Extracting sentence simplification pairs from French comparable corpora using a two-step filtering method

Ormaechea Grijalba, Lucía; Tsourakis, Nikolaos

Extracting sentence simplification pairs from French comparable corpora using a two-step filtering method

Lucía Ormaechea* and Nikos Tsourakis

TIM/FTI, University of Geneva, 40 Boulevard du Pont-d'Arve, Geneva, 1205, Switzerland

Abstract

Automatic Text Simplification (ATS) aims at simplifying texts by reducing their linguistic complexity while retaining their meaning. While being an interesting task from a societal and computational perspective, the lack of monolingual parallel data prevents an agile implementation of ATS models, especially in less resource-rich languages than English. For these reasons, this paper investigates how to create a general-language parallel simplification dataset for French using a method to extract *complex-simple* sentence pairs from comparable corpora like Wikipedia and its simplified counterpart, Vikidia. By using a two-step automatic filtering process, we sequentially address the two primary conditions that must be satisfied for a simplified sentence to be considered valid: (1) preservation of the original meaning, and (2) simplicity gain with respect to the source text. Using this approach, we provide a dataset of parallel sentence simplifications (WiViCo) that can be later used for training French sequence-to-sequence general-language ATS models.

Keywords

Automatic text simplification, Comparable corpora, Automatic sentence alignment, Wiki-based resources, Sentence semantic similarity, Low-resourced tasks

1. Introduction

Automatic Text Simplification (ATS) is an area of NLP that aims at automatically converting texts into simpler variants, by reducing their linguistic complexity, albeit preserving their original meaning [1, 2]. ATS plays an important role from a societal point of view, as it seeks to provide a comprehensibility aid for different target readers (*i.e.*, children [3], people with low literacy skills [4] or dyslexia [5], among others). Furthermore, it proves to be a valuable task from a machine-oriented perspective, as it can efficiently serve as a pre-processing step for other NLP applications such as Machine Translation (MT) [6].

However, the limited availability of simplification-related resources has proven to be a bottleneck for the advancement of this field. Automating Sentence-level Simplification (SS) is particularly reliant on the existence of large-scale parallel monolingual corpora, that associate *complex-simple* pairs, and can hence help train supervised ATS systems [7]. The paucity of such data collections has strongly influenced research on this under-resourced task, both method- and language-wise.

To mitigate this problem, new approaches have been proposed in recent years. By relying on unsupervised methods [8], researchers have proposed promising ways to exploit unlabeled corpora to generate simplified sentences and thus enormously lessen the need for aligned texts. Yet, not completely. Oftentimes, these unsupervised approaches are complemented with existing labeled data, either to help gain additional knowledge of simplification [9], or as a means to mine aligned pairs and thus generate more data to help improve the performance of simplification models [10, 11]. The existence of such aligned texts is often solely available in English (*i.e.*, WikiSmall [12], Ew-Sew [13], Newsela[1] [14], WikiLarge [15]), leading data-driven ATS in less resource-rich languages to be tougher to implement, due to the unavailability of high-scale corpora that facilitate the training of ATS models.

Given the above considerations, the research question that we have aimed to address is the following: how can we automatically extract *complex-simple* sentence pairs from comparable corpora that are relevant to the SS task? To answer this inquiry, we have sought to provide a method by which appropriate parallel data can be extracted for text simplification from comparable Wiki-based corpora. More particularly, we decided to focus on the French language, and to explore its versions of Wikipedia and Vikidia articles. The latter is a simplified version of the former and aims to make texts more easily accessible for children. At present[2], French Vikidia contains about 38k published articles, and is the language for which the largest number of Viki-articles are written, making it seemingly a non-negligible resource for ATS.

Our two-fold contribution can be summarized as follows: (1) we present a two-step filtering pipeline for mining suitable *complex-simple* pairs for ATS, and (2) we provide a general-language sentence simplification dataset in French (*Wikipedia-Vikidia Corpus*, WiViCo).

Swiss Text Analytics Conference, Neuchâtel, Switzerland, 2023.

* Corresponding author.

Lucia.OrmaecheaGrijalba@unige.ch (L. Ormaechea);

Nikolaos.Tsourakis@unige.ch (N. Tsourakis)

[1] A Spanish version of Newsela has recently made available.

[2] As of March 2023.

2. Background and Related Work

2.1. Automatic Text Simplification in French

Despite French being a well-resourced language for many NLP-related tasks, the body of literature regarding text simplification remains somewhat limited. ATS research on this language has explored rule-based methods for syntactic simplification [16, 17], and only more recently lexical simplification [18] and medical discourse adaptation [19]. SS remains a largely unexplored aspect, which merely includes a few experiments performed on a synthetic version of Newsela to train a neural-based ATS model [20].

