Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification

Lucía Ormaechea, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon, Benjamin Lecouteux

Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification

Lucía Ormaechea[1,2], **Nikos Tsourakis**[1], **Didier Schwab**[2],
Pierrette Bouillon[1] and **Benjamin Lecouteux**[2]
[1] TIM/FTI, University of Geneva, 40 Boulevard du Pont-d'Arve – Geneva, Switzerland
{firstName.lastName}@unige.ch
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG – Grenoble, France
{firstName.lastName}@univ-grenoble-alpes.fr

Abstract

Automatic text simplification models face the challenge of generating outputs that, while being indeed simpler, still retain some complexity. This stems from the inherently relative nature of simplification, wherein a given text is transformed into a relatively simpler version, which does not necessarily equate to simple. We thus aim to propose a finer-grained method to assess sentence complexity in French. Our solution comprises three models, in which two address absolute and relative sentence complexity assessment, while the third focuses on measuring simplicity gain. By employing this triad of models, we aim to offer a comprehensive approach to qualify and quantify sentence simplicity. Our approach utilizes FlauBERT, fine-tuned for classification and regression tasks. Based on our three-dimensional complexity analysis, we provide the WIVICO dataset, comprising 46,525 aligned *complex-simpler* pairs, which can be further leveraged to fine-tune large language models to automatically generate simplified texts, or to assess text complexity with greater granularity.

1 Introduction

Automatic Text Simplification (ATS) aims at producing a simpler version of a given input text, while still preserving its original information, semantic coherence and grammaticality (Horn et al., 2014). The resulting text is expected to be linguistically less complex, which can in turn have an interest from a *human-oriented perspective*, so as to provide with adapted texts for different target readers, like children (De Belder and Moens, 2010) or people with dyslexia (Rello et al., 2013); and a *machine-oriented perspective*, as a pre-processing step for other NLP applications like information extraction (Evans and Orasan, 2019).

Nevertheless, ATS models are subject to generating outputs that, while being indeed simpler, still retain a level of complexity. This arises from the inherently relative nature of simplification, in which a given reference text is rewritten into a comparatively simpler version. Yet, simpler does not necessarily equate to simple, and can result in outputs that still exhibit complex linguistic features.

Predicting sentence complexity seems a valuable ancillary task in this respect, as it can help evaluate the simplification effectiveness of the generated output. In addition, it can contribute to the automatic creation of monolingual *complex-simpler* pairs, which are a scarce resource in ATS, especially for less resource-rich languages than English. Prior research has often addressed sentence complexity assessment by relying on binary classification models (Paetzold and Specia, 2016; Stajner et al., 2017), through which an input is categorized as either *complex* or *simple* on an absolute basis. However, this approach proves somewhat coarse in the context of simplification, considering its acknowledged relative nature. Since ATS models operate based on a provided text, we believe that estimating the sentential complexity should also be conducted in a reference-aware manner.

In this paper, we aim to contribute with a BERT-based finer-grained method to assess sentence complexity, specifically in French. Despite its substantial resources, ATS research on this language remains largely unexplored given the scarcity of parallel simplification data. To alleviate this issue, we introduce a new triad of increasingly fine-grained models so as to: *i*) determine whether a sentence is inherently *complex* or *simple*; *ii*) assess if the second sentence in a pair is simpler than the first; and *iii*) measure the simplification gain achieved by the second sentence in comparison to the original one. Additionally, based on the proposed method, we provide a general-purpose parallel sentence simplification dataset for French language[1].

[1] Which is publicly released on the following GitHub repository: https://github.com/lormaechea/wivico.

2 Background and related work

2.1 Simple and simpler: a fundamental distinction often omitted in ATS

The performance of ATS models is normally judged upon three criteria (Martin, 2021): *i*) how *fluent* the simplified output is; *ii*) how well the *meaning* of source text is preserved in the output; and most notably, *iii*) how simple it is compared to the original unsimplified text. A successful model is thus expected to produce a fluent, lossless-meaning text that is comparatively simpler in form than its original counterpart. This implies that the system is not necessarily designed to generate *simple text*, but rather to achieve or satisfice a simplicity gain with respect to a given text. In other words, the model is aimed at producing a comparatively simpler version of a text, according to a provided input. Yet *simpler* does not equal *simple* by definition. A complex text can be transformed into a relatively simpler version, but still show complex features that would make them inadequate to the constraints of simple language.

Then the question that arises is: what is the notion of *simple*? Is there such a thing as an absolute and objective simplicity that defines one particular text? The concept of *simple language* has been extensively investigated in prior literature, especially in the context of text accessibility. It has been broadly defined as a variety of language that shows low lexical and syntactic complexity (Klaper et al., 2013). Nevertheless, providing proper simplified texts requires a more precise delineation, as it is greatly influenced by the needs of specific target readers (*e.g.*, individuals with cognitive disabilities, foreign language learners, children, *etc*.), which condition the preferred simplification operations accordingly. As can be noted, the audience is not a negligible factor, as it shows that text simplification is a strongly subject-dependent task: the perception of a text as being more easily accessible or comprehensible may vary substantially according to the target reader (Dmitrieva et al., 2021).

