# Syntactic Sentence Simplification for French

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, Thomas François

# Syntactic Sentence Simplification for French

**Laetitia Brouwers**
Aspirante FNRS
CENTAL, IL&C
UCLouvain
Belgium

**Delphine Bernhard**
LiLPa
Université de Strasbourg
France

**Anne-Laure Ligozat**
LIMSI-CNRS
ENSIIE
France

**Thomas François**
CENTAL, IL&C
UCLouvain
Belgium

## Abstract

This paper presents a method for the syntactic simplification of French texts. Syntactic simplification aims at making texts easier to understand by simplifying complex syntactic structures that hinder reading. Our approach is based on the study of two parallel corpora (encyclopaedia articles and tales). It aims to identify the linguistic phenomena involved in the manual simplification of French texts and organise them within a typology. We then propose a syntactic simplification system that relies on this typology to generate simplified sentences. The module starts by generating all possible variants before selecting the best subset. The evaluation shows that about 80% of the simplified sentences produced by our system are accurate.

## 1 Introduction

In most of our daily activities, the ability to read quickly and effectively is an undeniable asset, even often a prerequisite (Willms, 2003). However, a sizeable part of the population is not able to deal adequately with the texts they face. For instance, Richard et al. (1993) reported that, in 92 applications for an unemployment allowance filled by people with a low level of education, about half of the required information was missing (some of which was crucial for the processing of the application), mainly because of comprehension issues.

These comprehension issues are often related to the complexity of texts, particularly at the lexical and syntactic levels. These two factors are known to be important causes of reading difficulties (Chall and Dale, 1995), especially for young children, learners of a foreign language or people with language impairments or intellectual disabilities.

In this context, automatic text simplification (ATS) appears as a means to help various people access more easily the contents of the written documents. ATS is an application domain of Natural Language Processing (NLP) aiming at making texts more accessible for readers, while ensuring the integrity of their contents and structure. Among the investigations in this regard are those of Caroll et al. (1999), Inui et al. (2003) and, more recently, of Rello et al. (2013), who developed tools to produce more accessible texts for people with language disabilities such as aphasia, deafness or dyslexia. In the FIRST project, Barbu et al. (2013) and Evans and Orăsan (2013) implemented a simplification system for patients with autism, who may also struggle to understand difficult texts.

However, reading assistance is not only intended for readers with disabilities, but also for those who learn a new language (as first or second language). De Belder and Moens (2010) focused on ATS for native English schoolchildren, while Siddharthan (2006), Petersen and Ostendorf (2007) and Medero and Ostendorf (2011) focused on learners of a second language. Williams and Reiter (2008), Aluisio et al. (2008) and Gasperin et al. (2009) addressed ATS for illiterate adults. Most of these studies are dealing with the English language, with the exception of some work in Japanese (Inui et al., 2003), Spanish (Saggion et al., 2011; Bott et al., 2012), Portuguese (Aluísio et al., 2008) and French (Seretan, 2012).

ATS was also used as a preprocessing step to increase the effectiveness of subsequent NLP operations on texts. Chandrasekar et al. (1996) first considered that long and complex sentences were an obstacle for automatic parsing or machine translation and they showed that a prior simplification may result in a better automatic analysis of sentences. More recently, Heilman and Smith (2010) showed that adding ATS in the context

of automatic question generation yields better results. Similarly, Lin and Wilbur (2007) and Jonnalagadda et al. (2009) optimized information extraction from biomedical texts using ATS as a preprocessing step.

In these studies, the simplifications carried out are generally based on a set of manually defined transformation rules. However, ATS may also be solved with methods from machine translation and machine learning. This lead some researchers (Zhu et al., 2010; Specia, 2010; Woodsend and Lapata, 2011) to train statistical models from comparable corpora of original and simplified texts. The data used in these studies are often based on the English Wikipedia (for original texts) and the Simple English Wikipedia, a simplified version for children and non-native speakers that currently comprises more than 100,000 articles. Similar resources exist for French, such as Vikidia and Wikimini, but texts are far less numerous in these as in their English counterpart. Moreover, the original and simplified versions of an article are not strictly parallel, which further complicates machine learning. This is why, so far, there was no attempt to adapt this machine learning methodology to French. The only previous work on French, to our knowledge, is that of Seretan (2012), which analysed a corpus of newspapers to semi-automatically detect complex structures that has to be simplified. However, her system of rules has not been implemented and evaluated.