As for the available parallel corpora, to our knowledge, there exists only the Alector corpus [21], which is a collection of 79 aligned *complex-simple* literary and scientific texts that were simplified at the document-level for child audiences. In spite of being a high-quality corpus, its compact size makes it insufficient for the training of simplification models.

Additionally, prior studies have also pointed to the potential exploitation of Wiki-based comparable corpora in French such as the aforementioned Wikipedia and Vikidia [16]. However, it has been argued that Wiki-texts and their simplified counterparts, Viki-texts, are often written independently, which makes it challenging to identify parallel pairs between them [19]. In any case, the claim regarding the complexity of this task has not been empirically demonstrated. To the best of our knowledge, there has not been any study attempting to implement an automatic alignment method to construct a general-purpose parallel corpus from Wiki- and Viki-based articles[3].

2.2. Characterizing Sentence-level Simplification Operations

SS seeks to detect potentially complex constructions (*e.g.*, deviations from canonical linear order, long subordinate clauses, *etc.*) and rewrite them into simpler versions, but without entailing meaning loss. To a certain extent, we can imagine such process in a two-dimensional space, where we seek to reduce its linguistic complexity (reflected in its *form*) whilst maintaining intact, as much as possible, the original meaning, *i.e.*, its *substance*.

To achieve this, SS typically performs on two linguistic axes: (1) *syntagmatic*, thus transforming the syntax and grammatical structure of the input sentence; and (2) *paradigmatic*, hence replacing lexical terms with simpler ones. Nevertheless, these changes do not necessarily involve one single sentence, but can affect a broader scope, depending on the performed transformations. More specifically:

1. *Intra-sentential* operations, referring to the simplification changes that are produced within the scope of one single sentence (that is, on a *1:1* basis). These include word(s) *substitution*, *reordering*, *paraphrasing* as well as *deletion*, in the context of superfluous information.
2. *Inter-sentential* operations, referring to the changes involving several sentences, that is, on a *n:m* basis (two examples are shown in Table 1):
    - *Divergence* (or *splitting*), *viz.*, dividing long sentences into shorter and less complex segments (with $m>n$).
    - *Convergence* (or *compression*), namely, rewriting n sentences into a simpler and more compact version (with $n>1$ and $m<n$).

It should be noted that, when automating SS, the explicit implementation of such transformations may vary depending on the model being designed, as each system tends to target a specific linguistic axis or operation [12].

2.3. Automatic Sentence Alignment

Sentence alignment aims at mining comparable corpora to extract parallel sentences, *i.e.*, pairs whose semantic content is equivalent. While originally designed to align bilingual texts[4], sentence alignment mechanisms have also garnered attention in the context of monolingual tasks like summarization, style transfer and ATS. As a result, there have been developed alignment algorithms specifically geared towards simplification.

More precisely, there are language-independent tools that allow the alignment of *complex-simple* pairs from comparable monolingual documents, such as MassAlign [22], CATS [23] and LHA [24]. These employ various document and/or paragraph and sentence alignment methods to generate *n:m* pairs. Other alignment tools such as VecAlign [25], not initially intended for ATS, have also been used for this purpose [26].

In addition to alignment algorithms, similarity measures have been used as a proxy for semantic closeness between sentence pairs. Such is the case of SBERT [27], which modifies the pretrained BERT network [28] by using siamese and triplet network structures to compute sentence embeddings that can later be compared using a cosine similarity measure. SBERT has been applied in the context of standard and simplified sentence mapping, particularly as a means of obtaining alignments at a *1:1* [29] and *n:1* level [26, 30]. Nevertheless, so far, its implementation does not align *n:m* pairs, which appears crucial for extracting inter-sentential simplification operations such as splitting.

[3] Only articles that are indexed in the French Vikidia medical portal have been formerly utilized for automatic alignment purposes [19].

[4] And consequently be used for training MT systems.

Table 1
Inter-sentential examples extracted from the French versions of Wikipedia (*Original*) and Vikidia (*Simpler*). A gloss in English is provided below each segment for clarity purposes.

