>
<td>2</td>
</tr>
<tr>
<td>A l’école d’une alimentation saine</td>
<td>75.8</td>
<td>324.4</td>
<td>2</td>
</tr>
<tr>
<td>L’enseignement spécialisé</td>
<td>46.2</td>
<td>339.7</td>
<td>3</td>
</tr>
<tr>
<td>Lettre pour la semaine européenne de la vaccination</td>
<td>40.6</td>
<td>340.5</td>
<td>3</td>
</tr>
<tr>
<td>Cumuls de pensions</td>
<td>57.5</td>
<td>372.3</td>
<td>4</td>
</tr>
<tr>
<td>Liquidation des subventions ordinaires 2004</td>
<td>15</td>
<td>376.6</td>
<td>4</td>
</tr>
<tr>
<td>Déclaration de succession</td>
<td>57</td>
<td>379</td>
<td>5</td>
</tr>
<tr>
<td>Tax shelter</td>
<td>36.5</td>
<td>390</td>
<td>5</td>
</tr>
</tbody>
</table>

In the second step, an annotation guide was developed that includes five of the ten tested texts to serve as reference examples. The guide was used by experts from the FWB administration to manually annotate the 105 remaining texts according to our 5-level scale. Not all texts were seen by each expert; instead we got an average of 2.5 judgements per texts. The inter-rater agreement, assessed with Krippendorf’s (1980) alpha, was not very high (0.37), but we assumed that averaging judgements would help to get more reliable annotations.

Once the corpus was annotated, we were able to apply NLP-routines to extract as much as 344 readability features belonging to three main categories: lexical (e.g. word frequency, lexical diversity, type of orthographical neighbours, length of words, etc.); syntactic (e.g. length of the sentences in words, ratios of POS, verbal tenses and moods used in the text, etc.), and semantic (density of the ideas, cohesion of the text, personalization level, etc.). The efficiency of each of the 344 variables was evaluated as regards its Spearman correlation with the difficulty of the texts and the 10 best predictors were selected in the final model. The ‘AMesure’ readability model is then based on a support-vector machine (SVM) algorithm.

7 The variables used and their implementation details are the same as those reported in François (2011), except that we excluded variables specific to French as a foreign language.

© Publications linguistiques | Téléchargé le 18/03/2021 sur www.cairn.info (IP: 54.70.40.11)
using L2 regulation and a linear kernel. Its accuracy was estimated by a 10-fold cross-validation procedure to reach 58%, which is slightly better than the 50% accuracy obtained by the model of François and Faron (2012). However, the amount of training data being quite limited, serious concerns about the generalizability of the ‘AMesure’ model can be raised.

![Image](image.png)

**Figure 1. Example of results from the ‘AMesure’ platform.**

This readability formula has been integrated into the ‘AMesure’ platform, which also offers, besides this global difficulty index, two types of diagnosis about a given text complexity. The first consists of a more precise analysis of text readability addressing five dimensions of the texts, namely their lexical difficulty, their sentence length, a rough indication of their syntactic complexity, their level of personalization, and their cohesion level. The second diagnosis, which is still being worked on, aims to detect the specific lexical and syntactic forms that can be challenging for the reader. Currently, the system detects difficult words based on their frequencies as well as three types of subordinated clauses. An example of a short text analysis is presented at Figure 1. Future work on ‘AMesure’ will focus on detecting technical terms specific to the administrative or legal language, increasing the amount of complex structures covered, and possibly offering simplification advice.

### 6. Conclusion

In conclusion, it is important to stress once more the challenges that await the researcher within the field of readability. A large number of formulas have been designed for different uses (e.g. assessing the readability of textbooks, newspapers, administrative documents or even technical manuals in the army), for different populations (e.g. schoolchildren, L2 learners, people with language disabilities), and for different languages (see below). Some of them were very successful, in particular the Flesch formula, which was applied to control the difficulty of mainstream newspapers in the U.S. and later included within some word processors. Despite that, the field seems not to have been able to completely unravel the mysterious nature of text difficulty. We have learned a good deal about the variables that impact the reading process, but we still do not know how they relate to each other and a lot of questions remain.
Nevertheless, it is clear that the advent of AI readability has revived the field, offering new avenues for research, new algorithms to parametrize more complex text dimensions or to combine more efficiently linguistic characteristics into a readability model. We believe that there is still much to be done to improve current models as well as to better understand the nature of readability. To this aim, we believe less attention should be aimed at improving performance of models trained on a specific corpus and intended for a specific population and instead more attention should be paid to finding the answers to certain transversal questions or replicating the lessons learned so far on different corpora and for different languages. For instance, although semantic and discourse predictors appear very appealing, they do not seem to make much difference in some cases, or might even be redundant with more basic lexical variables. Another major challenge would be to design an efficient and valid, but simple and inexpensive criterion that could open the path to more accurate and, more importantly, more psychologically-valid formulas. Investigating sentence and word readability is also a promising area of research as is the development of less-informed models. There is at least one certainty: There is at least one certainty: readability is not a solved problem yet, but it is a one we should attempt to solve nonetheless.