In this paper, we aim to further investigate the issue of syntactic simplification for French. We assume a midway point between the two main tendencies in the field. We use parallel corpora similar to those used in machine learning approaches and analyse it to manually define a set of simplification rules. We have also implemented the syntactic part of our typology through a simplification system. It is based on the technique of overgeneration, which consists in generating all possible simplified variants of a sentence, and then on the selection of the best subset of variants for a given text with the optimization technique known as integer linear programming (ILP). ILP allows us to specify a set of constraints that regulate the selection of the output by the syntactic simplification system. This method has already been applied to ATS in English by Belder and Moens (2010) and Woodsend and Lapata (2011).

To conclude, the contributions of this paper are: (1) a first corpus-based study of simplification processes in French that relies on a corpus of parallel sentences, (2) the organization of this study's results in what might be the first typology of simplification for French based on a corpus analysis of original and simplified texts; (3) two new criteria to select the best subset of simplified sentences among the set of variants, namely the spelling list of Catach (1985) and the use of keywords, and finally (4) a syntactic simplification system for French, a language with little resources as regards text simplification.

In the next sections, we first present the corpora building process (Section 2.1) and describe a general typology of simplification derived from our corpora (Section 2.2). Then, we present the system based on the syntactic part of the typology, which operates in two steps: overgeneration of all possible simplified sentences (Section 2.3.1) and selection of the best subset of candidates using readability criteria (Section 2.3.2) and ILP. Finally, we evaluate the quality of the syntactically simplified sentences as regards grammaticality, before performing some error analysis (Section 3).

## 2 Methodology

### 2.1 Corpus Description

We based our typology of simplification rules on the analysis of two corpora. More specifically, since our aim is to identify and classify the various strategies used to transform a complex sentence into a more simple one, the corpora had to include parallel sentences. The reason why we analysed two corpora is to determine whether different genres of texts lead to different simplification strategies. In this study, we focused on the analysis of informative and narrative texts. The informative corpus comprises encyclopaedia articles from Wikipedia [1] and Vikidia [2]. For the narrative texts, we used three classic tales by Perrault, Maupassant and Daudet and their simplified versions for learners of French as a foreign language.

To collect the first of our parallel corpora, we used the MediaWiki API to retrieve Wikipedia and Vikidia articles with the same title. The

[1] http://fr.wikipedia.org

[2] This site is intended for young people from eight to thirteen years and gathers more accessible articles than Wikipedia, both in terms of language and content. It is available at the address http://fr.vikidia.org

WikiExtractor [3] was then applied to the articles to discard the wiki syntax and only keep the raw texts. This corpus comprises 13,638 texts (7,460 from Vikidia and only 6,178 from Wikipedia, since some Vikidia articles had no counterpart in Wikipedia).

These articles were subsequently processed to identify parallel sentences (Wikipedia sentence with a simplified equivalent in Vikidia). The alignment has been made partly manually and partly automatically with the monolingual alignment algorithm described in Nelken and Shieber (2006), which relies on a cosine similarity between sentence vectors weighted with the *tf-idf*. This program outputs alignments between sentences, along with a confidence score. Among these files, twenty articles or excerpts from Wikipedia were selected along with their equivalent in Vikidia. This amounts to 72 sentences for the former and 80 sentences for the latter.

The second corpus is composed of 16 narrative texts, and more specifically tales, by Perrault, Maupassant, and Daudet. We used tales since their simplified version was closer to the original than those of longer novels, which made the sentence alignment simpler. The simplified versions of these tales were found in two collections intended to learners of French as a foreign language (FFL): "Hachette - Lire en français facile" and "De Boeck - Lire et s'entrainer". Their level of difficulty ranges from A1 (Daudet) to B1 (Maupassant) on the CEFR scale (Council of Europe, 2001), with Perrault being A2. The texts were digitized by OCR processing and manually aligned, by two annotators, with an adjudication phase for the disagreement cases. In this corpus, we analysed 83 original sentences and their corresponding 98 simplified versions, which gives us a size roughly similar to the Wikipedia-Vikidia corpus.

