LC-Score: Reference-less estimation of Text Comprehension Difficulty

Paul Tardy
U31
https://u31.io
pltrdy@gmail.com

Charlotte Roze
U31
https://u31.io
charlotte.roze@u31.io

Paul Poupet
U31
https://u31.io
paul.poupet@u31.io

Abstract

Being able to read and understand written text is critical in a digital era. However, studies shows that a large fraction of the population experiences comprehension issues. In this context, further initiatives in accessibility are required to improve the audience text comprehension. However, writers are hardly assisted nor encouraged to produce easy-to-understand content. Moreover, Automatic Text Simplification (ATS) model development suffers from the lack of metric to accurately estimate comprehension difficulty We present LC-SCORE, a simple approach for training text comprehension metric for any French text without reference *i.e.* predicting how easy to understand a given text is on a $[0, 100]$ scale. Our objective with this scale is to quantitatively capture the extend to which a text suits to the *Langage Clair* (LC, *Clear Language*) guidelines, a French initiative closely related to English Plain Language. We explore two approaches: (i) using linguistically motivated indicators used to train statistical models, and (ii) neural learning directly from text leveraging pre-trained language models. We introduce a simple proxy task for comprehension difficulty training as a classification task. To evaluate our models, we run two distinct human annotation experiments, and find that both approaches (indicator based and neural) outperforms commonly used readability and comprehension metrics such as FKGL and SAMSA.

1 Introduction

The ability to understand text is essential for a wide range of daily tasks. It enables individuals to stay informed, understand administrative forms, and have a full, unimpeded access to social and medical care.

Studies shows that a large fraction of the population experiences comprehension issues in their daily life. Almost half of the OECD population shows reading and written information comprehension difficulties (OECD 2013; Štajner, 2021).

Such difficulties have a major impact in people's life. In France for example, the National Statistic Institute (INSEE 2012) reports that one person out of four has already abandoned an administrative procedure deemed too complicated to follow-along.

In order to improve written text accessibility, initiatives such as Plain Language[1] or *Language Clair* (LC, translates to *Clear Language*) defines writing guidelines to produce clearer texts. Moreover, comprehension makes its way into international standards and norms (ISO 24495; WCAG 2018) but still lacks of concrete solution and measurable objectives.

With the rise of deep-learning approaches in natural language processing, as well as its recent successes in a wide variety of tasks (transcription, translation, summarization, question answering), Automatic Text Simplification is an interesting candidate for accessibility improvements at scale. However, system performances are difficult to measure due to the limitations of current automatic metrics (Alva-Manchego et al., 2021).

We hypothesize that the development of better text comprehension metrics could provide Automatic Text Simplification researchers with a way of validating their models while also to giving measurable objectives for the content editors to write clearer texts.

In this context, we focus our work in developing models for reference-less text comprehension evaluation as a scoring function for French texts *i.e.* $s : \text{text} \mapsto [0, 100]$ reflecting how clearly written a text is.

In this paper, we present the following contributions:

[1] https://plainlanguagenetwork.org/plain-language/what-is-plain-language

- We introduce a simple approach to address comprehension evaluation as a classification task

- We introduce a set of linguistically motivated lexical, syntactic and structural indicators

- We train both indicator based models and text-based Neural Models

- We evaluate our experiments thanks to two human annotation experiments using crowd sourced human judgement for one and expert rating for the second.

2 Related Work

Defining what makes a text difficult to understand is a complex task by itself. Multiple approaches are explored, like studying the age at which children acquires complex syntactic constructions in French (Canut, 2014); or relying on standardized foreign language levels such as the Common European Framework of Reference (CEFR), ranging from A1 to C2. Wilkens et al. (2022) uses this scale to study French as a Foreign Language difficulty.

In order to improve texts clarity, some organizations produced redaction guidelines *i.e.* suggestions of good practices to write clear texts, such as Plain Language (PLAIN) and, in French, (Leys, 2011). Gala et al. (2020) also published guidelines for adapting French texts to increase readability and comprehension. More closely related to our work, Francois and Fairon (2012) introduced a readability formula for French as a foreign language.

Automatic Text Simplification aims at generating simpler versions of a source texts. In literature, such models are usually evaluated using automatic metrics. Therefore, standard language level and redaction guidelines are hardly suitable to evaluate simplification models since it would require an expert judgement. Automatic evaluation instead mostly rely on readability metrics such as FKGL (Kincaid et al., 1975), SMOG (McLaughlin, 1969) and Gunning fog Index (Gunning, 1952). Such metrics were designed with English in mind but can be used on French in practice. On the other hand, SAMSA (Sulem et al., 2018), a semantic metric, is currently not implemented for French, as discussed in section 3.1.