In recent years, the growing awareness of the eventual reading comprehension difficulty arisen by some types of documents (*e.g.*, technical, administrative, but also general-domain) (Stajner, 2021), as well as the regulations ratified from institutional frameworks (Nomura et al., 2010), has fostered the definition of easy-to-understand manual simplification style guides, such as *Easy Language* or *Plain Language* (Maaß, 2020). These initiatives were created to provide standards for the writing of comprehensibility-enhanced texts, and to guarantee the quality and appropriateness of the resulting simplifications. Nonetheless, such guidelines often advise the use of overly broad or imprecise simplification-oriented rules, such as the usage of short sentences and simple words, or the avoidance of non-essential information (Candido et al., 2009). Such haziness hinders their eventual applicability within automated text simplification solutions. And, more importantly, it makes it difficult to objectively quantify the extent to which a text complies with a specific guideline (Fajardo et al., 2013; Sutherland and Isherwood, 2016), thus obfuscating a consensual definition of *simple language* and a common characterization of *simple text*.

2.2 Existing approaches for building parallel text simplification corpora

The creation of relevant resources for text simplification is a crucial procedure for the subsequent training and evaluation of data-driven ATS models. However, it poses a significant challenge due to the intricacies associated with defining *simplicity*, as discussed earlier, and also the strong reliance on monolingual parallel corpora comprising representative simplified texts and their corresponding complex references. The paucity of such data collections has significantly hindered progress on this task, both method- and language-wise. To mitigate this issue, previous research has employed two approaches for building parallel *complex-simple(r)* text resources: *manual* and *automatic*, with a special focus on sentence-level simplifications.

Manually-created Manually crafted monolingual parallel corpora for ATS are usually created from scratch, by asking experts (*i.e.*, teachers, translators or speech therapists) to simplify a set of texts (usually genre- or domain-specific), for a particular audience (Brunato et al., 2022). By relying on pre-existing or *ad hoc* target-aware style guidelines, and professional editors' expertise, the resulting sentence simplification pairs are expected to provide a reliable and high-quality parallel dataset.

On this basis, several datasets have been released, such as NEWSELA (Xu et al., 2015), in English and Spanish, PORSIMPLES (Aluisio and Gasperin, 2010) in Brazilian Portuguese, or ALECTOR (Gala et al., 2020) in French. Parallel corpora derived from this approach are notable for their highly reliable simplification operations performed on the

original text. However, this process is costly, both economically and time-wise, due to the requirement of trained human editors. Furthermore, it has an impact on the reduced size of the resulting dataset, which with the exception of NEWSELA, does not easily support the implementation of ML algorithms that are able to infer the transformations to generate simplified text.

Automatically-created With the goal of providing with ATS-oriented high-scale parallel monolingual datasets, automatic data acquisition approaches rely on existing comparable corpora (usually Wiki-based) that associate standard texts with their simplified versions. These resources are later used to extract *complex-simple(r)* sentence pairs, giving rise to labeled data collections, like WIKISMALL (Zhu et al., 2010), EW-SEW (Hwang et al., 2015) or WIKILARGE (Zhang and Lapata, 2017).

While being widely used in the training of ATS models in prior literature (Nisioi et al., 2017; Martin et al., 2020; Sheang and Saggion, 2021), the adequacy of the simplifications within these datasets has been called into question (Xu et al., 2015). This is due to the eventual disparity between the source text and its comparatively simpler counterpart, given the fact that comparable corpora being used are often written independently. In addition to this, their limited controllability has also been debated, since it appears difficult to determine to what extent they observe any style manual, or whether the performed simplifications are target-aware or target-oblivious. Nor is it any less of an impediment that such resources are often solely existing in English, leading data-driven ATS in less resource-rich languages to be harder to implement.

Yet, the main reason to emphasize the unsuitability of these datasets is based on the eventual suboptimality of the methods used to mine register-diversified comparable corpora. So as to capture monolingual parallel data that is relevant for ATS, prior research has typically relied on automatic alignment algorithms and semantic similarity scores (Paetzold et al., 2017; Stajner et al., 2018; Nikolov and Hahnloser, 2019; Sun et al., 2023). Although these strategies are prone to error, they aid in assessing the semantic closeness between two sentences, and thus serve as a proxy for meaning preservation. However, they do not suffice on their own, as they fail to ascertain whether the target text genuinely constitutes a simpler version with respect to the corresponding input. Given that simplicity gain is a *sine qua non* condition for a simplified text to be considered valid, recent studies have explored the use of classification and regression models to estimate sentence complexity, as we will see below.

2.3 Automatic assessment of sentence complexity

Automatically determining the complexity of a sentence proves to be a valuable ancillary task for ATS, as it can potentially serve as a preliminary step in creating labeled simplification data. Additionally, it can aid in evaluating the simplification effectiveness of the generated output.