The two corpora created are relevant for a manual analysis, as done in the next section, but they are too small for automatic processing. We plan to implement a method to align automatically the narrative texts in the near future and thus be able to collect a larger corpus.

### 2.2 Simplification Typology

The observations carried out on these two corpora have made it possible to establish a typology organised according to three main linguistic levels of transformation: lexical, discursive and syntactic, which can be further divided into subcategories. It is worth mentioning that in previous work, simplification is commonly regarded as pertaining to two categories of phenomena: lexical and syntactic (Carroll et al., 1999; Inui et al., 2003; De Belder and Moens, 2010). Little attention has been paid to discourse in the area of automatic simplification (Siddharthan, 2006).

The typology is summarized in Table 1. As regards the lexicon, the phenomena we observed involve four types of substitution. First, difficult terms can be replaced by a synonym or an hypernym perceived as simpler. Second, some anaphoric expressions, considered simpler or more explicit, are preferred to their counterparts in the original texts. For example, in our three tales, simplified nominal anaphora are regularly used instead of pronominal anaphora. Third, rather than using synonymy, the authors of the simplified texts sometimes replace difficult words with a definition or an explanatory paraphrase. Finally, in the particular case where the original texts contain concepts in a foreign language, these non-French terms are translated.

At the discourse level, the authors of simple texts pay particular attention to the organization of the information which has to be clear and concise. To this end, clauses may be interchanged to ensure a better presentation of the information. In addition, information of secondary importance can be removed while explanations or examples are added for clarity. These two phenomena can appear to be contradictory (deletion and addition), but they actually operate in a common goal: make the main information more comprehensible. Particular attention is also placed on the coherence and cohesion of the text: Authors tend to explain the pronouns and explicit the relations between sentences. The last observed strategy is that impersonal structures are often personalized.

Finally, at the syntactic level, five types of changes are observed: tense modification, deletion, modification, splitting and grouping. The last two types can be considered together since they are two opposite phenomena.

- First, the tenses used in the simplified versions are more common and less literary than those used in the original texts. Thus, the present and present perfect are preferred to the simple past, imperfect and past perfect.

[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

| Lexicon | Discourse | Syntax |
|---|---|---|
| Translation | Reorganisation | Tense |
| Anaphoric synonyms | Addition | Modification |
| Definition and paraphrase | Deletion | Grouping |
| Synonym or hypernym | Coherence and cohesion | Deletion |
| | Personalisation | Splitting |

Table 1: Typology of simplifications

- Secondary or redundant information, that is generally considered removable at the syntactic level, is not included in the simplified texts. Adverbial clauses, some adverbs and adjectives and subordinate clauses, among others, are omitted.
- When some complex structures are not deleted, then they are often moved or modified for better clarity. Such structures include negative sentences, impersonal structures, indirect speech and subordinate clauses.
- The authors sometimes choose to divide long sentences or conversely merge several sentences into one. The grouping of elements is much less frequent than the division of sentences. To split a sentence, the authors generally transform a secondary clause–be it relative, coordinate, subordinate, participial or adjectival–into an independent clause.

This classification can be compared with that of Medero et al. (2011) who propose three categories – division, deletion and extension – or that of Zhu et al. (2010), which includes division, deletion, reorganization, and substitution.

Among those transformations, some are hardly implementable. This is the case when a change requires the use of semantics. For example, noun modifiers may sometimes be removed, but in other cases, they are necessary. However, there are often neither typographical nor grammatical marked differences between the two cases.

Another issue is that other syntactic changes should be accompanied by lexical transformations, which are difficult to generalize. For example, transforming a negative sentence into its affirmative equivalent requires to find a verb whose affirmative form includes the meaning of the negative construction to replace.