Other approach include learning regression and classification models (Martin et al., 2018) or pre-trained language models (Zhang et al., 2020). However, (Alva-Manchego et al., 2021) found that automatic metrics remains unsuitable to evaluate progress in Automatic Text Simplification.

3 Methods

3.1 Baseline metrics

In order to evaluate our work with respect to the literature we take the following existing readability metrics as baselines: FKGL (Kincaid et al., 1975), SMOG (McLaughlin, 1969), Gunning Fog (Gunning, 1952).

The SAMSA metric (Sulem et al., 2018) takes semantic into consideration. Even though it would be theoretically possible to adapt this metric for french, it is not yet implemented. We tried adapting existing implementation from EASSE (Alva-Manchego et al., 2019) based on CoreNLP (Manning et al., 2014) but it turned out to fail due to the lack of French lemmatization model.

3.2 Evaluate text comprehension difficulty as a classification task

Training a model to predict comprehension difficulty would require a text corpus annotated with comprehension scores. However, to the best of our knowledge, their is no such corpus for the general audience and of sufficient size to envision model training. In this context, we suggest to rely on a simpler proxy task consisting of a classification between *simple* and *complex* texts. Defining what makes a text simple or complex here is difficult. In order to bypass this question, we uses pairs of content sources such as one is roughly a simplified version of the other:

Encyclopedia articles based on French Wikipedia (*complex*) and its simpler alternative, Vikidia (*simple*), designed for 8-13 years old readers. We only took into consideration the introduction paragraph as it is a concise and synthetic presentation of the article. Articles are aligned *i.e.* the corpus consists in $(simple, complex)$ pairs.

International Radio Journal Transcriptions with *France Culture international press review* (*complex*) [2] and *RFI Journal En Français Facile* (*simple*), [3] aimed at french speakers that do not speak the language on a daily basis. Articles

[2] https://www.radiofrance.fr/franceculture/podcasts/revue-de-presse-internationale

[3] https://francaisfacile.rfi.fr/fr/podcasts/journal-en-fran%C3%A7ais-facile/

Corpus	#T	#W/#T	#W/#S
Wikipedia	25812	144	26.0
Vikidia	25812	80	18.9
France Culture	1402	1106	28.8
Journal en Français Facile	1555	1494	19.0

Table 1: Comprehension Classification Datasets: number of texts per corpus ($\#T$), average word per text ($\#W/\#T$) and average word per sentence $\#W/\#S$).

have similar subjects (international news) but are not aligned strictly speaking *i.e.* there is no $(complex, simple)$ pairs for a given article. We report statistics about this new corpus in table 1.

3.3 Linguistic Indicators

Deriving from works on Langage Clair we introduce a set of complexity indicators. Indicators varies from lexical difficulties (*i.e.* a word difficulty score) to syntactic difficulties or sentences parse tree height. Indicators are detailed below.

Indicators are detected based on our own rules implementation using SpaCy pipeline based on both dependency and constituency parsing respectively using `fr-dep-news-trf`[4] and `benepar`[5].

Lexical Indicators (5) These are indicators of difficulties at word level. We use a word difficulty score based on word frequencies in corpora of different difficulty levels: elementary school textbooks of various grades from Manulex (Lété et al., 2004) and French as a Foreign Language textbooks of various CEFR (Common European Framework of Reference for Languages) levels from FLELex (François et al., 2014). Lexical indicators also include abbreviations, acronyms, named entities and numerical expressions.

Sentence Length Indicators (3) We measure sentences lengths with averages of words per sentence; dependency and constituency tree heights.

Syntactic Indicators (17) Several difficulties on the syntactic level in sentences are identified, which are related to sentence structure: coordinate clauses, relative clauses, adverbial clauses, participle clauses, cleft structures, interpolated clauses, appositive phrases, enumerations, etc.). Information about verb forms are also detected: non-finite clauses, passive voice, complex verbal tenses, conditional mood. Negations marks, complex noun phrases and text spans between brackets are also included in syntactic indicators.

