

# Coreference-Based Text Simplification

Rodrigo Wilkens, Bruno Oberle, Amalia Todirascu

LiLPa – University of Strasbourg

rswilkens@gmail.com, b.oberle@zoho.eu, todiras@unistra.fr

## Abstract

Text simplification aims at adapting documents to make them easier to read by a given audience. Usually, simplification systems consider only lexical and syntactic levels, and, moreover, are often evaluated at the sentence level. Thus, studies on the impact of simplification in text cohesion are lacking. Some works add coreference resolution in their pipeline to address this issue. In this paper, we move forward in this direction and present a rule-based system for automatic text simplification, aiming at adapting French texts for dyslexic children. The architecture of our system takes into account not only lexical and syntactic but also discourse information, based on coreference chains. Our system has been manually evaluated in terms of grammaticality and cohesion. We have also built and used an evaluation corpus containing multiple simplification references for each sentence. It has been annotated by experts following a set of simplification guidelines, and can be used to run automatic evaluation of other simplification systems. Both the system and the evaluation corpus are freely available.

**Keywords:** Automatic Text Simplification (ATS), Coreference Resolution, French, Annotated and evaluation corpus

## 1. Introduction

Text cohesion, a crucial feature for text understanding, is reinforced by explicit cohesive devices such as coreference (expressions referring to the same discourse entity: *Dany Boon—the French actor—his film*) and anaphoric (an anaphor and its antecedent: *it—the fox*) chains. Coreference chains involves at least 3 referring expressions (such as proper names, noun phrases (NP), pronouns) indicating the same discourse entity (Schneidecker, 1997), while anaphoric chains involves a directed relation between the anaphor (the pronoun) and its antecedent. However, coreference and anaphora resolution is a difficult task for people with language disabilities, such as dyslexia (Vender, 2017; Jaffe et al., 2018; Sprenger-Charolles and Ziegler, 2019). Moreover, when concurrent referents are present in the text, the pronoun resolution task is even more difficult (Givón, 1993; McMillan et al., 2012; Li et al., 2018): the pronouns may be ambiguous and their resolution depends on user knowledge about the main topic (Le Bouëdec and Martins, 1998). This poses special issues to some NLP tasks, such as text simplification.

Automatic text simplification (ATS) adapts text for specific target audience such as L1 or L2 language learners or people with language or cognitive disabilities, as autism (Yaneva and Evans, 2015) and dyslexia (Rello et al., 2013). Existing simplification systems work at the lexical or syntactic level, or both. Lexical simplification aims to replace complex words by simpler ones (Rello et al., 2013; François et al., 2016; Billami et al., 2018), while syntactic simplification transforms complex structures (Seretan, 2012; Brouwers et al., 2014a). However, these transformations change the discourse structure and might violate some cohesion or coherence constraints.

Problems appear at discourse level because of lexical or syntactic simplifications ignoring coreference. In the following example, the substitution of *hyène* ‘hyena’ by *animal* ‘animal’ introduces an ambiguity in coreference resolution, since the animal might be *le renard* ‘the fox’ or *la hyène* ‘the hyena’.

Original: *Le renard se trouvait au fond du puits et appelait.*

*La hyène l’approcha.* ‘The fox was at the bottom of the well. The hyena approached it.’

Simplified: **Le renard se trouvait au fond du puits. L’animal l’approcha.** ‘The fox was at the bottom of the well. The animal approached it.’

However, few existing syntactic simplification systems (e.g. Siddharthan (2006) and Canning (2002)) operate at the discourse level and replace pronouns by antecedents or fix these discourse constraints after the syntactic simplification process (Quiniou and Daille, 2018).

In this paper, we evaluate the influence of coreference in the text simplification task. In order to achieve this goal, we propose a rule-based text simplification architecture aware of coreference information, and we analyse its impact at the lexical and syntactic levels as well as for text cohesion and coherence. We also explore the use of coreference information as a simplification device in order to adapt NP accessibility and improve some coreference-related issues. For this purpose, we have developed an evaluation corpus, annotated by human experts, following discourse-based simplification guidelines.

This paper is organised as follows. We present related work on cohesion markers such as coreference chains, as well as lexical and syntactic simplification systems that take into account these elements (Section 2). Then, we present the architecture of our rule-based simplification system alongside the corpus used to build it and the corpus used to evaluate it (Section 3). The rules themselves, and the system evaluation, are presented in Section 4. Finally, Section 5 presents final remarks.

## 2. Related Work

Few systems explore discourse-related features (e.g. entity densities and syntactic transitions) to evaluate text readability alongside other lexical or morphosyntactic properties (Štajner et al., 2012; Todirascu et al., 2016; Pitler and Nenkova, 2008).