There are also changes that are very particular and require a manual rather than an automatic processing of the text, in the sense that each case is different (even if part of a more global rule). In addition, they usually involve discourse or lexical information and not just syntactic one.

Finally, the syntactic changes impacting other parts of the text or concerning elements that depend on another structure require more comprehensive changes to the text. Therefore, they are also difficult to handle automatically. Thus, to change the tense of a verb in a sentence, we must ensure that the sequence of tenses agree in the entire text.

### 2.3 The Sentence Simplification System

We used this typology to implement a system of syntactic simplification for French sentences. The simplification is performed as a two-step process. First, for each sentence of the text, we generate the set of all possible simplifications (overgeneration step), and then, we select the best subset of simplified sentences using several criteria.

#### 2.3.1 Generation of the Simplified Sentences

The sentence overgeneration module is based on a set of rules (19 rules), which rely both on morpho-syntactic features of words and on syntactic relationships within sentences. To obtain this information, the texts from our corpus are analyzed by MELT [4] (Denis and Sagot, 2009) and Bonsai [5] (Candito et al., 2010) during a preprocessing phase. As a result, texts are represented as syntax trees that include the information necessary to apply our simplification rules. After preprocessing, the set of simplification rules is applied recursively, one sentence at a time, until there is no further structure to simplify. All simplified sentences produced by a given rule are saved and gathered in a set of variants.

The rules for syntactic simplification included in our program are of three kinds: deletion rules (12 rules), modification rules (3 rules) and splitting rules (4 rules). With regards to our typology, it can be noted that two types of rules have not been implemented: aggregation rules and tense simplification rules. The merging strategies (in which several sentences are aggregated into one) were

[4] https://gforge.inria.fr/projects/lingwb

[5] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

not observed consistently in the corpus. Moreover, aggregation rules could have come into conflict with the deletion rules, since they have opposite goals. Concerning tense aspects, some of them are indeed more likely to be used than others in Vikidia. However, this strategy has not been implemented, since it implies global changes to the text. For instance, when a simple past is replaced by a present form, we must also adapt the verbs in the surrounding context in accordance with tense agreement. This requires to consider the whole text, or at least the paragraph that contains the modified verbal form, and be able to automatically model tense agreement. Otherwise, we may alter the coherence of the text and decrease its readability.

This leaves us with 19 simplification rules.[6] To apply them, the candidate structures for simplification first need to be detected using regular expressions, via `Tregex` [7](Levy and Andrew, 2006) that allows the retrieval of elements and relationships in a parse tree. In a second step, syntactic trees in which a structure requires simplification are modified according a set of operations implemented through `Tsurgeon`.

The operations to perform depend on the type of rules:

1. For the deletion cases, simply deleting all the elements involved is sufficient (via the `delete` operation in Tsurgeon). The elements affected by the deletion rules are adverbial clauses, clauses between brackets, some of the subordinate clauses, clauses between commas or introduced by words such as "comme" (*as*), "voire" (*even*), "soit" (*either*), or similar terms, some adverbs and agent prepositional phrases.
2. For the modification rules, several operations need to be combined: some terms are dropped (via `Tsurgeon delete`), others are moved (operation `Tsurgeon move`) and specific labels are added to the text to signal a possible later processing. These labels are useful for rules implying a modification of tense or mode aspects for a verb. In such cases, tags are added around the verb to indicate that it needs to be modified. The modification is performed later, using the conjugation system `Verbiste`.[8] For instance, to change a passive into an active structure, not only the voice must be changed, but sometimes also the person, so that the verb agrees well with the agent that has become the new subject. As regards modification rules, three changes were implemented: moving adverbial clauses at the beginning of the sentence, transforming passive structures into active forms, and transforming a cleft to a non-cleft.
3. For the splitting rules, we followed a two-step process. The subordinate clause is first deleted, while the main clause is saved as a new sentence. Resuming from the original sentence, the main clause is, in turn, removed to keep only the subordinate clause, which must then be transformed into an independent clause. In general, the verbal form of the subordinate clause needs to be altered in order to operate as a main verb. Moreover, the pronoun governing the subordinated clause must be substituted with its antecedent and the subject must be added when missing. In the case of a relative clause, the relative pronoun thus needs to be substituted by its antecedent, but it is also important to consider the function of the pronoun to find out where to insert this antecedent. Our splitting rules apply when a sentence includes either relative or participle clauses, or clauses introduced by a colon or a coordinating conjunction.