Prior literature has approached sentence complexity prediction in various ways, depending on the ultimate objective. This typically includes: *i*) detecting the complex sentences needing to be simplified, and *ii*) quantifying the degree of simplification achieved within a pair. As a result, it has had an impact on the approach used for such assessment. So as to address the first goal, previous works have mainly employed absolute complexity classifiers. These models assign a discrete label to an input text that represents its difficulty. This can in turn be treated as a binary classification problem (Paetzold and Specia, 2016; Stajner et al., 2017) or a multi-class discrimination problem, if a greater granularity is considered (Vajjala and Meurers, 2014; Khallaf and Sharoff, 2021). On the other side, relative sentence complexity classifiers (Ambati et al., 2016) and, more particularly, regression models have been prioritized to address the second objective (Iavarone et al., 2021), as they can represent linguistic complexity in a continuum, and help predict the degree of complexity reduction obtained by a simplified sentence.

It is also worth noting that such regressors have commonly been used from the perspective of automatic readability assessment (Lee and Vajjala, 2022). While it is a complementary notion to that of simplification, they are not equivalent concepts. Readability primarily focuses on language clarity and accessibility, and it does not strictly target the *meaning preservation* and *simplicity gain* relation. In addition to this, readability formulae were designed for a document-level application, which means that they may not be completely reliable on a sentential-level (Stajner et al., 2017). This suggests the need to introduce new metrics within ATS, so as to properly quantify the gain or loss of simplicity in a *complex-simpler* pair.

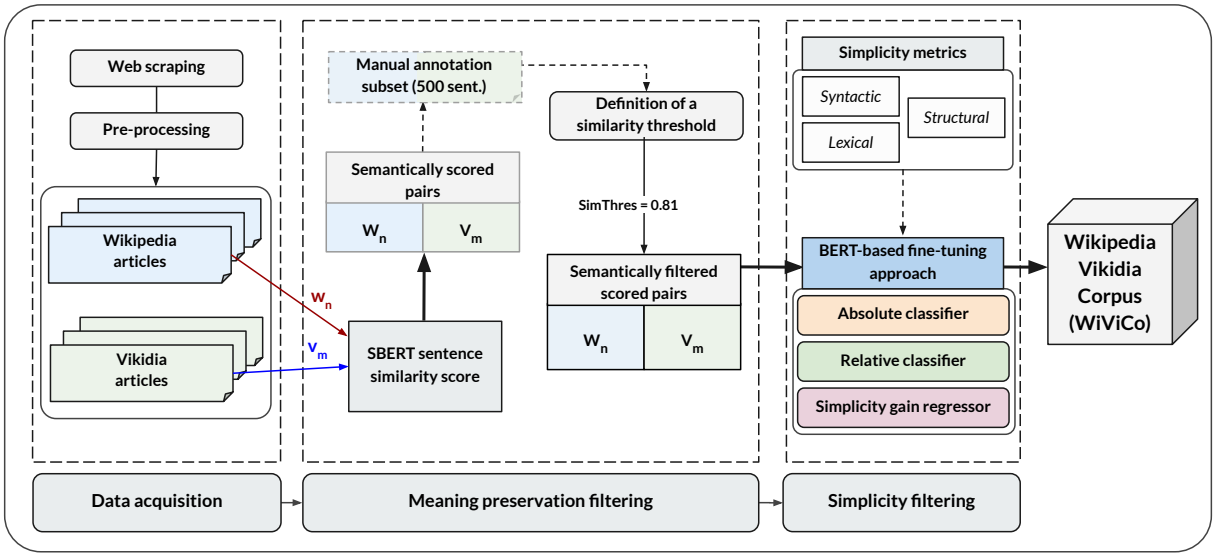

Figure 1: Overview of the pipeline to obtain *complex-simpler* sentence pairs from the French Wikipedia and Vikidia.

3 Corpora

As previously stated, automatically determining the complexity of a sentence (or a pair of sentences) can potentially serve as a helpful preliminary step in creating labeled simplification data in languages such as French, where ATS-specific aligned data is scarce. In this section, we showcase the corpora we used to make such prediction as well as to automatically mine *complex-simpler* pairs.

3.1 WIKILARGE-FR

Assessing sentence simplicity in an automatic manner is generally based on data-driven approaches. Considering this, we opted to rely on WIKILARGE (Zhang and Lapata, 2017), a well-established dataset that has been utilized to develop and refine simplification models in previous ATS research. However, a significant obstacle was encountered since the texts in WIKILARGE were originally written in English, requiring to be translated into French. To tackle this issue, we employed Google Translate to obtain the respective translations for every pair and produced WIKILARGE-FR.

WIKILARGE-FR	
Train size	105,420
Dev size	13,177
Test size	13,179
Total	131,776

Table 1: Overview of size (in sentence pairs) and data distribution of the WIKILARGE-FR dataset.

We identified that certain pairs were too similar during this process, so we kept those with a Levenshtein distance of less than 0.95. We then split the data into a train, validation, and test set using an 80:10:10 split and stratification (see Table 1).

3.2 Wikipedia-Vikidia data compilation

Prior studies have highlighted the potential use of Wiki-based articles for the creation of ATS resources (Brouwers et al., 2012). For this reason, we decided to use the French-language editions of register-differentiated comparable corpora to subsequently extract parallel simplification pairs. More precisely, we relied on Wikipedia and Vikidia, where the latter constitutes an adapted version of the former, and was created to provide with texts that can be more easily understandable by children between 8 and 13 years old. At present, French Vikidia comprises about 40*k* articles, which makes it a significant resource for ATS. Notwithstanding French is a reasonably well-resourced natural language, the available aligned data for this task is limited (Seretan, 2012; Cardon and Grabar, 2019).