Structure Indicators (3) Two indicators are related to the presence of connectives and their potential complexity, estimated by syntactic information (*e.g.* clause position for conjunction connectives, sentence initial position for adverbial connectives) and information from a French connectives lexicon (Roze et al., 2010). A third indicator counts temporal breaks (*i.e.* a tense change) within text paragraphs.

We train models using `sklearn`: two linear models (Linear SVC and Ridge) for fairer comparison to linear readability metrics, and 2 non-linear (Random Forest and Multi Layer Perceptron)

3.4 Neural Methods based on Text

Even though indicator-based approaches rely on linguistic motivations, they lack the possibility to learn from deeper relationships throughout the text such as the subject, the context and the semantic that might carry essential information to infer comprehension difficulty. This is the reason why we chose to compare indicator-based methods with deep learning approaches directly relying on text.

We use two French pre-trained language models such as BARThez (Eddine et al., 2020) and CamemBERT (Martin et al., 2020) fine-tuned with a classification (C) or a regression objective (R).

4 Comprehension Difficulty Annotation

We ran two human annotation experiments in two different contexts: the first one using Mechanical Turk, a crowd-sourcing platform to receive annotations of French speakers from general audience (4.1); the second based on the feedback of Langage Clair experts in our team (4.2).

4.1 Crowd-sourced Human Annotation

In order to get the most reliable annotations we follow (Kiritchenko and Mohammad, 2017) and use a Best-Worst Scaling (BWS) technique. They recommend to use comparison task instead of direct assessment *i.e.* directly giving a note to a given text. More specifically, BWS compares k (typically $k =$

[4] https://spacy.io/models/fr#fr_dep_news_trf

[5] https://github.com/nikitakit/self-attentive-parser

4) simultaneous examples and asks the annotator to select the best one and the worst one with respect to the dimension of interest (text comprehension difficulty in our context).

When annotating texts of up to 200 words, preliminary experiments showed us that comparing $k = 4$ simultaneous texts was too long and fastidious. In this light, we reduce to $k = 3$.

The annotation counts $T = 48$ news articles (up to 200 words). Each text is present in $e = 12$ different examples of $k = 3$ texts. Examples are annotated by $a = 3$ separate annotators in a total of 26. We end up with a total of $E = (T \times e)/k = 192$ examples, and $E \times a$ annotation *i.e.* for any three texts $\{T_a; T_b; T_c\}$ the annotation task consist in submitting an ordered set *e.g.* $T_c > T_a > T_b$.

Each text T_i is associated with an annotation score by $score(i) = \#best\%(i) - \#worst\%(i)$ with $\#best\%(i)$ (resp. $\#worst\%(i)$) representing the frequency at which T_i was evaluated the best (resp. worst) text out of the 3.

In order to measure the reliability of an annotation experiment, a common practice is to measure inter-annotation agreement. However, in a BWS process, each annotators is presented with a different set of examples which makes the concept of annotator agreement less relevant. Moreover, disagreement is even beneficial to produce accurate annotation: for two items A and B of similar difficulty, we can expect half of the annotator to rate $A > B$ and the other half $B > A$. From this apparent disagreement emerges diversity that actually reinforce score accuracy. For this reason, BWS is instead evaluated in terms of reproductibility metrics like Split Half Reliability (SHR). SHR is the correlation between two randomly sampled half of the annotation. In practice, we average SHR over 1000 iterations to rule out randomness.

4.2 Expert Annotation

In addition to crowd-sourced corpus, our team built a small corpus of 74 texts annotated with difficulty scores. We selected 37 texts originating from news articles, literature, and customer support mails. In addition, we provide 37 manually simplified versions following Langage Clair methodology. Each of the 74 resulting texts were then scored on a $[0, 100]$ scale by 4 LC experts from our team.

To make sure we obtained good quality annotation, we measure annotator agreement with Intraclass Correlation Coefficient (ICC2, Shrout and Fleiss, 1979). ICC2 ranges from 0 (no agreement) to 1 (perfect agreeement).

5 Results

5.1 Annotation results

Annotations experiments text length metrics and reliability measure are reported in table 2.

Good reliability from MTurk and Expert even though our annotation experiments are very different in terms of annotators and process, both shows high reliability measures achieving respectively an SHR correlation of 64.7 (MTurk) and an Intraclass Correlation Coefficient of 74.6 (Experts).