All these simplification rules are applied recursively to a sentence until all possible alternatives have been generated. Therefore, it is common to have more than one simplified variant for a given sentence. In this case, the next step consists in selecting the most suitable variant to substitute the original one. The selection process is described in the next section.

### 2.3.2 Selection of the Best Simplifications

Given a set of candidate simplified sentences for a text, our goal is to select the best subset of simplified sentences, that is to say the subset that maximizes some measure of readability. More precisely, text readability is measured through different criteria, which are optimized with an Integer Linear Programming (ILP) approach (Gillick and Favre, 2009). These criteria are rather simple in

[6] These 19 rules are available at `http://cental.fltr.ucl.ac.be/team/lbrouwers/rules.pdf`

[7] `http://nlp.stanford.edu/software/tregex.shtml`

[8] This software is available at the address `http://sarrazip.com/dev/verbiste.html` under GNU general public license and was developed by Pierre Sarrazin.

this approach. They are used to ensure that not only the syntactic difficulty, but also the lexical complexity decrease, since syntactic transformations may cause lexical or discursive alterations in the text.

We considered four criteria to select the most suitable sentences among the simplified set: sentence length (in words) ($h_w$), mean word length (in characters) in the sentence ($h_s$), familiarity of the vocabulary ($h_a$), and presence of some keywords ($h_c$). While the first two criteria are pretty obvious as regards implementation, we measured word familiarity based on Catach's list (1985).[9] It contains about 3,000 of the most frequent words in French, whose spelling should be taught in priority to schoolchildren. The keywords were in this study simply defined as any term occurring more than once in the text.

These four criteria were combined using integer linear programming as follows: [10]

$$\begin{array}{ll} \text{Maximize :} & h_w + h_s + h_a + h_c \\ \text{Where :} & h_w = \text{wps} \times \sum_i s_i - \sum_i l_i^w s_i \\ & h_s = \text{cpw} \times \sum_i l_i^w s_i - \sum_i l_i^c s_i \\ & h_a = \text{aps} \times \sum_i s_i - \sum_i l_i^a s_i \\ & h_c = \sum_j w_j c_j \\ \text{Subject to:} & \sum_{i \in g_k} s_i = 1 \; \forall g_k \\ & s_i \text{occ}_{ij} \leq c_j \; \forall i, j \\ & \sum_i s_i \text{occ}_{ij} \geq c_j \; \forall j \end{array} \quad (1)$$

The above variables are defined as follows:

- `wps`: desired (mean) number of words per sentence
- `cpw`: desired (mean) number of characters per word
- `aps`: desired (mean) number of words absent from Catach's list for a sentence
- $s_i$: binary variable indicating whether the sentence $i$ should be kept or not, with $i$ varying from 1 to the total number of simplified sentences
- $c_j$: binary variable indicating whether keyword $j$ is in the simplification or not, with $j$ varying from 1 to the total number of keywords
- $l_i^w$: length of sentence $i$ in words
- $l_i^c$: number of characters in sentence $i$
- $l_i^a$: number of words absent from Catach's list in sentence $i$
- $w_j$: number of occurrences of keyword $j$
- $g_k$: set of simplified sentences obtained from the same original sentence $k$
- $occ_{ij}$: binary variable indicating the presence of term $j$ in sentence $i$

`wps`, `cpw` and `aps` are constant parameters whose values have been set respectively to 10, 5 and 2 for this study. 5 for `cpw` corresponds to the value computed on the Vikidia corpus, while for `wps` and `aps`, lower values than observed were used to force simplification (respectively 10 instead of 17 and 2 instead of 31).

However, these parameters may vary depending on the context of use and the target population, as they determine the level of difficulty of the simplified sentences obtained.