Convergence	***Original***	La Ligue internationale contre le racisme et l'antisémitisme (LICRA) est une association luttant contre le racisme et l'antisémitisme en France, mais également sur le plan international. Elle est fondée en 1927 sous le nom de Ligue internationale contre l'antisémitisme (LICA).
	Gloss	The International League against Racism and Antisemitism (LICRA) is an association fighting against racism and antisemitism in France, as well as internationally. It was founded in 1927 under the name of International League against Antisemitism (LICA).
	Simpler	La Ligue internationale contre le racisme et l'antisémitisme (LICRA) est une association antiraciste et anti-discrimination créée en 1927 qui agit dans le monde entier.
	Gloss	The International League against Racism and Antisemitism (LICRA) is an antiracist and antidiscrimination association created in 1927 that operates worldwide.
Divergence	***Original***	Lio, de son vrai nom Vanda Maria Ribeiro Furtado Tavares de Vasconcelos, née le 17 juin 1962 à Mangualde au Portugal, est une chanteuse et actrice luso-belge francophone.
	Gloss	Lio, whose real name is Vanda Maria Ribeiro Furtado Tavares de Vasconcelos, was born on June 17 1962, in Mangualde, Portugal, and is a French-speaking Luso-Belgian singer and actress.
	Simpler	Lio, de son vrai nom Vanda Maria Ribeiro Furtado Tavares de Vasconcelos, est une chanteuse et actrice luso-belge francophone. Elle est née le 17 juin 1962 à Mangualde au Portugal.
	Gloss	Lio, whose real name is Vanda Maria Ribeiro Furtado Tavares de Vasconcelos, is a French-speaking Luso-Belgian singer and actress. She was born on June 17 1962, in Mangualde, Portugal.

3. Method

3.1. Data Acquisition

In order to mine monolingual bi-texts from the French-language edition of Wikipedia and its simplified counterpart, Vikidia, we firstly needed to perform a web scraping procedure to extract the article texts from both sources. Due to the unavailability of Vikidia dumps[5], we were compelled to parse all article pairs using alternative tools.

In the context of the compilation of our Wiki-based dataset, we decided to take into account the total number of parallel articles between the two encyclopedias. For each of them, we extracted the summaries included in the article's preface (also known as *lead section*), on the hypothesis that there might be greater chances of finding aligned sentences, given its prevalent definitional style. The implemented pipeline was the following:

1. We initiated the extraction process by parsing the URL list of all available articles from Vikidia. To carry out this step, we resorted to the `vikitext` Python library[6]. The output yielded a total of 34,806 article links, although this number has increased subsequently[7].
2. We later parsed the HTML content of the extracted URLs to find the corresponding Wikipedia articles, with the use of inter-language links. It should be noted that a naive replacement from "vikidia" to "wikipedia" in the link does not necessarily work as expected: a Vikidia page may not have a corresponding Wikipedia one, or may redirect to a disambiguation page rather than a genuine article.
3. Subsequently, we pre-processed the extracted lead sections using the SpaCy library to clean the text and segment it into sentences. A more detailed description of the corpus obtained is shown in Table 2.

Table 2
Data collected from the lead sections of Wikipedia and Vikidia.

Dataset	Wiki-texts	Viki-texts
# documents	34,806	
# sentences	165,806	134,348
# tokens	4,030,148	2,373,045
# types	294,979	195,791
Type/token ratio	7.32	8.25
Avg. word length	5.32	5.08
Avg. sent. length	25.22	18.16

[5] Seemingly, a portal for Viki-dumps exists (https://dumps.vikidia.org/), but we were unable to gain access to this service.
[6] https://pypi.org/project/vikitext/.
[7] More precisely, to 38,429 (https://www.vikidia.org/), visited on: 21/03/2023.

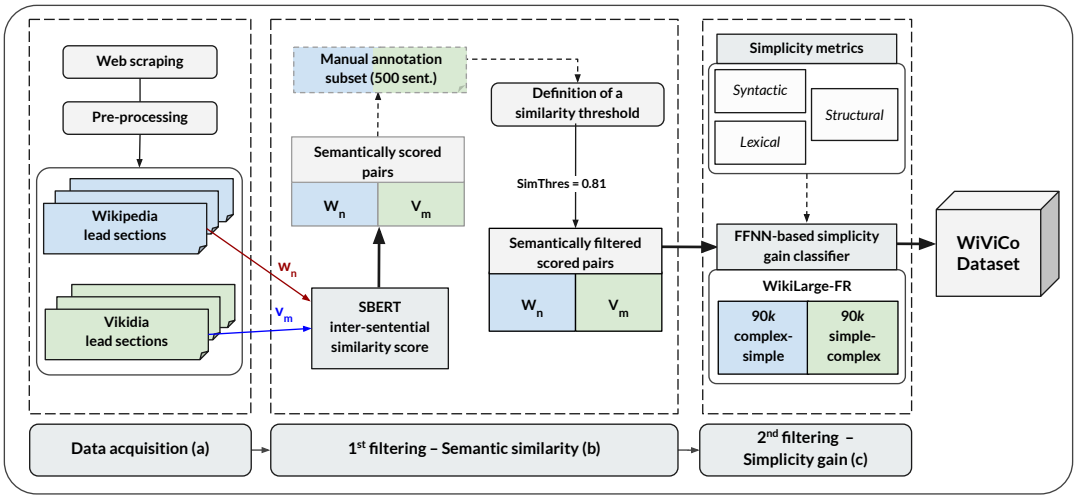

Figure 1: Overview of the complete filtering pipeline to obtain *complex-simple* sentence pairs from the French versions of Wikipedia and Vikidia.