In order to retrieve the textual content from the articles of both sources, we extracted the complete URL list of articles from Vikidia using the web scraping pipeline described in Ormaechea and Tsourakis (2023). The output yielded a total of 34,357 article links[2]. We later parsed the HTML content to find the corresponding Wikipedia articles, by relying on inter-language links. Afterwards, we tokenized the text content and segmented

[2] As of April 14th, 2023.

it into sentences. We finally filtered out the sentences exceeding 128 word pieces, so as to avoid an eventual truncation when encoded into a sentence embedding.

4 Meaning preservation filtering

As discussed in Section 2.1, the output produced by an ATS model is expected to meet two primary conditions: *i*) retain the meaning and information conveyed in the input text, and *ii*) obtain a linguistic simplicity gain with respect to the reference. Based on this definition, we addressed these two dimensions sequentially. In order to determine suitable *complex-simpler* pairs for ATS, we must first assess whether they are semantically equivalent[3].

We thus implemented a meaning preservation filtering method to identify the Wiki-Viki pairs exhibiting a high semantic overlap. To this effect, we relied on SBERT (Reimers and Gurevych, 2019), which modifies the pretrained BERT network (Devlin et al., 2019) by using a siamese architecture to compute sentence embeddings[4]. After mapping the sentences to a 768-dimensional dense vector space, we computed the cosine similarity for the resulting encoded pairs.

Once such values were obtained, we needed to assess which pairs showed sufficient semantic consistency. To this end, we chose to rely on a manual annotation of 500 randomly picked sentence pairs from our initial dataset. Two subjects were selected for this purpose. They were given three judgment labels to conduct the annotation: *valid*, where the meaning from source to target is fully preserved; *partially valid*, where information is partially lost from source to target or vice versa; and *non-valid*, where information between the two sentences diverges. After the first annotation round, the two experts convened to discuss and reached a consensus, resulting in a Cohen's kappa score of 0.87. With 500 annotated sentence pairs at our disposal, we plotted the distribution of the SBERT scores for each judgment label. On average, *valid* pairs show higher SBERT-derived values, which confirms a direct correlation between SBERT scoring and human judgments on sentence similarity. The mean score for *valid* pairs was 0.81, which we consider the cutoff threshold for the semantic filtering step.

[3] If their meaning is divergent, no assessment on simplicity gain is applicable.

[4] We used multilingual sentence transformers: https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1.

5 Simplicity filtering

After addressing the meaning preservation dimension, we focused on how to extract the simplicity gain obtained by the target sentence with respect to the reference. Our approach consists of three distinct steps to assess absolute and relative simplification and estimate a gain score (as shown in Figure 1), and aims to properly address the relative nature of *simplification*. An absolute binary categorization of a sentence as *complex* or *simple* seems somewhat insufficient and not suited for ATS. Indeed, a complex sentence (C) being transformed into a simple one (S) results in a *simplification*. Conversely, a S→C process gives rise to a *complexification*. Nevertheless, an absolute classifier can equally categorize a source and target sentences as C→C or S→S. Given that *simplification* and *complexification* operations are reference-dependent, they may validly occur in both cases.

Because there are several phenomena involved within simplicity assessment, we split the problem into an increasingly fine-grained approach. First, we incorporated the WIKILARGE-FR dataset to elicit pairs of *complex-simpler* sentences that can be used to fine-tune different versions of FlauBERT (Le et al., 2020). For the classification task, we created two models: one to assess the simplicity of each sentence in the pair, and another to determine whether the target sentence is simpler than the corresponding source. Subsequently, based on a set of features, we calculated the simplicity gain for each pair that allowed the creation of a regressor model to automate this process. For a clearer depiction of the specific steps involved, refer to Figure 2.

5.1 Classification models for sentence complexity

Fine-tuning pre-trained classification models can help leverage their learned knowledge and transfer it to a new classification task. By adapting the model to the target task with labeled data, we can improve its generalization, capture domain-specific nuances, and achieve better results. In our work, we incorporated a specific architecture based on the FlauBERT language model to perform sentence complexity classification. It is a variant of the model that has been adapted specifically for sequence classification. In this architecture, the model is combined with additional layers and a classification head to enable it to classify sequences into different categories.