The constraints specify that (i) for each original sentence, at most one simplification set should be chosen, (ii) selecting a sentence means selecting all the terms it contains and (iii) selecting a keyword is only possible if it is present in at least one selected sentence.

We illustrate this process with the Wikipedia article entitled *Abel*. This article contains 25 sentences, from which 67 simplified sentences have been generated. For the original sentence (1a) for example, 5 variants were generated and simplification (2) was selected by ILP.

(1a) Original sentence[11] : *Caïn, l'aîné, cultive la terre et Abel ( étymologie : de l'hébreu ≪ souffle ≫, ≪ vapeur ≫, ≪ existence précaire ≫) garde le troupeau.*

(1b) Possible simplifications :
Simplification 1 : *Caïn, l'aîné, cultive la terre et Abel garde le troupeau.*
Simplification 2 : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*
Simplification 3 : *Caïn, l'aîné, cultive la terre.*
Simplification 4 : *Abel garde le troupeau.*
(...)

(1c) Selected simplification (2) : *Caïn, l'aîné, cultive la terre. Abel garde le troupeau.*

## 3 Evaluation

Syntactic simplification involves substantial changes within the sentence both in terms of contents and form. It is therefore important to check that the application of a rule does not cause errors that would make the sentences produced unintelligible or ungrammatical. A manual evaluation of our system's efficiency to generate correct simplified sentences was carried out on our two corpora. In each of them, we selected a set of texts that had not been previously used for

[9] This list is available at the site `http://www.ia93.ac-creteil.fr/spip/spip.php?article2900`.

[10] We used an ILP module based on `glpk` that is available at the address `http://www.gnu.org/software/glpk/`

[11] *Caïn, the eldest brother, farms the land and Abel (etymology : from Hebrew ≪ breath ≫, ≪ steam ≫, ≪ fragile existence ≫) looks after the flock.*

| | Sentence length | Word length | Word familiarity | Keywords |
|---|---|---|---|---|
| Expected values | 10 | 5 | 2 | / |
| Original | 19 | 6.1 | 11 | 5 |
| Simplification 1 | 11 | 4.3 | 5 | 5 |
| **Simplification 2** | **5** | **4.6** | **2** | **5** |
| Simplification 3 | 6 | 4.5 | 3 | 3 |
| Simplification 4 | 4 | 4.7 | 2 | 2 |
| Simplification 5 | 9 | 6.3 | 5 | 5 |
| Simplification 6 | 12 | 7.3 | 8 | 2 |

Table 2: Values of the criteria in IPL for example (1).

the typological analysis, that is to say 9 articles from Wikipedia (202 sentences) and two tales from Perrault (176 sentences). In this evaluation, all simplified sentences are considered, not only those selected by ILP. The results are displayed in Table 3 and discussed in Section 3.1. Two types of errors can be detected: those resulting from morpho-syntactic preprocessing, and particularly the syntactic parser, and the simplification errors *per se*, that we discuss in larger details in Section 3.2.

### 3.1 Quantitative Evaluation

Out of the 202 sentences selected in the informative corpus for evaluation, 113 (56%) have undergone one or more simplifications, which gives us 333 simplified variants. Our manual error analysis revealed that 71 sentences (21%) contain some errors, among which we can distinguish those due to the preprocessing from those actually due to the simplification system itself. It is worth mentioning that the first category amounts to 89% of the errors, while the simplification rule are only responsible for 11% of those. We further refined the analysis of the system's errors distinguishing syntactic from semantic errors.

The scores obtained on the narrative corpus are slightly less good: out of the 369 simplified variants produced from the 154 original sentences, 77 (20.9%) contain errors. This value is very similar to the percentage for the informative corpus. However, only 50.7% of these errors are due to the preprocessing, while the remaining 49.3% come from our rules. It means that our rules yield about 10.3% incorrect simplified variants compared to 2.7% for the informative corpus. Nevertheless, these errors are caused mostly by 2 or 3 rules: the deletion of subordinate clauses, of infinitives or of clauses coordinated with a colon. This loss in efficiency can be partly explained by the greater presence of indirect speech in the tales that include more non-removable subordinate clauses, difficult to distinguish from removable clauses.