Dr. Thomas François  
FNRS-CENTAL/IL&C  
Université Catholique de Louvain  
Place Blaise Pascal, 1 bte L3.03.12  
B-1348 Louvain-la-Neuve, Belgium  
+32 10 47 37 89  
thomas.francois@uclouvain.be

Acknowledgments: Thomas François is supported by the Belgian Fund for Scientific Research (FNRS). We would like to thank Serge Bibauw and Jean-Léon Bouraoui for their valuable comments during the writing process.

References


Readability and computational linguistics: a new paradigm in readability


Gray, W.S., Leary A.B. (1935),–


Readability and computational linguistics: a new paradigm in readability


Appendix

This appendix provides 5 excerpts from the 10 tested texts in the 'AMesure' project and aimed to be representative of the five difficulty levels defined.

Niveau 1 : La santé de votre enfant : les programmes de prévention
Pour avoir de bonnes dents...
Il est conseillé de prendre l'habitude de brosser les premières dents dès leur apparition. Mais surtout, pour réduire l'apparition des caries, il faut éviter de donner des biberons contenant des boissons sucrées, les tétines trempées dans du miel ou de la confiture et limiter les sucreries.
Le premier examen chez le dentiste est conseillé vers l'âge de 3 ans. Ensuite, deux visites préventives sont recommandées annuellement.
Les enfants âgés de moins de 12 ans bénéficient de la gratuité des soins dentaires, à l'exception de l'orthodontie (la correction des défaux de la dentition). Toutefois, vérifiez si votre dentiste applique bien cette gratuité. N'hésitez pas à lui poser la question !

Niveau 2 : Des chaussures confortables jour après jour.
Les pieds de votre enfant grandissent à un rythme soutenu et la voûte plantaire se développe progressivement jusqu'à l'âge de 6 ans. L'enfant aura donc besoin de 2 à 3 paires de chaussures successives la première année de marche.
Les pieds transpirent moins dans des chaussettes en coton. Si l’enfant transpire beaucoup, il est recommandé de laisser aérer les chaussures en alternant un jour sur deux avec une autre paire de même taille.
Les chaussures de fantaisie comme les bottes, sandales et autres chaussures qui tiennent moins bien aux pieds doivent être limitées à un usage occasionnel.
Certaines baskets ont un revêtement synthétique, une semelle rigide, une hauteur de contrefort insuffisante. Elles ne répondent pas à tous les critères de qualité d’une chaussure pour un petit enfant.

Niveau 3 : Le "Clair Logis" - Des enfants auteurs de projets !
Située à proximité du centre-ville dans un cadre verdoyant et sécurisant, cette école conviviale et familiale a pour priorité le dialogue et l'écoute des enfants et des parents. Elle accueille des enfants de niveau maternel et primaire de types 1, 2, 3 et 8 et répond aux besoins de chaque élève. Par une pédagogie dynamique, variée et adaptée, l'école se fixe pour objectif de donner à chacun la possibilité de devenir acteur de son évolution, de lui permettre de s'intégrer, de réaliser des projets.
Pour y parvenir, l’école met en place des classes moins nombreuses, des espaces de manipulations, des animations bibliothèque, des ateliers, des séances de psychomotricité et des projets tels que l'hippothérapie, le jardinage, les petits déjeuners malins, la réalisation d'un journal, des classes de dépaysement et d’autres activités, toutes ayant pour but de renforcer l’estime de soi et l’autonomie.

Niveau 4 : Liquidation des subventions ordinaires
Je vous prie d'excuser le ton impersonnel de la présente, qui revêt pour des raisons pratiques la forme d'une circulaire.
Comme vous le savez, le décret du 17 juillet 2003 prévoit, en disposition transitoire, qu'en attente de l'aboutissement de leur demande de reconnaissance en vertu de ce dispositif, les associations antérieurement reconnues selon le décret de 1976 perçoivent annuellement, et ce durant trois années maximum, un montant total de subventions ordinaires équivalent à celui alloué en 2003. La subvention annuelle en période transitoire sera versée en deux tranches : la première de 80% au cours du premier semestre, le solde au cours du second semestre. Pour rappel cependant, en vertu de l'article 39 dudit décret, sur avis de l'Inspection constatant une baisse sensible du volume annuel d'activités d'une asbl, la subvention annuelle de l'association en question peut être revue à la baisse sur décision du Ministre ; décision motivée qui est alors dûment notifiée par l'administration à l'asbl concernée.
Niveau 5 : Tax shelter : agrément des œuvres par la Communauté française de Belgique

Le texte de définition de l'œuvre européenne selon la directive Télévision sans frontières reprend les points suivants :

1. Une œuvre est européenne quand elle est originaire d'Etats membres ou d'Etats tiers européens parties à la Convention TVTF, réalisée essentiellement avec le concours de professionnels résidant dans ces États, et qui remplit une de ces trois conditions :
   a. établissement du (des) producteur(s) dans un ou plusieurs de ces États (le producteur est établi dans un État européen si son entreprise est une entreprise permanente employant un personnel stable à des activités commerciales et de production en Europe ; orientations d'application de la Directive, juin 1999) ;
   b. contrôle de la production par un (des) producteur(s) établi(s) dans un de ces États ;
   c. contribution majoritaire au coût et contrôle de la coproduction par un (des) producteur(s) établi(s) dans un des ces États.