3.2. Semantic Similarity Filtering

3.2.1. Automatic Sentence Alignment

As discussed in Section 2.2, the output produced as a result of automatically simplifying an input sentence needs to meet two primary conditions: (1) retention of the original meaning and information, and (2) a linguistic simplicity gain with respect to the reference[8].

Based on this definition, we opted to tackle these two dimensions in a sequential manner, so as to properly proceed to the extraction of pertinent sentence pairs for the ATS task. To put it another way, if we seek to determine which *complex-simple* pairs are suitable for ATS, we must first ascertain whether they are semantically equivalent or not. If they are semantically divergent, then no assessment on simplicity gain is applicable.

In this manner, we first implemented a semantic filtering method, so as to extract the Wiki- and Viki-pairs that exhibit a high semantic overlap. To do so, we relied on SBERT (see Section 2.3), which has previously been employed to identify standard and simplified parallel sentences. Using SBERT allowed us to compute the similarity measure between two sentence embeddings without incurring significant computational and time costs.

In our case, though, we opted to introduce a modification to the way in which it has been applied to date. Since we intended to capture all simplification operations (both intra- and inter-sentential), we decided to compute sentence embeddings on a multi-sentence basis, and thereby apply a *n:m*-aware sentence alignment. We thus fed SBERT with W_n Wiki-sentences and V_m Viki-sentences, with $1 \leq n, m \leq 3$, where $n, m \in \mathbb{N}$. We decided not to conceive a sentence alignment algorithm, in reason of the typically short length of the input lead sections.

For its implementation, we used multilingual sentence transformers[9], to tokenize our sentences and to map them to a 768-dimensional dense vector representation. Since the model has an upper limit of 128 word pieces, we excluded the input sentences exceeding that maximum to avoid truncation. We then computed the cosine similarity values for the remaining W_n:V_m encoded pairs (as shown in region *b* in Figure 1).

3.2.2. Manual Annotation

Indeed, using SBERT-derived cosine similarity values as a proxy for semantic equivalence is a suitable option within our framework, especially considering that it has surpassed the performance of previous state-of-the-art models on many sentence-pair regression tasks [27]. Yet, we still need to assess which pairs are sufficiently semantically consistent. For this reason, in this section, we discuss how to extract an appropriate value to filter sentence pairs according to semantic similarity.

After randomly selecting 500 samples from the initial dataset, we relied on two annotators to determine to which extent each pair of W_1:V_1 sentences conveyed the same meaning. The two subjects were given three

[8] A third dimension is *fluency* [2]. Given the fact that we are extracting human-crafted texts, and not producing machine-generated outputs, we deemed it appropriate to assume without verification that the input texts are already grammatical.

[9] More precisely: https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1.

Table 3
Manually annotated examples on sentence semantic similarity. The dissimilar fragments from the *partially* and *non-valid* examples are red colored. A gloss in English is provided below each segment for clarity purposes.

Label	Wikipedia sentence	Vikidia sentence
Valid	L'expression « Maison-Blanche » est souvent employée pour désigner, par métonymie, l'administration du président.	Par métonymie, la Maison-Blanche désigne aussi le gouvernement américain et son entourage.
Gloss	The term "White House" is often used as a metonym for the president's administration.	By metonymy, the White House also refers to the US government and its entourage.
Partially valid	Neal McDonough est un acteur et producteur américain né le 13 février 1966 à Dorchester (Massachusetts).	Neal McDonough est un acteur américain.
Gloss	Neal McDonough is an American actor and producer born on February 13, 1966 in Dorchester, Massachusetts.	Neal McDonough is an American actor.
Non-valid	L'information désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire.	Les écrits, les sons, les images, les odeurs ou les goûts contiennent de l'information.
Gloss	The information refers both to the message to be communicated and the symbols used to write it.	The written words, sounds, images, smells or tastes contain information.

judgment labels to conduct the annotation (Table 3 shows an example of each case):

- *valid*, where the meaning and information from W_1 to V_1 is fully preserved;
- *partially valid*, where information is partially lost from W_1 to V_1 or vice versa;
- *non-valid*, where information between W_1 and V_1 is dissimilar.