Globally, our results appear to be in line with those of similar systems developed for English.[12] Yet, few studies have a methodology and evaluation close enough to ours to allow comparison of the results. Siddharthan (2006) assessed his system output using three judges who found that about 80% of the simplified sentences were grammatical, while 87% preserved the original meaning. These results are very similar to our findings that mixed the syntactic and discourse dimensions. Drndarević et al. (2013) also presented the output of their system to human judges who estimated that 60% of the sentences were grammatical and that 70% preserved the initial meaning. These scores appear lower than ours, but Drndarević et al. also used lexical rules, which means that their error rate includes both grammatical and lexical errors.

### 3.2 Error Analysis

As regards syntax, the structure of a sentence can be modified so that it becomes grammatically incorrect. Three simplification rules are concerned. Deletion rules may cause this kind of problem, because they involve removing a part of the sentence, considered as secondary. However, sometimes the deleted element is essential, as in the case of the removal of the referent of a pronoun. This type of problem arises both with the deletion of a subordinate clause or that of an infinitive clause. Deletion rules are also subject to a different kind of errors. During the reconstruction of the sentence resulting from the subordinate clause, some constituents, such as the subject, may not be properly identified and will be misplaced in the new sentence.

At the semantic level, the information conveyed by the original sentence may be modified or even removed. When an agent or an infinitive clause

[12] We do not discuss French here, since no simplification system were found for French, as explained previously.

| Wikipedia-Vikidia corpus | | | | |
|---|---|---|---|---|
| **nb. sent.** | **% correct** | **% preproc. errors** | **% simplification errors** | |
| 333 | 262 (78.7 %) | 63 (18.9%) | 8 (2.4 %) | |
| | | | syntax: 6 (1.8%) | semantics: 2 (0.6%) |
| **Narrative corpus** | | | | |
| **nb. sent.** | **% correct** | **% preproc. errors** | **% simplification errors** | |
| 369 | 292 (79.1 %) | 39 (10.6%) | 38 (10.3 %) | |
| | | | syntax: 20 (5.4%) | semantics: 18 (4.9%) |

Table 3: Performance of the simplification system on both corpora

are suppressed, the meaning of the sentence may be disrupted or some of the content lost. For instance, in the following sentence – extracted from the Wikipedia article *abbé* (abbot) – the infinitive clause explaining the term is dropped:

> (2a) *C'est aussi depuis le XVIIIe siècle le terme en usage pour désigner un clerc séculier ayant au moins reçu la tonsure.*[13]
> (2b) *C'est aussi depuis le XVIIIe siècle le terme en usage.*

To fix the errors identified above, our rules should be refined and developed, with the addition of better tools for sentence regeneration as well as some exclusion criteria for the incorrect sentences within the ILP module, as discussed in the next section.

## 4 Perspectives and Conclusions

This article describes an automatic syntactic simplification system for French intended for children and language learners. It is based on a set of rules defined after a corpus study, which also led to the development of a typology of simplifications in French. It would be easy to extend our typology to other target users based on other appropriate corpora, such as people with language disorders.

Our approach also uses the technique of overgeneration, which makes it possible to retain the best set of simplifications based on readability criteria. Note that among those employed, some had not been considered previously and produce interesting results. Finally, we showed that the performance of our system is good (about 80 % of the generated sentences are correct) and in line with previous studies.

[13] *It is also the term in use since the 18th to refer to a secular cleric who, at least, received the tonsure.*

The evaluation showed that the rules implemented are more suitable for expository texts, probably because they are more explicit, as style there is of a minor importance. In addition, the system set up was first tested on and therefore adapted to Wikipedia. It was only subsequently applied to narratives, that revealed new challenges, especially concerning the deletion rules. The information provided in secondary clauses or complements indeed seems most essential to understanding the story, especially when it comes to direct or indirect speech. In order to comprehend the differences in terms of efficiency and rules to be applied between genres, it would be necessary to extend our study to other texts collected in the corpora.