Filtering MTurk workers does not increase reliability A common practice when involving crowd-sourced annotation is to filter-out users that shows the lowest agreement. Even though we discussed in 4.2 that agreement is not considered to be the most relevant metric for BWS annotation, we challenge this hypothesis by calculating worker agreement rate based on how often a given user submits the same result than another worker. Then, we suppose that workers with the lowest agreement rate might add noise to the experiment so we might want to exclude them. However, results showed the opposite: filtering out workers does not increase reliability in terms of SHR, no matter the agreement rate of each. This observation is in line with the hypothesis that annotator disagreement is expected and beneficial in a BWS annotation experiment.

	MTurk	**Expert**
#T	48	37 / 37
#W/#T	183	190 / 209
#W/#S	25	28 / 13
#Annotators	26	4
Type	BWS	RS
Reliability Measure	SHR	ICC2
Reliability	64.7	74.6

Table 2: Human Annotation Experiments. Corpus are reported with number of texts per corpus ($\#T$), average word per text ($\#W/\#T$) and average word per sentence $\#W/\#S$). Since Expert is aligned, metrics are reported for both sides. Experiments uses two different annotation processes (i) Best Worst Scaling (BWS) evaluated in term of Split Half Reliability (SHR) and (ii) Rating Scale in $[0, 100]$ (RS, 100 is best) evaluated with Intraclass Correlation Coefficient (ICC2).

Model	**Valid** *acc%*	**MTurk** ρ	**Expert** ρ
SMOG	-	-18.68	-73.09
Gunning Fog	-	-12.59	**-82.14**
FKGL	-	**-19.66**	-77.54
Linear SVC	73.07	20.94	69.37
Ridge	-	27.58	86.44
MLP	75.31	32.56	85.73
Random Forest	**77.20**	**34.42**	**88.09**
BARThez	79.64	23.16	58.41
Camembert(R)	**91.01**	**28.35**	75.85
Camembert(C)	90.15	18.44	**84.73**

Table 3: Scoring models Spearman correlations (ρ) with human judgement. (C) and (R) respectively indicates classification and regression training objective.

5.2 Scoring results

First, we evaluate model performances with respect to their own training by measure accuracy on their validation set: a 10% held-out subset from the training set. Validation accuracy is used to select the best hyper-parameters and training iterations for each models.

Models are then evaluated against human annotations from MTurk and Experts using Spearman Rank Correlations (ρ).

Results are reported in Table 3. Our approaches show better correlations with the human judgement than readability metrics. Models trained from indicators achieves the highest correlations, with Random Forest being the best on both evaluation sets, MTurk and Expert.

It is also interesting that even simple linear statistical models based on our indicators outperforms readability metrics therefore arguing in favor of this indicator set. In particular, the Ridge Regression model outperform FKGL by 14.76 and 10.55 correlation point respectively on MTurk and Expert.

Readability metrics seems complementary in that FKGL achieve better correlation on MTurk evaluation while Gunning Fog does on Expert.

Similarly, we observe sensible differences between Camembert training objectives, with the regression (R) being better on MTurk and classification (C) on Expert.

6 Discussions

Results shows a large improvement of human judgement correlation in favor to our approaches over existing readability metrics. Moreover, indicator based method outperform neural models fine-tuned from pre-trained model. Neural models' results are promising and could be extended with longer training time and adapting their training objective to produce equally distributed scores.

In addition to outperforming neural models, indicator based model are far cheaper to train and predict with since they does not require GPU. Being indicator-based makes it easier to interpret and more predictable than neural models, and thus might deliver a better user experience. We observed Neural models we trained tend to produce very polarized output probabilities *i.e.* either very close to 0 or to 1. That's not a problem to quantitatively evaluate the resulting score, but it should probably be adapted to output equally distributed scores in order to be more intuitive.

7 Conclusion

Developing methods to accurately measure written text comprehension difficulty is a key challenge that would help better assessing the quality of Automatic Text Simplification models, and provide with a tool for editors to produce texts that are simpler to understand.

We explore multiple approaches for training a reference-less metric based on a simple classification task. Our systems rely either on linguistic indicators or directly from text.

To evaluate our models, we two human annotation experiments. The first involves crowd-sourced workers, asked to compare text based on their comprehension difficulties using Best Worst Scaling with $k = 3$. In the second experiment, texts are simplified then rated on a $[0, 100]$ scale by experts from our team.

Both neural and indicator based methods shows promising results and largely outperform other broadly used readability metrics, on both crowd-sourced and expert human annotations. Even simple linear models largely outperform readability metrics which adds an evidence against using it to estimate text comprehension complexity.