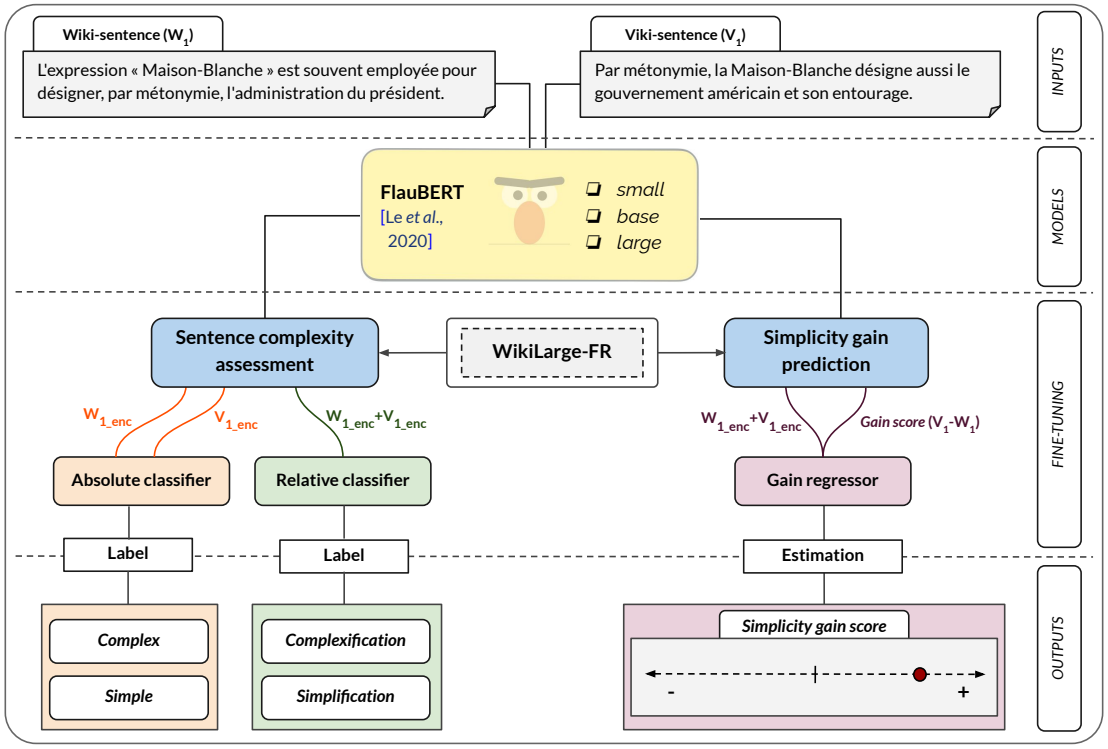

Figure 2: Overview of the simplicity assessment task.

5.1.1 Absolute sentence complexity assessment

In the first experiment, we treated each sentence in the input pairs independently to determine whether it is categorized as *simple* or *complex*. To achieve this, we assigned a binary label for each of the sentences in the WIKILARGE-FR dataset (see Section 3.1). The performance on the test set is presented on the left side of Table 2. Utilizing different variants of the FlauBERT model, we contrasted the performance between each baseline model (*untuned*) and the one after training (*tuned*). We observe significant improvement in the second case, which is similar to all three variants. The baseline untuned models' performance was no better than random chance in distinguishing between the two classes (∼50%) versus the tuned ones (∼70%). It is worth noting that the small version of the untuned FlauBERT model is partially trained, which may impact its performance. Nevertheless, it was included for debugging purposes.

5.1.2 Relative sentence complexity assessment

The second classifier aims to assess the relative simplification between the source and target sentence pairs, answering the question of whether the second is a *simpler* version of the first. To accomplish this, we juxtaposed the sentences alternating their order into two sets of pairs to signify either *simplification* or *complexification*. This time, we significantly improved the baseline performance (∼50% versus ∼93%). To reinforce the validity of the previous outcome, we also utilized the manually annotated dataset of Section 4, which included human annotations of relative simplification. The results shown on the right side of Table 2 corroborate our previous assessment. As the dataset is imbalanced, the baseline classifiers' performance mirrors the class distribution and can largely be attributed to chance. However, the tuned models improve those significantly (∼94%).

5.2 A regression model for simplicity gain

The classification models presented above allow us to discern in a binary manner whether a sentence is *complex* or *simple*, or whether a pair of sentences has undergone a process of *simplification* or *complexification*. However, these models lack the capacity to indicate to what extent a target sentence is *simpler* than its original counterpart. For these reasons, we have aimed to quantify the simplification shift produced within a pair of classically categorized *complex-simple* sentences, with the training of a regression model. In this way, we have sought to measure the *simplicity gain* achieved from the original sentence to its simplified version.

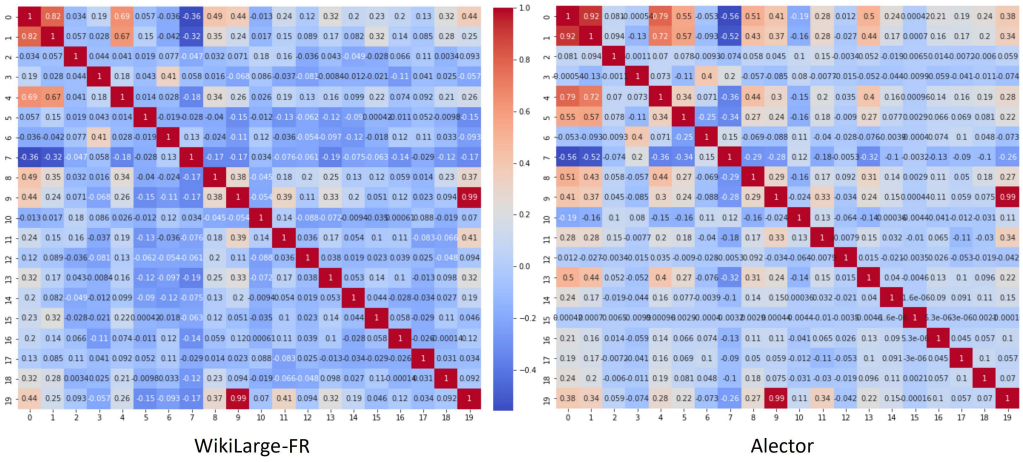

Figure 3: Correlation heatmaps among the feature gains for the WIKILARGE-FR and ALECTOR datasets.