Following the completion of the first annotation round, they convened to discuss their disagreements and reach a consensus. Based on this procedure, their final agreement yields a Cohen's kappa score equal to 0.87. Having 500 annotated sentence pairs at our disposal, we then plotted the distribution of SBERT scores for each judgment label (as shown in Figure 2). On average, *valid* pairs exhibit higher SBERT-derived values, which confirmed a positive correlation between SBERT scoring and human judgments on sentence similarity.

The mean score for the *valid* case was equal to 0.81, which we consider as the cutoff threshold for the semantic filtering of W_1:V_1 pairs. Conversely, applying the same threshold for the other combinations necessitated additional processing. In Figure 3, we plot the distribution of scores for all samples in the corpus. The aim was to shift all mean values proportionally based on the cutoff threshold extracted earlier. This process is described in the following steps:

- For the list of mean score values, $score$, with length n, we shift each element of the list by a certain amount.

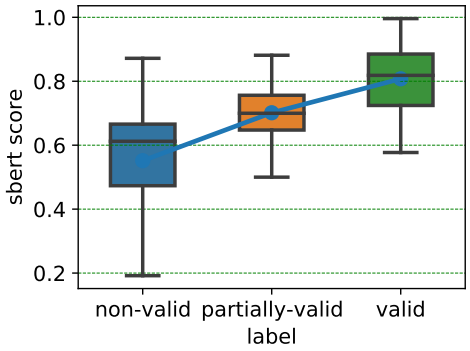

Figure 2: Box and whisker plot distribution of SBERT-derived cosine similarity values for each human judgment label.

- We then use a decreasing shift factor as we move down the list. Let s be the initial shift amount, which equals to the threshold (0.81). Then:
  - $new\_score[0] = score[0] + s$
- For each subsequent element i in the list, we calculate the shift amount as:
  - $shift = s \cdot (1 - \frac{i}{n-1}) \cdot p$
- Then, we calculate the new shifted value for the remaining elements as:
  - $new\_score[i] = score[i] + shift$

We used the factor p so that the shift applied to each element of the list decreases proportionally. The steps are shown below:

- We calculate the differences between adjacent elements in $scores$:
  - $diffs = scores[i+1] - scores[i]$
- The list of ratios of the absolute differences to the next element in the list is:
  - $ratios[i] = abs(\frac{diffs[i]}{scores[i+1]})$
- Finally, the proportion p is calculated as:
  - $p = \frac{sum[ratios]}{n-1}$

At the end of this process, we acquired nine cutoff thresholds, one per W_n:V_m configuration.

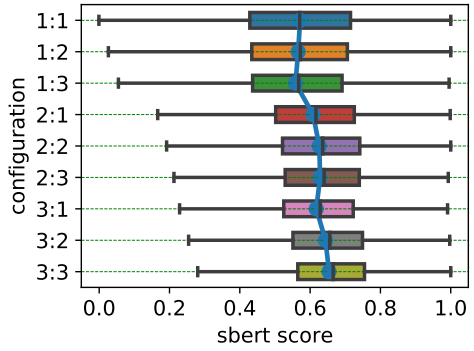

Figure 3: Distribution of SBERT scores for all sentence pairs according to W_n:V_m configurations.

3.3. Simplicity Gain Filtering

The filtering of the previous section provides a set of semantically similar sentences. The next step consists in determining which of those constitute valid simplification pairs. For this purpose, we exploited around 180k example pairs from WikiLarge [15]. This dataset has been extensively used to develop and refine text simplification models. An obvious impediment is that WikiLarge texts are written in English, which required translation into French. To accomplish this, we resorted to Google Translate to obtain the corresponding translations for each sentence pair.

Next, we trained a simplicity gain classifier based on pertinent features, shown in Table 5. The features describe the dataset along three dimensions and are grouped into structural, lexical, and syntactic groups. As for the CEFR (*Common European Framework of Reference for Languages*) score, we employed the FLELex lexicon [31], that associates French lemmas with their corresponding CEFR levels (A1, A2, B1, *etc.*). We assigned weight values to each level, on the basis of a non-linear scale that places greater importance on increasingly challenging vocabulary (see Figure 4). It is based on the following formula:

$$f(x) = 0.1 * 1.5^x$$

To acquire positive and negative examples of simplification, we split the dataset into two halves interchanging the source/target order in the second subset. Identical pairs were also removed. Then, we extracted the values of the features independently for each article and combined them for a single pair. This process uses the notion of *gain*, which signifies the absolute difference between a feature value in the source and target sentence, including polarity. For instance, if the *number of words* feature is *8* in the source and *6* in the target, the gain becomes *-2*. The outcome of this process is a table of gains for each article pair that we then standardized. Finally, as in every typical ML pipeline, we split the data into a train, validation, and test set using an 80:10:10 split and stratification.