Linguistic theories such as Accessibility theory (Ariel, 2001) organise referring expressions and their surface forms into a hierarchy that predicts the structure of cohe-

sion markers such as coreference chains. In this respect, a new discourse entity is introduced by a low accessibility referring expression, such as a proper noun or a full NP. On the contrary, pronouns and possessive determiners are used to recall already known entities. This theory is often used to explain coreference chain structure and properties (Todirascu et al., 2017). Other linguistic theories such as Centering theory (Grosz et al., 1995) predict discourse centres following a typology of centre shift or maintenance and explains linguistic parameters related to coherence issues. Simplification systems frequently ignore existing cohesive devices. This aspect is however taken into account by, for instance, Siddharthan (2006), Brouwers et al. (2014a) and Quiniou and Daille (2018). Canning (2002) replaces anaphor by their antecedent for a specific target audience. Siddharthan (2004) first uses anaphora detection to replace pronouns by NP. Then a set of ordered hand-made syntactic rules is applied (e.g. conjunctions are simplified before relative clauses). Rhetorical Structure Theory (Mann and Thompson, 1988) is used to reorder the output of the syntactic simplification and anaphoric relations are checked after simplification. Moreover, Siddharthan (2006) proposes a model based on Centering theory (Grosz et al., 1995) to recover broken cohesion relations, by using a specific pronoun resolution system for English. The model allows the replacement of a pronoun by its immediate antecedent. Few systems use a coreference resolution module to solve coreference issues (Barbu et al., 2013). For French, Quiniou and Daille (2018) develop a simple pronoun resolution module, inspired by (Mitkov, 2002) (e.g. searching antecedents in two sentences before the pronoun). This system previously detects expletive pronouns to exclude them from pronoun resolution. Brouwers et al. (2014a) mainly propose syntactic simplification using hand-made rules implemented with Tregex and Tsurgeon (Levy and Andrew, 2006). The only rules handling anaphora replace pronouns with NP from the previous or the current sentence. To sum up, only a few ATS approaches, mostly for English, propose discourse simplification rules or rules checking discourse constraints.<sup>1</sup>

### 3. Methodology

Taking into account our goal of analysing the impact of coreference in text simplification, we compiled two different types of corpora (one for the evaluation and other for the simplification reference), described in Section 3.1., we propose a coreference-aware architecture in Section 3.2.

#### 3.1. Corpora

One of the most critical elements in text simplification is the target audience since it defines what types of operations should be performed. In this regard, we compiled a reference corpus composed of parallel texts manually adapted for dyslexic children in the context of the Methodolodys association<sup>2</sup>. This corpus consists of five manually adapted

paired tales (1,143 words and 84 sentences for the dyslexic texts and 1,969 words and 151 sentences for the original texts). This corpus helps us to better understand simplifications targeting dyslexic children both for coreference chains and at the lexical and syntactic levels.

The reference corpus has been preprocessed in two steps. First, we aligned the corpus using the MEDITE tool (Fenoglio and Ganascia, 2006). This process identifies the transformations performed (phrase deletion or insertion, sentence splitting, etc.) as well as their level (i.e. lexical, syntactic or discourse). The second step consisted in the manual annotation of coreference chains (mentions and coreference relations) (Todirascu et al., 2017; Schnedecker, 1997; Todirascu et al., 2016) and referring expressions accessibility (Ariel, 1990; Ariel, 2001). Then, we compared coreference chains properties: chain size, average distance between mentions, lexical diversity (with the stability coefficient defined by Perret (2000)), annotation (mention density, link (relation between consecutive mentions) count and density, grammatical categories of the mentions).

The reference corpus provides several meaningful descriptions of the simplification phenomenon. However, it is limited in the sense of system evaluation since it provides only one valid simplification, and it may require resources other than those currently available in NLP technology. In order to build an evaluation corpus, we manually collected simplified alternatives to the original texts (3 texts from the reference corpus and 2 new texts). We used the online PsyToolkit tool<sup>3</sup>, and 25 annotators (master students in linguistics and computational linguistics) participated. They all provided information on age, mother tongue and education level, and replied to questionnaires to check reading time and text understanding. Additionally, we summarised the discursive observations identified in the reference corpus (presented in Section 4.1. and 4.2.) as simplification guidelines<sup>4</sup> provided to the annotators. The purpose of these guidelines was to drive the annotators' attention to discourse operations.

To create an evaluation corpus, the students proposed simplified alternatives to texts from the original corpus (we replaced 2 texts to broaden the text coverage). These alternatives had to follow the provided guidelines, but the students could also suggest other simplification proposals. Taking into account the task complexity and the time required to simplify a text, we ask them to simplify only some short paragraphs (894 words per person on average). We excluded from our data the responses from 6 students who did not fully understand the task. We aligned the source text and then, we identified ungrammatical transformations and typos, and replaced these answers with the original text. The evaluation corpus also offers complementary simplifications for each text. Thus, it can also be used to select the most significant simplifications required. We obtained several simplified versions for each sentence. The analysis of the simplifications performed in both reference and evaluation corpora is presented in Section 4. Furthermore, the

dyslexic children.