Classification task	AC		RC			
Evaluation dataset	Test set		Test set		Manual set	
Transformer model	untuned	tuned	untuned	tuned	untuned	tuned
`flaubert-small`	49.54	70.11	49.78	92.99	34.58	92.52
`flaubert-base`	50.97	69.82	49.88	93.82	36.45	93.46
`flaubert-large`	52.29	69.19	52.18	94.16	75.71	95.33

Table 2: Accuracy results in % obtained for the *absolute complexity classifier* (AC) on the test set, and for the *relative complexity classifier* (RC) on the test and manual evaluation sets.

As noted in Section 2.3, similar regression models have been used from a readability perspective, but they prioritize the measurement of clarity and accessibility aspects, and do not explicitly address the challenges of ATS. This is why we sought to examine the quantification of the simplicity gain.

5.2.1 Definition of features

We extracted a set of pertinent features, shown in Table 4, that were chosen on the basis of previous literature regarding sentence simplicity assessment (Tanguy and Tulechki, 2009; Brunato et al., 2022). These describe the WIKILARGE-FR dataset along three dimensions and are grouped into structural, lexical, and syntactic groups. Based on these features, we calculated their values for each sentence in the pair and performed an element-wise subtraction. The result is a list containing the differences between the elements in the same positions of the original feature lists that we also standardized.

While using a predictive model to estimate the *simplicity gain* from *complex-simpler* pairs might not be necessary when a direct calculation process is available, there are potential benefits to consider. Predictive models can assist in quality assessment by identifying cases where direct calculations may falter due to assumptions or heuristics. They offer generalization capabilities, making predictions for new data and variations that the direct process may not cover. Additionally, these models can uncover hidden patterns, adapt to changes in data distributions, and provide robustness against noisy or imperfect data, enhancing their value in real-world scenarios. For that reason, LLMs can be beneficial by leveraging their capacity to comprehend and learn from intricate language patterns in the data.

To tackle the challenge of collinearity, we calculated the correlation of the simplicity gains shown in the left heatmap of Figure 3. This heatmap aids in detecting patterns and dependencies among the features. This helps to identify the impact of each one on the overall simplicity gain and to decide on which to keep in the subsequent analysis. We observe that certain pairs demonstrate a high correlation, like *Sentence length* and *Number of words* (row: 0 – col: 1) or *IDT* and *IDT-DLT* (row: 9 – col: 19). We therefore excluded the second feature in each pair, ending with 18 features in total.

We also performed a symmetric analysis on the aforementioned ALECTOR dataset (shown in the right heatmap of Figure 3). Given that it was manually created by expert linguists, the produced simplifications are expected to be highly reliable. This

in turn helps to reinforce our decision to maintain or exclude features according to their relevance to the simplicity assessment. Interestingly, we observe similar patterns of correlation, indicating that the features have a similar effect in both datasets.

5.2.2 Simplicity gain estimation

Similarly to the classification tasks, we fine-tuned FlauBERT for regression. By utilizing the Mean Squared Error (MSE) as the loss function, Adam optimizer and a batch size of 16, we trained FlauBERT to learn to map its linguistic representations to continuous target variables. The input received by the regressor consisted on the *complex-simpler* pairs appended to their simplicity gain score, with a maximum input size of 512 tokens.

GR		
Evaluation dataset	Test set	
Transformer model	untuned	tuned
`flaubert-small`	1.89	0.39
`flaubert-base`	1.18	0.35
`flaubert-large`	4.59	0.23

Table 3: MSE scores from the *gain regressor* (GR).

Table 3 contrasts the performance on the test set using either an untuned or a tuned FlauBERT model. We observe a significant improvement in all three cases. Specifically, the tuned models achieved a much lower MSE, demonstrating their ability to capture underlying patterns in the data and provide more accurate predictions. The `flaubert-large` model yields the best performance with an MSE equal to 0.23, which seems still insufficient in the context of our application. These results may suggest further exploration in the optimization of the model hyperparameters, but they may also point towards a broader categorization of each pair based on a range of gain values.

5.3 Wikipedia-Vikidia Corpus (WIVICO)

Having this triad of models in place, we were able to finally implement our fine-grained method on sentence simplicity to extract relevant pairs for ATS. To do so, we implemented our best performing models on the compiled data introduced in Section 3.2. As a result, we were able to generate the Wikipedia-Vikidia Corpus (WIVICO), that contains 46,525 aligned sentence pairs[5]. These include standard C→S labeled examples, but also C→C and S→S ones, where a simplification operation was performed (as can be seen in Appendix B).

[5] Appendix C provides a detailed description of the dataset).