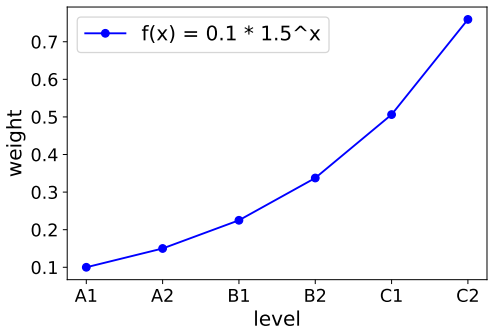

Figure 4: Weight values assigned to each CEFR level.

The classifier was a Feed-Forward Neural Network (FFNN) with four hidden layers of 256 nodes each and ReLU for activation. We used a batch size of 64 and a dropout rate equal to 0.5. Before training, we used the *SelectKBest* feature selection technique to identify the most relevant features from the dataset. We measured the correlation between features and the target variable, eliciting those attributes with the highest scores. Specifically, we aimed to extract feature subsets starting with a single item until the whole feature space, which can be used to train the classifier. The combination of the best features in each case and the corresponding accuracy of the classifier is shown in Figure 5.

We observe that using the whole set of features provides the best performance, and even solely the sentence length is a good predictor (SL). The confusion matrix in Figure 6 indicates a balance between the number of misclassifications for the two classes.

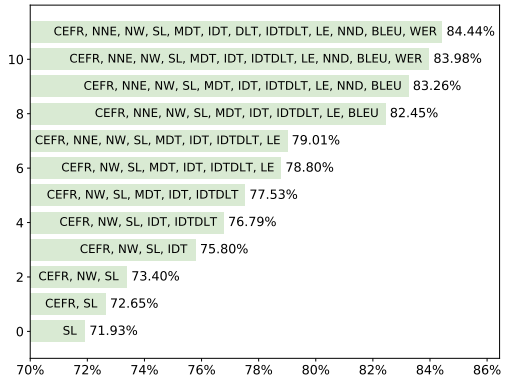

Figure 5: Accuracy of the classifier for the different mixtures of features.

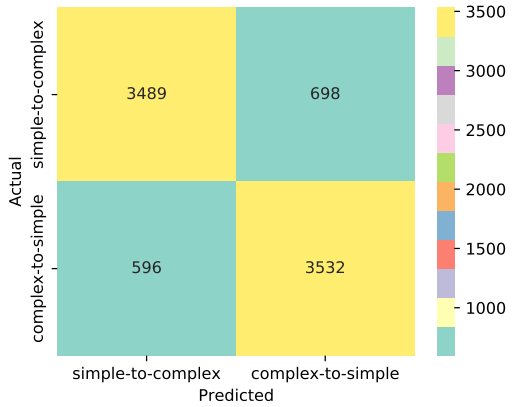

Figure 6: Confusion matrix illustrating the distribution of accurate and erroneous classifications for the two classes.

4. Results and Discussion

Through the implementation of our two-step filtering method, we managed to create a sentence simplification French parallel corpus, comprising relevant sentence pairs for the text simplification task. From the application of the first step of our filtering method, we were able to extract sentences from comparable corpora exhibiting a high semantic overlap. Based on the output of this stage, we performed a second pass to identify the text pairs in which Vikidia represents a simpler sentence in comparison to its Wikipedia counterpart.

An obvious deficiency in the first step is the absence of a sentence alignment algorithm, which would facilitate the detection of parallel sentences in a more efficient manner. As for the second step, although feature-based approaches are a compelling choice, they cannot easily capture the contextual understanding required to comprehend the meaning of words throughout the entire text and thus enhance classification accuracy.

With the implementation of the proposed two-step filtering method we created the WiViCo[10] monolingual parallel corpus. Utilizing the sigmoid output layer of the classifier we provide simplification pairs based on lenient or stricter thresholds (see results on Table 4). Based on different cutoff probability thresholds, we enumerate all W_n:V_m samples in each class. In this manner, other researchers can benefit from this incremental approach and make use of the subset that suits their needs.

Table 4
Results obtained by the automatic simplicity gain classifier.

	Label	
Probability	**0** (*non-simplified*)	**1** (*simplified*)
>0.9	44,049	20,692
>0.8	22,556	42,185
>0.7	18,642	46,099
>0.6	11,087	53,654
>0.5	7,302	57,439

5. Conclusions and Further Work

Through our research, we have developed a method to mine comparable corpora so as to extract relevant parallel *complex-simple* sentences for text simplification. By using a two-step automatic filtering process, we sequentially addressed the two main conditions that need to be fulfilled for a simplified sentence to be considered valid: (1) preservation of the original meaning, that we implemented with the use of *n:m*-aware SBERT-based cosine similarities, and (2) simplicity gain with respect to the original text, that was treated with a text simplicity classifier. Using this approach, we produced a dataset of parallel sentence simplifications, WiViCo. Our intention is to subsequently utilize it to train a sequence-to-sequence general-language ATS model for French, or to fine-tune a pretrained LLM for our downstream task.