<sup>3</sup>psytoolkit.org

<sup>4</sup>The guidelines are available on the Web site of Alector project <https://alectorsite.wordpress.com/>

<sup>1</sup>For an overview of simplification studies, including systems for different needs arguing for discourse phenomena processing, see Saggion (2017).

<sup>2</sup>methodolodys.ch/ is an association providing texts and exercises to improve reading and comprehension skills for

system that uses the result of this analysis is introduced in Section 3.2., and its evaluation is presented in Section 4.3.

## 3.2. Architecture

In this paper, we propose a rule-based approach to the simplification task since, on one hand, the original and simplified parallel corpora are small, which makes applying machine learning methods difficult; and, on the other hand, this kind of approach allows us to study the impact of each type of transformation on the comprehension and reading capabilities of the target audience. In this case, it is possible to decompose the simplification rules into various levels and to evaluate them separately. In this work, we are particularly interested in discursive simplification, which aims at preserving textual cohesion markers, such as coreference chains (Schneidecker, 2017).

The proposed architecture is composed of four modules, as illustrated in Figure 1. The first module preprocesses the text in order to facilitate the application of the simplification rules. This module starts by annotating the text with a parser (Qi et al., 2019) and with coreference information, which consist of the delimitation of referring expressions (e.g. proper names or named entities, NP and pronouns) and identification of coreference relations between these expressions. It is based on the architecture proposed by Kantor and Globerson (2019) but trained on the DEMOCRAT corpus<sup>5</sup> (Landragin, 2016). Our trained model achieved 85,04% of CoNLL score (the standard evaluation metric for automatic coreference resolution) with predicted mentions<sup>6</sup>. The syntactic simplification module is inspired by the work of Siddharthan (2003) and Brouwers et al. (2014b), applying deletion and rewriting rules described in the next sections. Then the data is processed by the third module, the discursive simplification module, which modifies the structure of coreference chains detected by the first module. Finally, the last module applies lexical and morphological simplifications by replacing words. This module is based on ReSyf (Billami et al., 2018) and its API<sup>7</sup>, which allows to query by the easiest alternative synonym to a given target word. Since ReSyf proposed different alternatives for each word sense, we selected as output only those that are the simplest in all senses and the most frequent across the senses.

To evaluate the system, taking advantage of the alternative simplification references in the evaluation corpus, we used the SARI measure that correlates to some level with human judgement of simplicity (Xu et al., 2016). Moreover, as a point of comparison, we also present the results of BLEU (Papineni et al., 2002), used in MT-based simplification methods.

A key element in our architecture is the rewriting tool (see Section 3.2.1.), that allows to search for both lexical and morphosyntactic patterns as well as to modify the syntactic parse structure.

<sup>5</sup>We trained the model on text from the 19th to the 21st century: 295,978 tokens, 81,506 relations, 43,211 chains.

<sup>6</sup>In NLP, a mention is a referring expressions.

<sup>7</sup>[gitlab.com/Cental-FR/resyf-package](https://gitlab.com/Cental-FR/resyf-package)

### 3.2.1. Text rewriting tool

The text rewriting tool applies several text transformations and changes to the structure of the sentences: deletion of secondary information, sentence splitting and phrase changes. However, we have to transform the text without violating the grammar. We compared available rewriting tools such as Tregex and Tsurgeon (Levy and Andrew, 2006), Semgrex (Chambers et al., 2007), and Semgrex-Plus (Tamburini, 2017).

Levy and Andrew (2006) provide tree query (Tregex) and manipulation (Tsurgeon) tools that can operate on constituent trees. Tree query tools have proven invaluable to NLP research for both data exploration and corpus-based research. Complementary to Tregex queries, Tsurgeon operates at node and edge levels, to change the structure of the trees, allowing, for example, node renaming, deletion, insertion, movement and replacement.

Chambers et al. (2007) proposed Semgrex to handle dependencies instead of constituents. The tool identifies semantic patterns supporting inference in texts as an alternative to writing graph traversal coded by hand for each desired pattern. Semgrex allows inter- and extra-dependency graphs relations. For instance, queries may be used to identify direct or indirect governor associations, with or without limitation of the distance between the elements, or even the node positional relation (e.g. immediately precedes, right sibling, right immediate sibling, same nodes).

Making a step forward into graph modification alike to Tsurgeon, Tamburini (2017) developed Semgrex-Plus to convert dependency treebanks into various formats. It supports three rewriting operations: replacing the tag of a graph node, and inserting or deleting a dependency edge between two graph nodes.