We envision multiple perspectives to improve our system. First, syntactic simplification could be supplemented by lexical simplification, as is done in some studies for English (Woodsend and Lapata, 2011). Moreover, our error analysis has highlighted the need to add or repeat words when a sentence is split. It would therefore be useful to use a tool that manages references in order to improve the quality of simplified text. In addition, the sentence selection module could include additional selection criteria, based on the work done in readability of French (François and Fairon, 2012). A final perspective of improvement would be to make the rule system adapt to the target audience and the genre of the texts. This would require assessing the relevance of various transformations and selection criteria of the best simplifications. This perspective would also require assessing the effectiveness of the rules by means of comprehension tests both on the original and simplified sentences, which we plan to do.

## References


S. Aluísio, L. Specia, T. Pardo, E. Maziero, and R. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

E. Barbu, P. de Las Lagunillas, M. Martın-Valdivia, and L. Urena-López. 2013. Open book: a tool for helping asd users' semantic comprehension. *NLP4ITA 2013*, pages 11–19.

S. Bott, L. Rello, B. Drndarevic, and H. Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING 2012*, pages 357–374.

M. Candito, B. Crabbé, and P. Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1840–1847.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying Text for Language-Impaired Readers. In *Proceedings of EACL*, pages 269–270.

N. Catach. 1985. *Les listes orthographiques de base du français*. Nathan, Paris.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics*, pages 1041–1044.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

J. De Belder and M.-F. Moens. 2010. Text Simplification for Children. In *Proceedings of the Workshop on Accessible Search Systems*.

P. Denis and B. Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC*.

B. Drndarević, S. Štajner, S. Bott, S. Bautista, and H. Saggion. 2013. Automatic text simplification in spanish: a comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500.

R. Evans and C. Orăsan. 2013. Annotating signs of syntactic complexity to support sentence simplification. In *Text, Speech, and Dialogue*, pages 92–104.

T. François and C. Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of EMNLP 2012*, pages 466–477.

C. Gasperin, E. Maziero, L. Specia, T. Pardo, and S. Aluisio. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

D. Gillick and B. Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*.

M. Heilman and N. A. Smith. 2010. Extracting Simplified Statements for Factual Question Generation. In *Proceedings of the 3rd Workshop on Question Generation*.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

S. Jonnalagadda, L. Tari, J. Hakenberg, C. Baral, and G. Gonzalez. 2009. Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. In *Proceedings of NAACL-HLT 2009*.

R. Levy and G. Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC*, pages 2231–2234.

L. Lin and W. J. Wilbur. 2007. Syntactic sentence compression in the biomedical domain: facilitating access to related articles. *Information Retrieval*, 10(4):393–414, October.

J. Medero and M. Ostendorf. 2011. Identifying Targets for Syntactic Simplification. In *Proceedings of the SLaTE 2011 workshop*.

R. Nelken and S.M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of EACL*, pages 161–168.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of SLaTE2007*, pages 69–72.

L. Rello, C. Bayarri, A. Gòrriz, R. Baeza-Yates, S. Gupta, G. Kanvinde, H. Saggion, S. Bott, R. Carlini, and V. Topac. 2013. Dyswebxia 2.0!: more accessible text for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 25.

J.F. Richard, J. Barcenilla, B. Brie, E. Charmet, E. Clement, and P. Reynard. 1993. Le traitement de documents administratifs par des populations de bas niveau de formation. *Le Travail Humain*, 56(4):345–367.

H. Saggion, E. Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text simplification in simplext. making text more accessible. *Procesamiento del lenguaje natural*, 47:341–342.

V. Seretan. 2012. Acquisition of syntactic simplification rules for french. In *LREC*, pages 4019–4026.

A. Siddharthan. 2006. Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109, jun.

L. Specia. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (Propor-2010).*, pages 30–39.

S. Williams and E. Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(4):495–525.

J.D. Willms. 2003. Literacy proficiency of youth: Evidence of converging socioeconomic gradients. *International Journal of Educational Research*,

39(3):247–252.

K. Woodsend and M. Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of EMNLP*, pages 409–420.

Z. Zhu, D. Bernhard, and I. Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of COLING 2010*, pages 1353–1361.