While the size of our resulting dataset is not negligible, we envision to enlarge it, by incorporating the full Wiki- and Viki-articles into our filtering pipeline. To that end, we will design a *n:m*-aware SBERT-based sentence algorithm, that can help capture both intra- and inter-sentential simplification operations.

Lastly, we intend to conduct further investigations in order to improve the accuracy of the simplicity gain classifier. To achieve this, we plan to compare the performance between the feature-based approach already in use with pretrained BERT models fine-tuned for a text classification problem.

[10]The WiViCo dataset is hosted on the following GitHub repository: https://github.com/lormaechea/wivico.

Table 5
Selected features for the automatic classification of text simplicity *gain*, which refers to the absolute difference between a feature value in the source and target sentence.

Feature group	Feature	Description
Structural	Sentence Length (SL)	Difference in the number of characters between the target and source sentences.
	Number of Words (NW)	Difference in the number of words between the target and source sentences.
	Word Error Rate (WER)	Word-based similarity between the source and target sentences.
	BLEU score	*n*-gram overlap via precision of the target sentence with its corresponding source sentence.
Lexical	Number of Named Entities (NNE)	Difference in the number of named entities (organizations, people, places, *etc.*) between the target and source sentences.
	CEFR score	Within a sentence, sum of the frequencies of CEFR levels of all non-stop words multiplied by their complexity weight value.
Syntactic	Maximum Depth Tree (MDT)	Difference in the maximum depth of the dependency tree between the target and source sentences.
	Incomplete Dependency Theory (IDT)	Within a phrase, the average number of incomplete dependencies between the current and next token.
	Dependency Locality Theory (DLT)	For every head token in a sentence, the number of discourse referents starting from the current token and ending to its longest leftmost dependent [32]. Values are then combined using an average function.
	Combined IDT+DLT	Sum of IDT+DLT metrics for all tokens in a sentence. Resulting values are then combined using an average function.
	Left Embeddedness (LE)	Within a sentence, the number of tokens on the left-hand-side of the root verb that are not verbs.
	Noun Nested Distance (NND)	The average nested distance of all nouns within a phrase that have as ancestor another noun in the dependency tree.

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

[1] A. Candido, E. Maziero, L. Specia, C. Gasperin, T. Pardo, S. Aluisio, 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, 2009, pp. 34–42. URL: https://aclanthology.org/W09-2105/.

[2] C. Horn, C. Manduca, D. Kauchak, Learning a Lexical Simplifier Using Wikipedia, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 458–463. URL: http://aclweb.org/anthology/P14-2075.

[3] J. De Belder, M.-F. Moens, Text Simplification for Children, in: Workshop on Accessible Search Systems, 2010, pp. 19–26. URL: https://core.ac.uk/works/73269324.

[4] S. Aluisio, L. Specia, C. Gasperin, C. Scarton, Readability Assessment for Text Simplification, in: Proceedings of the NAACL HLT Fifth Workshop on Innovative Use of NLP for Building Educational Applications, 2010, pp. 1–9. URL: https://aclanthology.org/W10-1001.

[5] L. Rello, R. Baeza-Yates, H. Saggion, DysWebxia: Textos Más Accesibles Para Personas con Dislexia, Procesamiento del Lenguaje Natural 51 (2013). URL: http://rua.ua.es/dspace/handle/10045/30664.

[6] S. Stajner, M. Popovic, Can Text Simplification Help Machine Translation?, Proceedings of the 19th Annual Conference of the European Association for Machine Translation (2016) 230–242. URL: https://aclanthology.org/W16-3411.

[7] S. Nisioi, S. Stajner, S. P. Ponzetto, L. P. Dinu, Exploring Neural Text Simplification Models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017, pp. 85–91. URL: http://aclweb.org/anthology/P17-2014.

[8] J. Qiang, X. Wu, Unsupervised Statistical Text

Simplification, IEEE Transactions on Knowledge and Data Engineering 33 (2021) 1802–1806. URL: https://ieeexplore.ieee.org/document/8871118/.

[9] S. Surya, A. Mishra, A. Laha, P. Jain, K. Sankaranarayanan, Unsupervised Neural Text Simplification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2058–2068. URL: https://aclanthology.org/P19-1198.