Additionally to those, generic graph processing tools might be adapted for our task. For instance, Bonfante et al. (2018) present GREW, a graph rewriting tool that can perform similar queries to Semgrex, while providing graph operations close to those proposed by Tsurgeon. However, as pointed by Tamburini (2017), intricacies of the generic tools might have a significant impact on the sentence rewriting process. For querying parsed data, we selected Semgrex because it precisely fits our needs. But, regarding the sentence rewriting goal, we opted to create a new Semgrex-based sentence processing tool, given the parser restrictions and the small set of operations available on Semgrex-Plus. Concerning the operations, we developed the following: (1) *Insert* injects a node (or tree) in another node; (2) *Delete* removes a node and its subtree from the sentence graph; (3) *Split* detaches a node and its subtree; (4) *Move* detaches a node and its subtree from a tree node, attaching it to another node of the same tree; (5) *Replace tag label* replaces the node information (e.g. surface and PoS-tag); (6) *Replace node* substitutes a node by another one; and (7) *Copy subgraph* creates a deep copy of a node or a tree.

The *insert*, *delete*, *move*, and *replace node* operations are directly based on Tsurgeon while *replace label* is based both on Tsurgeon and Tsurgeon-plus. The *split* method is inspired by the Tsurgeon *excise* and *adjoin* operations. On the contrary, the *copy* operation was developed because we needed to copy parts of a sentence into different trees. In

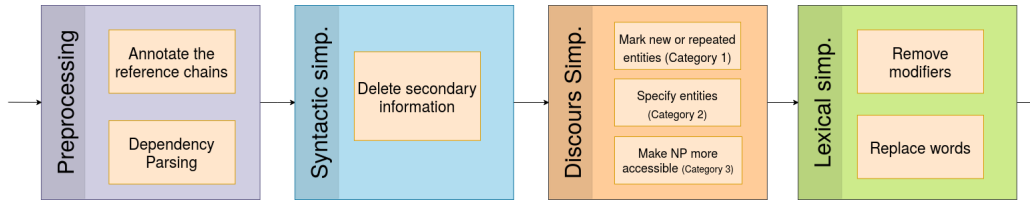


Figure 1: The architecture of the simplification system.

addition to these graph operations, we also extended Semgrex to read coreference information when available, and we simplified the morphology feature query by allowing to search by sub-elements without regular expressions. These operations are combined into rules in order to rewrite the text. We detail the process of defining the cohesion rules necessary for our discourse simplification system in Section 4.1. and Section 4.2.

## 4. Results

In this section, we present and explain the reference corpus analyses of cohesion changes that have been used to design simplification rules.

### 4.1. Cohesion changes during the simplification

The compilation of several observations from the reference corpus is presented in this section. It supports the discourse level transformations applied in this work. For our purpose, we use Accessibility theory (Ariel, 1990; Ariel, 2001), which proposes a hierarchy of referring expressions, from those with low accessibility (such as proper nouns or definite NP; these are usually newly introduced expressions) to highly accessible ones (such as pronouns or determiners; these have usually been introduced previously). Moreover, we use Centering theory (Grosz et al., 1995), which predicts situations when the attention centre shifts to a new one, resulting in a change of the syntactic function of the centre. By exploiting these observations, we propose three categories of rules. First, we present the results from our analysis, and then we detail these rules.

As discussed in Section 3.1., we manually enriched the reference corpus with the annotation proposed in Todirascu et al. (2017), Schnedecker (1997) and Todirascu et al. (2016). This properties are presented in Section 3.1. and in Table 1. Next, we compared these annotations in both the simplified and the original texts to find discourse simplification (cohesion) rules. The comparison between the coreference chain properties in the original and simplified texts is the first step to define the cohesion rules. Next, we have to identify changes in the structure of coreference chains induced by simplifications before defining the cohesion rules. We start our study of the cohesive elements by comparing the properties and transformations of five text pairs. Each of those was manually annotated with coreference chains. Due to the lack of available data containing original and simplified texts for dyslexic people, our corpus is relatively small when compared to others simplification corpora. Moreover, manual coreference annotation is a time-consuming and challenging task, in terms of referring ex-

pression identification (delimiting expressions and finding their type) and of chain identification (linking all the referring expressions belonging to the same chain).

The adapted texts present some specific coreference property statistically differences when compared to the original ones (Table 1): link count ( $p=0.01$ ), stability coefficient ( $p=0.01$ ), chain density ( $p=0.04$ ), link density ( $p=0.008$ ), and annotation density ( $p=0.02$ ). Additionally, the average distance between two consecutive referring expressions is higher in original than in adapted texts, as a consequence of text deletions.

We also observe interesting correlations for most of the properties (0.74 for link count, 0.81 for stability coefficient, 0.72 for chain density, and 0.74 for link density). We observe differences between original and adapted texts at the coreference level, but despite this, the correlations between the properties are still valid. Besides, a negative correlation (-0.717) is found between the length of the chains and the number of chains. In the adapted texts, longer chains are correlated with a lower number of chains (on average 10.62 against 7.0). Some referents were deleted in adapted versions, which explains this result.