6 Conclusions and further work

This paper presents an increasingly fine-grained approach for assessing sentence simplicity. Through a comprehensive three-dimensional analysis, our objective was to estimate sentence simplicity in a manner suitable for ATS, which is an inherently relative operation. Additionally, we believe that our work can serve as a relevant and reproducible method to automatically create parallel simplification datasets. This can in turn be of great interest for reasonably well-resourced natural languages like French that still lack sufficient resources for the ATS task. Consequently, we provide public access to the dataset that derives from the application of our approach, WIVICO. This may allow other researchers interested in this field to further use this resource to fine-tune LLMs for the task at hand, or to assess text complexity in a finer-grained manner.

As for the limitations of this work, it is important to note that due to the volume of the WIKILARGE corpus, we had to resort to Google Translate to obtain the corresponding French texts, without manually assessing the correctness of the produced outputs. A possible workaround to this drawback would be to compare a subset of the produced WIKILARGE-FR with its original counterpart and conduct a human evaluation of translation quality.

On another note, an extension of our investigations points to the creation of configurable ATS models. We could incorporate our triad of models into a larger pipeline designed for text simplification and use them to rank a set of candidate simplified sentences, with the goal of selecting the most simplified sentence that best preserves the original meaning of the input. Similarly, the fine-tuned model can serve as a guide during the simplification process by providing a continuous feedback signal to a generative ATS model and therefore adjust its output to attain a desired level of simplification.

Last but not least, we also intend to work on improving the interpretability of the assigned score for simplicity gain. While based on a calculation resulting from established linguistic features for text simplicity, we believe it is also necessary to contrast such scores to human judgments. By doing so, we can examine the correlation between the two in more depth, and measure the significance of each feature in the simplicity gain estimation.

Acknowledgements

This work is part of the PROPICTO (French acronym standing for *PRojection du langage Oral vers des unités PICTOgraphiques*) project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005).

References

Sandra Aluisio and Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing Relative Sentence Complexity using an Incremental CCG Parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057. Association for Computational Linguistics.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2012. Simplification Syntaxique de Phrases pour le Français. In *Actes de la Conférence Conjointe JEP-TALN-RECITAL*, pages 211–224.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian. *Frontiers in Psychology*, 13.

Arnaldo Candido, Erick Maziero, Lucia Specia, Caroline Gasperin, Thiago Pardo, and Sandra Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: A Text Simplification Editor for Brazilian Portuguese. In *NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42.

Rémi Cardon and Natalia Grabar. 2019. Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification. In *Proceedings - Natural Language Processing in a Deep Learning World*, pages 168–177.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Workshop on Accessible Search Systems*, pages 19–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186. Association for Computational Linguistics.

Anna Dmitrieva, Antonina Laposhina, and Maria Lebedeva. 2021. A Comparative Study of Educational Texts for Native, Foreign, and Bilingual Young Speakers of Russian: Are Simplified Texts Equally Simple? *Frontiers in Psychology*, 12.

Richard Evans and Constantin Orasan. 2019. Sentence Simplification for Semantic Role Labelling and Information Extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 285–294.

Inmaculada Fajardo, Vicenta Clemente, Antonio Ferrer, Gema Tavares, Marcos Gómez, and Ana Hernández. 2013. Easy-to-read Texts for Students with Intellectual Disability: Linguistic Factors Affecting Comprehension. *Journal of Applied Research in Intellectual Disabilities (JARID)*, 27:212–225.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 458–463.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence Complexity in Context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199. Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic Difficulty Classification of Arabic Sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114. Association for Computational Linguistics.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier

Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association.

Justin Lee and Sowmya Vajjala. 2022. A Neural Pairwise Ranking Model for Readability Assessment. In *Findings of the Association for Computational Linguistics*, pages 3802–3813. Association for Computational Linguistics.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*. Frank & Timme.

Louis Martin. 2021. *Automatic Sentence Simplification using Controllable and Unsupervised Methods*. Ph.D. Thesis, Sorbonne Université.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable Sentence Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698.

Nikola I. Nikolov and Richard Hahnloser. 2019. Large-Scale Hierarchical Alignment for Data-driven Text Rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 844–853.

Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 85–91.

Misako Nomura, Gyda Skat Nielsen, International Federation of Library Associations and Institutions, and Library Services to People with Special Needs Section. 2010. *Guidelines for Easy-to-Read Materials*. IFLA Headquarters.

Lucía Ormaechea and Nikos Tsourakis. 2023. Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method. In *Proceedings of the 8th Swiss Text Analytics Conference 2023*. Association for Computational Linguistics.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and Annotation of Comparable Documents. In *Proceedings of the IJCNLP, System Demonstrations*, pages 1–4.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 560–569. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013. DysWebxia: Textos Más Accesibles Para Personas con Dislexia. *Procesamiento del Lenguaje Natural*, 51.

Violeta Seretan. 2012. Acquisition of Syntactic Simplification Rules for French. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 4019–4026.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352. Association for Computational Linguistics.

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

Sanja Stajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A Tool for Customized Alignment of Text Simplification Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 3895–3903.

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic Assessment of Absolute Sentence Complexity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4102.

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023. Exploiting Summarization Data to Help Text Simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–51. Association for Computational Linguistics.

Rebekah Sutherland and Tom Isherwood. 2016. The Evidence for Easy-Read for People With Intellectual Disabilities: A Systematic Literature Review: The Evidence for Easy-Read for People With Intellectual Disabilities. *Journal of Policy and Practice in Intellectual Disabilities*, 13:297–310.