[10] L. Martin, É. de la Clergerie, B. Sagot, A. Bordes, Controllable Sentence Simplification, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 4689–4698. URL: https://aclanthology.org/2020.lrec-1.577.

[11] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, B. Sagot, MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1651–1664. URL: https://aclanthology.org/2022.lrec-1.176.

[12] Z. Zhu, D. Bernhard, I. Gurevych, A Monolingual Tree-based Translation Model for Sentence Simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING), 2010, pp. 1353–1361. URL: https://aclanthology.org/C10-1152.

[13] W. Hwang, H. Hajishirzi, M. Ostendorf, W. Wu, Aligning Sentences from Standard Wikipedia to Simple Wikipedia, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 211–217. URL: https://aclanthology.org/N15-1022.

[14] W. Xu, C. Callison-Burch, C. Napoles, Problems in Current Text Simplification Research: New Data Can Help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://direct.mit.edu/tacl/article/43283.

[15] X. Zhang, M. Lapata, Sentence Simplification with Deep Reinforcement Learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 584–594. URL: https://aclanthology.org/D17-1062.

[16] L. Brouwers, D. Bernhard, A.-L. Ligozat, T. François, Simplification Syntaxique de Phrases pour le Français, in: Actes de la Conférence Conjointe JEP-TALN-RECITAL, 2012, pp. 211–224. URL: https://hal.archives-ouvertes.fr/hal-00790862.

[17] V. Seretan, Acquisition of Syntactic Simplification Rules for French, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), 2012, pp. 4019–4026. URL: https://aclanthology.org/L12-1138/.

[18] E. Rolin, Q. Langlois, P. Watrin, T. François, FrenLyS: A Tool for the Automatic Simplification of French General Language Texts, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2021, pp. 1196–1205. URL: https://aclanthology.org/2021.ranlp-1.135.

[19] R. Cardon, N. Grabar, Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification, in: Proceedings - Natural Language Processing in a Deep Learning World, 2019, pp. 168–177. URL: https://aclanthology.org/R19-1020/.

[20] S. Abdul Rauf, A.-L. Ligozat, F. Yvon, G. Illouz, T. Hamon, Simplification Automatique de Texte dans un Contexte de Faibles Ressources, in: JEP/-TALN/RECITAL, 2020, pp. 332–341. URL: https://aclanthology.org/2020.jeptalnrecital-taln.33.

[21] N. Gala, A. Tack, L. Javourey-Drevet, T. François, J. C. Ziegler, 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, 2020, pp. 1353–1361. URL: https://aclanthology.org/2020.lrec-1.169.

[22] G. Paetzold, F. Alva-Manchego, L. Specia, MASSAlign: Alignment and Annotation of Comparable Documents, in: Proceedings of the IJCNLP, System Demonstrations, 2017, pp. 1–4. URL: https://aclanthology.org/I17-3001.

[23] S. Stajner, M. Franco-Salvador, P. Rosso, S. P. Ponzetto, CATS: A Tool for Customized Alignment of Text Simplification Corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), 2018, pp. 3895–3903. URL: https://aclanthology.org/L18-1615.

[24] N. I. Nikolov, R. Hahnloser, Large-Scale Hierarchical Alignment for Data-driven Text Rewriting, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2019, pp. 844–853. URL: https://aclanthology.org/R19-1098.

[25] B. Thompson, P. Koehn, Vecalign: Improved Sentence Alignment in Linear Time and Space, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019, pp. 1342–1348. URL: https://aclanthology.org/D19-1136/.

[26] S. Ebling, A. Battisti, M. Kostrzewa, D. Pfütze, A. Rios, A. Säuberli, N. Spring, Automatic Text Simplification for German, Frontiers in Communication 7 (2022) 706–718. URL: https://www.zora.uzh.ch/id/eprint/218829/.

[27] N. Reimers, I. Gurevych, 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 (EMNLP-IJCNLP), Association for Com-

putational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.

[28] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.

[29] D. Aumiller, M. Gertz, Klexikon: A German Dataset for Joint Summarization and Simplification, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 2693–2701. URL: https://aclanthology.org/2022.lrec-1.288.

[30] R. Sun, Z. Yang, X. Wan, Exploiting Summarization Data to Help Text Simplification, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 39–51. URL: https://aclanthology.org/2023.eacl-main.3.

[31] A. Pintard, T. François, Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words, in: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), 2020, pp. 85–92. URL: https://aclanthology.org/2020.readi-1.13/.

[32] L. Zou, M. Carl, M. Mirzapour, H. Jacquenet, L. Nunes Vieira, AI-Based Syntactic Complexity Metrics and Sight Interpreting Performance, Springer International Publishing, 2022, pp. 534–547. URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-98404-5_49.