Properties	Adapted	Original
Avg chain size	10.376	10.86
Avg link-to-link distance	14.550	11.920
Avg link length	1.500	1.450
Avg chain count	6.200	7.800
Avg link count	55.600	83.4
Avg chain density	0.012	0.009
Avg stability coefficient	0.607	0.471
Avg link density	0.113	0.093
Avg annotation density	0.162	0.139

Table 1: Coreference chains properties.

The composition of the chains varies with the complexity of the texts, as shown in Figure 2. In the simplified texts, the pronouns have been deleted or replaced by their referent: this explains that the percentage of personal pronouns (PRO.PER) included in coreference chains is larger in the original texts (36.5% of the mentions) than in the adapted texts (19.4%). This observation is in line with the significant difference for definite noun (NP.DEF) usage (36.0% in the simplified texts but only 18.7% in the original ones) or for proper noun (NP.NAM) usage (3.95% in simplified and 1.91% in original texts). The possessive determiners represent 10.1% in simpler texts but 12.9% in the original texts. This observation is related to our third category

(see at the end of the section), concerning possessive NP replacement by a specific referent. Moreover, concerning referring expression accessibility, we observed a significant change in determinant accessibility. This may be observed in the increase of indefinite NPs (NP . INDEF) from 3.11% to 5.03%, while demonstrative NPs (NP . DEM) decrease from 0.48% to 0.36% in simpler texts. This accessibility changing is related with our second category (see end of section), and it is exemplified in cases such as:

Original: *le<sub>1</sub> loup; cette<sub>2</sub> hyène.* ‘The<sub>1</sub> wolf; This<sub>2</sub> hyena’  
Simplified: **un<sub>1</sub> loup; la<sub>2</sub> hyène.** ‘A<sub>1</sub> wolf; The<sub>2</sub> hyena.’

Studying the stability coefficient<sup>8</sup> (Perret, 2000), we observed more stable chains in the dyslexic texts (0.47) than in the original texts (0.60). Thus, the coreference chains present less lexical variation (i.e. more repetitions) in the simple text versions than in the original ones. These observations support the first but also the second category of cohesion rules (see below).

To reduce coreference ambiguity, the pronoun *il* ‘it’ is replaced by the subject of the previous sentence (*le hérisson* ‘the hedgehog’):

Original: *Le hérisson voit le loup arriver, mais il<sub>1</sub> n’a pas le temps de se cacher.* ‘The hedgehog sees the wolf coming, but it<sub>1</sub> has no time to hide himself.’

Simplified: **Le hérisson voit le loup arriver, mais le hérisson<sub>1</sub> n’a pas le temps de se cacher.** ‘The hedgehog sees that the wolf arriving, but the hedgehog<sub>1</sub> has no time to hide himself.’

To reduce working memory, the repeated pronoun is replaced by the referent:

Original: *Le renard<sub>1</sub> avait très soif. Il<sub>2</sub> aperçut un puits. Sur la poulie, il y avait une corde, et, à chaque bout de la corde, il y avait un seau. Il<sub>3</sub> s’assit dans un des seaux et fut entraîné au fond. Heureux, il<sub>4</sub> but pendant de longues minutes.* ‘The fox<sub>1</sub> was very thirsty. It<sub>2</sub> saw a well. On the pulley, there was a rope, and at each end of the rope, there was a bucket. It<sub>3</sub> sat in one of the buckets and was dragged to the bottom. Happily, it<sub>4</sub> drank for long minutes.’

Simplified: **Le renard<sub>1</sub> avait très soif. Le renard<sub>2</sub> aperçut un puits. Sur la poulie, il y avait une corde, et, à chaque bout de la corde, il y avait un seau. Le renard<sub>3</sub> s’assit dans un des seaux et fut entraîné au fond. Heureux, le renard<sub>4</sub> but pendant de longues minutes.** ‘The fox<sub>1</sub> was very thirsty. The fox<sub>2</sub> saw a well. On the pulley there was a rope, and at each end of the rope there was a bucket. The fox<sub>3</sub> sat in one of the buckets and was dragged to the bottom. Happily, the fox<sub>4</sub> drank for long minutes.’

Moreover, we define rules from Category 2 to reflect differences between possessive determiners (12.95% vs 10.07%) and proper noun (1.91% vs 3.95%). For instance, the possessive NP (e.g. *son mari*) should be replaced by its referent (e.g. *M. Dupont*) in the example:

Original: *Mme Dupont a préparé sa soupe. Son mari<sub>1</sub> dit, pour la première fois, qu’il n’aime pas sa soupe.* ‘Mrs Dupont had prepared her soup. Her husband<sub>1</sub> says, for the first time, that he does not like her soup.’

Simplified: **Mme Dupont a fait sa soupe. M. Dupont<sub>1</sub> dit, pour la première fois, qu’il n’aime pas sa soupe.** ‘Mrs Dupont cooked her soup. Mr. Dupont<sub>1</sub> said for the first time that he does not like her soup.’