Ludovic Tanguy and Nikola Tulechki. 2009. Sentence Complexity in French: a Corpus-Based Approach. In *Intelligent Information Systems (IIS)*, pages 131–145.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the Relative Reading Level of Sentence Pairs for Text Simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1353–1361.

A Set of features for simplicity gain

Table 4: Selected features for the definition of the simplicity gain score.

Group	#	ID	Feature	Description
Structural	0	**SL**	Sentence length	*n* of characters comprising a sentence.
	1	**NW**	Number of words	*n* of words comprising a sentence.
	2	**VSL**	Verbal subject length	*n* of words comprising the verbal subject.
	3	**ATL**	Average token length	Average *n* of characters per token in a sentence.
Lexical	4	**CEFR**	CEFR score	Within a sentence, sum of the frequencies of CEFR levels of all non-stop words multiplied by their lexical complexity weight value (Ormaechea and Tsourakis, 2023).
	5	**NE**	Incidence of named entities	*n* of named entities (organizations, people, places, *etc*.) in a sentence.
	6	**LD**	Lexical density	Ratio between the *n* of content words (*i.e.*, nouns, adjectives, adverbs and verbs) and the total *n* of tokens in a sentence.
	7	**TTR**	Type-token ratio	*n* of unique words divided by the total *n* of words in a sentence.
Syntactic	8	**MDT**	Maximum depth tree	Maximum depth of the dependency tree.
	9	**IDT**	Incomplete dependency theory	Average number of incomplete dependencies between the current and next token.
	10	**DLT**	Dependency locality theory	For every head token in a sentence, *n* of discourse referents starting from the current token and ending to its longest leftmost dependent. Values are then combined using an average function.
	11	**LE**	Left embeddedness	*n* of tokens on the left-hand-side of the root verb that are not verbs.
	12	**NND**	Noun nested distance	Average nested distance of all nouns within a phrase that have as ancestor another noun in the dependency tree.
	13	**CC**	Use of coord. clauses	*n* of clauses linked by a coordinating conjunction.
	14	**SC**	Use of subord. clauses	*n* of clauses linked by a subordinating conjunction.
	15	**PR**	Use of parenthetical remarks	*n* of parenthesized information items in a sentence.
	16	**NEG**	Number of negations	*n* of negative adverbs in a sentence (that implies a slower processing with respect to affirmative ones).
	17	**PAS**	Incidence of passive forms	*n* of passive voice verbs in a sentence (that implies a longer reading time with respect to active ones).
	18	**CT**	Incidence of complex tenses	*n* of complex or unusual verb tenses, *i.e.*, those other than infinitive or present, present perfect, imperfect, future indicative.
	19	**IDT-DLT**	Combined IDT-DLT	Sum of IDT-DLT metrics for all tokens in a sentence. Resulting values are then combined using an average function.

B Application of classification and regression models to Wikipedia-Vikidia pairs

Table 5: Applying the triad models to Wikipedia-Vikidia sentence pairs. A gloss in English is provided below each segment for clarity purposes.

	Wikipedia sentence	**Vikidia sentence**
Pair$_1$	En France, ce lézard est strictement protégé par la loi.	En France, il est protégé par la loi.
Gloss	In France, this lizard is strictly protected by law.	In France, it is protected by law.
AC	`Complex`	`Simple`
RC	`Simplification`	
GR	`0.84`	
Pair$_2$	Praticien précoce et représentant éminent du concept français de la haute gastronomie, il est considéré comme le fondateur de ce style grandiose, recherché à la fois par les cours royales et les nouveaux riches de Paris.	Il est considéré comme l'un des pionniers, sinon le fondateur, de la gastronomie française.
Gloss	As an early practitioner and leading exponent of the French concept of *haute gastronomie*, he is considered the founder of this grandiose style, sought after by both the royal courts and the newly rich of Paris.	He is considered one of the pioneers, if not the founder, of French gastronomy.
AC	`Complex`	`Complex`
RC	`Simplification`	
GR	`2.45`	
Pair$_3$	Makassar ou Macassar est une ville d'Indonésie et la capitale de la province de Sulawesi du Sud.	Macassar ou Makassar est une ville d'Indonésie, située sur l'île de Sulawesi (ou Célèbes), en bordure du détroit du même nom.
Gloss	Makassar or Macassar is a city in Indonesia and the capital of the province of South Sulawesi.	Macassar or Makassar is a city in Indonesia, on the island of Sulawesi (or Celebes), bordering the strait of the same name.
AC	`Simple`	`Complex`
RC	`Complexification`	
GR	`-2.65`	

C Detailed description of the Wikipedia-Vikidia Corpus (WIVICO) dataset

Table 6: Detailed description of WIVICO. We purposely use *texts* and not *sentences* because our dataset includes intersentential examples (*i.e.*, texts comprising more than one sentence).

WIVICO dataset	**Original texts**	**Simpler texts**
# texts	46,525	
# tokens	1,730,277	1,321,139
# types	100,357	73,926
Type/token ratio	5.80	5.60
Average word length	5.27	5.04
Average sentence length	38.63	29.08