As further researches, we suggest exploring multi-lingual neural training. This would have the obvious benefit of overcoming the language restriction of our work while also mutualizing learning from each language and unifying comprehension difficulties estimation accross languages.

8 Lay Summary

Nowadays, most services use the Internet as their primary way of communicating. Therefore, being able to read and understand texts is really important. But a lot of people have difficulties reading and understanding so it is not simple for them to access information or complete administrative procedures.

We introduce a method to calculate a difficulty score for French texts. A score of 0 means that the text is really difficult to understand, whereas a score of 100 means it is really clear. We suggest that developing such a score is a first step toward helping people write easier texts. We gathered two categories of texts: some that we consider easy to understand and others that we consider difficult to understand. Then, we trained models to predict whether a text is categorized as "easy" or not. After training, we use the predictions as our scoring method: the score corresponds to the probability (multiplied by 100) that a text is categorized as easy by the model.

We explored two kinds of models. For the first one, we count different kinds of linguistic difficulties and give them to the model to predict the difficulty. The second kind of model is deep neural networks that have already been trained to learn French. We specialize it in predicting the difficulty based on the text by providing examples of texts and their difficulties.

To measure how relevant our models are, we asked people on the Internet as well as experts to give their opinions on texts. In particular, they were given texts and should determine how difficult they are. We found that people agreed more with our method's scores than with other existing scoring methods.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier Automatic Sentence Simplification Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Emmanuelle Canut. 2014. Acquisition des constructions syntaxiques complexes chez l'enfant français entre 2 et 6 ans. *SHS Web of Conferences*, 8:1437–1452.

Moussa Kamal Eddine, Antoine J. P. Tixier, and Michalis Vazirgiannis. 2020. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model.

Thomas Francois and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, pages 466–477.

Thomas François, Nùria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Núria Gala, Amalia Todirascu, Ludivine Javourey-Drevet, Delphine Bernhard, Rodrigo Wilkens, Jean-Paul Meyer, and Al Recommandations. 2020. Recommandations pour des transformations de textes français afin d'améliorer leur lisibilité et leur compréhension.

Robert Gunning. 1952. The technique of clear writing. *1952*, page 289.

INSEE 2012. 2012. Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul. Standard, INSEE.

ISO 24495. 2023. Plain language – Part 1: Governing principles and guidelines. Standard, International Organization for Standardization.

J. P. Kincaid, R. P Fishburne, R. L Rogers, and B. S Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 465–470, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bernard Lété, Liliane Sprenger-Charolles, and Pascale Colé. 2004. MANULEX: a grade-level lexical database from French elementary school readers. *Behavior Research Methods Instruments and Computers*, 36(1):156–66.

Michel Leys. 2011. Écrire pour être lu : comment rédiger des textes administratifs faciles à comprendre? Technical report, Fédération Wallonie-Bruxelles.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2014-June, pages 55–60. Association for Computational Linguistics (ACL).

Louis Martin, Samuel Humeau, Pierre Emmanuel Mazare, Antoine Bordes, Eric De La Clergerie, Sagot Benoit, Pierre-Emmanuel Mazaré, and Antoine Bordes. 2018. Reference-less Quality Estimation of Text Simplification Systems. In *ATA 2018 - 1st Workshop on Automatic Text Adaptation, Proceedings of the Workshop*, pages 29–38. Association for Computational Linguistics (ACL).

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. Association for Computational Linguistics (ACL).

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

OECD 2013. 2013. OECD Skills Outlook 2013 – First Results from the Survey of Adult Skills. Standard, OECD.

PLAIN. 2023. Federal Plain Language Guidelines. Standard, Plain Language Action and Information Network (PLAIN).

Charlotte Roze, Danlos Laurence, and Philippe Muller. 2010. LEXCONN: a French Lexicon of Discourse Connectives. In *MAD 2010 - 8th Workshop Multidisciplinary Approaches to Discourse*, Proceedings of the 8th Workshop Multidisciplinary Approaches to Discourse (MAD 2010), pages 114–125, Moissac, France.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Sanja Štajner. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 685–696.

WCAG 2018. 2018. Web Content Accessibility Guidelines (WCAG) 2.1 – 3.1 Understanding. Standard, W3C.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment toolkit.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating Text Generation With Bert. *8th International Conference on Learning Representations, ICLR 2020*.