The reference corpus alignment also contains pronoun deletions due to the suppression of secondary information. For example, the relative pronoun *qui* ‘who’ and the personal pronoun *eux* ‘them’ were deleted because the relative clause *qui se dirigent vers eux* ‘who went to them’ was deleted. We add a rule to Category 3, concerning information suppression:

Original: *En chemin, ils aperçoivent, au loin, des bandits qui se dirigent vers eux.* ‘In their way, they saw, far away, bandits who went to them.’

Simplified: **En chemin, ils aperçoivent au loin des bandits.** ‘In their way, they saw, far away, bandits.’

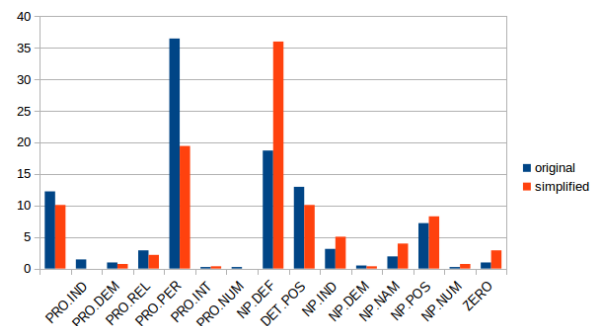


Figure 2: The distribution of referring expression types in the chains for original and simplified texts.

All the differences observed in the corpus are summarised in the three following categories, each one containing different rules.

**Category 1 mark new or repeated entities.** A referent should be found for ambiguous pronouns (i.e. several referents might be selected) or successive pronouns in the same chain. This operation decreases the number of processing inferences done by the reader to solve coreference relations.

**Category 2 specify entities.** New entities should be introduced by either an indefinite NP or a proper noun, while definite nouns phrases (formed with a definite article or a demonstrative determiner), being highly accessible, refer to known entities. The change of determiner for a more highly accessible one modifies the accessibility of the referring expression.

**Category 3 make NP more accessible.** Secondary information, such as relative or oblique clauses, should be removed. As a consequence, mentions of coreference chains are deleted (e.g. indefinite pronouns) as well as non-coreferent pronouns, such as *chacun*, *quelqu’un*. Possessive NPs are replaced by their explicit referent (a proper noun or another NP).

The rules have been written as simplification guidelines and applied by human annotators (Master students from Linguistics and Computational Linguistics) to create an evaluation corpus for our system.

<sup>8</sup>A low stability coefficient means that there is a large variety of referring expressions in a given chain, in terms of synonyms.

## 4.2. Proposed simplifications

Concerning the evaluation corpus, we annotated and ranked the multiple simplification references proposed by the annotators which followed the simplification guidelines. The proposals are not unanimous, in other words, there is not a single case in which all the annotators agreed with the simplification. Moreover, we observed several parts of the texts without any simplification suggestion from the annotators. These observations are also supported by the low value of 0.189 in the Krippendorff inter-annotator agreement, which combines several annotations and annotators.

We build a typology for simplification that also includes the simplification rules. At the lexical level, one of the most applied rules concerns the deletion of modifiers (adjectives and adverbs) and the replacement of words by simpler synonyms. At the morphological level, we consistently observe a change in the tense of the verbs (usually replacing the simple past (*passé simple*) by the composed past (*passé composé*), but sometimes replacing the simple/composed past by the present). The change of the most frequent words of a morphological family by a word from the same family is observed at a lower frequency. Concerning the syntactic modifications, the suppression of secondary information, such as relative or adverbial subordinate clauses, is noticeable, followed by the sentence reduction (e.g. sentence split at conjunctions and punctuation marks). We also observed some cases of sentence rewriting in order to ensure an SVO (subject-verb-object) structure. These rewriting operations address the cleft and passive sentences.

Additionally, we also observed transformations of negative sentences into positive ones, but at a lower frequency. Furthermore, as expected, we identified several cases of discourse simplification. The most applied rules are from Category 2, followed by those from Category 1, and finally Category 3. Additionally, we also identify 10% of discourse simplifications, such as insertion of pronoun where there is a zero subject, that are not present in the guidelines.

After these observations, we coded the most recurrent rules in the rewriting tool presented in Section 3.2.1. At the syntactic level, we addressed the secondary information removal and sentence reduction. For the former, the system searches for conjunctions linking full sentences or NP, splitting them into two separated sentences. At the coreference level, the NP splits require to repeat some elements to keep reference information. Adverbial clauses are deleted when they are not required by the sentence structure.

The coded rules at the discourse level consisted of five different strategies. At first, some pronouns (e.g. *chaque* and *tout*) when non-coreferent and the subordinate pronouns with their clauses are removed. Then, determiners that are in a coreference chain are changed in order to indicate their position in the chain. Moreover, other determiners are changed following Accessibility theory. Similarly, the third rule explicits coreference relations in possessive determiners. The next rule searches for ambiguous pronouns replacing them by their referents. The last rule solves all anaphoric relations of subject pronouns.

## 4.3. System evaluation

The rules proposed in the last section feed the simplification system. They are coded using the operations presented in Section 3.2.1. and, as indicated in Section 3.2., the text is annotated with syntactic dependency and coreference information before the simplification pipeline starts. This pipeline stacks the syntactic, discursive and lexical simplifications shown in Figure 1.

Aiming to better understand the impact of the simplification on the coreference, we analysed the errors produced by the system. This evaluation is based on the judgement of three judges (two native speakers and one non-native, but advanced, speaker) who evaluated the grammaticality and familiarity of the system output. During this process, they first focused on text cohesion (without lexical simplifications), and then judged the choice of words (lexical simplifications). This approach was adopted to help them to concentrate on the cohesion aspects without distractions from the lexical issues. The total inter-annotator agreement was 56.59% (41.78% for the cohesion and 68.34% for the lexical judgements). Furthermore, we considered only simplification errors spotted by at least two judges.

Concerning the cohesion evaluation, we observed that most errors come from the application of rules from Category 2. It creates referential inconsistencies since it changes the determiner. These errors are caused by coreference annotation tool errors and miss-identification of idioms and collocations. The coreference tool also contributed to errors in rules from Category 1 and 2. These errors may have been caused by both coreference chain divisions (causing determination issues) or merging (mixing different entities). Errors related to Category 3 rules were less frequent, and they are mostly related to coreference chain merging.

The syntactic transformations do not generate noticeable errors. However, during preliminary evaluations, we identified that they mostly contribute to two error types: they caused cascade errors related to ambiguity if the coreference information was not kept in sentence splitting operations. The sentence deletion transformations may over-delete central elements due to parsing errors.

All these transformations generated a total of 180 errors spread into 207 sentences. Taking into account only the lexical simplifications, the systems produced a total of 96 errors (62.35% of accuracy). Considering that these errors have an undesirable impact on simplification evaluation, we changed back all incorrect transformations. Given the grammatical output and the evaluation corpus (described in Section 3), we can move to simplicity evaluation.

We evaluate the simplification using the SARI measure (Xu et al., 2016) (presented in Table 2). However, this measure, is still new, and it lacks in-depth studies. We selected random manual simplifications from the evaluation corpus and set it as a reference. The results of both the system output and the manual simplification are presented in Table 2. This table shows the SARI and BLEU scores as well as other measures related to transformations at the sentence level.

The result of the BLEU score points out a low n-gram variability in the evaluation corpus. Thus, a smaller number of operations may be a useful strategy for this corpus. The SARI score does not indicate a big difference. Moreover,



	System	Manual annotation
SARI	38.124	44.720
BLEU	74.084	91.986
Compression ratio	0.984	1.008
Sentence splits	1.026	1.056
Additions proportion	0.124	0.108
Deletions proportion	0.126	0.104

Table 2: System evaluation.

the same behaviour is observed in sentence-level measures. To better understand the results, we analysed the best and worse SARI’s results by sentence, which lead us to two sources of noise: syntactic and lexical. These issues expose a contradiction in the simplification evaluation. The syntactic noise is related to the removal of secondary information. On one hand, the judges read and understand the texts without significant loss of information, on the other the candidate simplifications tend to keep the secondary information; even if this operation is one of the most performed at the syntactic level. The lexical issues are related to ReSyf. This dictionary contains lexical information graded by complexity, although most of the replacements indicated by this resource are not present in the evaluation corpus.

## 5. Conclusion and further work

We have presented a study of discourse-level transformations to simplify French texts.<sup>9</sup> This study focuses on cohesion issues related to text simplification. From the analysis of a corpus of simplified vs not simplified texts, we have first written guidelines for discourse-level simplifications. We have then designed a system to automatically applied these simplification guidelines. Our system has been evaluated with a corpus containing alternative simplifications proposed by 19 annotators. This corpus also supported the selection of lexical and syntactic simplification rules.

We also presented a proposal for a rule-based coreference-aware simplification system. It was evaluated in terms of text coherence and lexical substitutions by three judges. An automatic evaluation gives a SARI score of 38.13.

During the system evaluation, we identified that most of the miss-simplifications are caused by a lack of language resources. This indicates that the proposed rules seem appropriate, but that extra-linguistic resources are required or should be improved, as the graded lexicon that we used. In a purely rule-based system like ours, tuning further the rules would require a significant development time.

As future work, we intend to improve the system performance. We will explore other coreference properties, such as the negative correlation between the length and the number of chains. We will start with the inclusion of more language resources, but we also intend to explore other approaches than rule-based methods, as well as increase the number of rules through the analysis of other corpora and the use of rules tested in other works, such as Drndarevic and Saggion (2012). A comparison with baseline systems will also complete the evaluation of our system.

<sup>9</sup>The corpora and systems are available at <https://github.com/rswilkens/text-rewrite>.

We also plan to validate the simplification with a larger group of annotators, including dyslexic children. Moreover, we would like to include feedback from the simplification target-group.

## 6. Bibliographical References

- Ariel, M. (1990). *Assessing noun-phrase antecedents*. Routledge.
- Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.
- Barbu, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2013). Open book: a tool for helping asd users’ semantic comprehension. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 11–19.
- Billami, M. B., François, T., and Gala, N. (2018). ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*. Wiley Online Library.
- Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2014a). Syntactic sentence simplification for french. In *3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations*.
- Brouwers, L., Bernhard, D., Ligozat, A., and François, T. (2014b). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014, Gothenburg, Sweden, April 27, 2014*, pages 47–56.
- Canning, Y. M. (2002). *Syntactic simplification of Text*. Ph.D. thesis, University of Sunderland.
- Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., De Marneffe, M.-C., Ramage, D., Yeh, E., and Manning, C. D. (2007). Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170. Association for Computational Linguistics.
- Drndarevic, B. and Saggion, H. (2012). Reducing text complexity through automatic lexical simplification: An empirical study for spanish. *Procesamiento del lenguaje natural*, 49:13–20.
- Fenoglio, I. and Ganascia, J.-G. (2006). Edite, un programme pour l’approche comparative de documents de genèse. *Genesis (Manuscrits-Recherche-Invention)*, 27(1):166–168.
- François, T., Billami, M., Gala, N., and Bernhard, D. (2016). Bleu, contusion, ecchymose: tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL 2016*, volume 2, pages 15–28.
- Givón, T. (1993). Coherence in text, coherence in mind. *Pragmatics & Cognition*, 1(2):171–227.

- Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.
- Jaffe, E., Shain, C., and Schuler, W. (2018). Coreference and focus in reading times. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 1–9.
- Kantor, B. and Globerson, A. (2019). Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677.
- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMO-CRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 92:11–15.
- Le Bouëdec, B. and Martins, D. (1998). La production d'inférences lors de la compréhension de textes chez des adultes : une analyse de la littérature. *L'année psychologique*, 98.
- Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.
- Li, J., Fabre, M., Luh, W.-M., and Hale, J. (2018). The role of syntax during pronoun resolution: Evidence from fmri. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 56–64.
- Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281, 01.
- McMillan, C. T., Clark, R., Gunawardena, D., Ryant, N., and Grossman, M. (2012). fmri evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*, 50(5):674–687.
- Mitkov, R. (2002). *Anaphora Resolution*. Oxford University Press.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Perret, M. (2000). Quelques remarques sur l'anaphore nominale aux xive et xve siècles. *L'Information grammaticale*, 87(1):17–23.
- Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.
- Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2019). Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.
- Quiniou, S. and Daille, B. (2018). Towards a diagnosis of textual difficulties for children with dyslexia. In *11th International Conference on Language Resources and Evaluation (LREC)*.
- Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. (2013). Simplify or help?: text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 15. ACM.
- Saggion, H. (2017). Automatic text simplification: Synthesis lectures on human language technologies, vol. 10 (1). California, Morgan & Claypool Publishers.
- Schnedecker, C. (1997). *Nom propre et chaînes de référence*. Centre d'Etudes Linguistiques des Textes et Discours de l'Université de Metz, Paris.
- Schnedecker, C. (2017). Les chaînes de référence: une configuration d'indices pour distinguer et identifier les genres textuels. *Langue française*, 195(3):53–72.
- Seretan, V. (2012). Acquisition of syntactic simplification rules for french. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4019–4026.
- Siddharthan, A. (2003). Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.
- Siddharthan, A. (2004). *Syntactic simplification and Text Cohesion*. Number 597 in Technical Reports. University of Cambridge, 10.
- Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.
- Sprenger-Charolles, L. and Ziegler, J. C. (2019). Apprendre à lire : contrôle, automatismes et auto-apprentissage. In A. Bentollila & B. Germain, editor, *L'apprentissage de la lecture*. Nathan, September.
- Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.
- Tamburini, F. (2017). Semgex-plus: a tool for automatic dependency-graph rewriting. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 248–254.
- Todirascu, A., François, T., Bernhard, D., Gala, N., and Ligozat, A.-L. (2016). Are cohesive features relevant for text readability evaluation? In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 987–997.
- Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A.-L., and Khobzi, R. (2017). Chaînes de référence et lisibilité des textes : Le projet ALLuSIF. *Langue française*, 195(3):35–52, September.
- Vender, M. (2017). *Disentangling Dyslexia: Phonological and Processing Impairment in Developmental Dyslexia*. Frankfurt: Peter Lang.
- Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Yaneva, V. and Evans, R. (2015). Six good predictors of autistic text comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 697–706.