Acquisition of Syntactic Simplification Rules for French

Seretan, Violeta

Acquisition of Syntactic Simplification Rules for French

Violeta Seretan

FTI/TIM, University of Geneva
40 Bd. du Pont-d'Arve, CH-1211 Genève 4
violeta.seretan@unige.ch

Abstract

Text simplification is the process of reducing the lexical and syntactic complexity of a text while attempting to preserve (most of) its information content. It has recently emerged as an important research area, which holds promise for enhancing the text readability for the benefit of a broader audience as well as for increasing the performance of other applications. Our work focuses on syntactic complexity reduction and deals with the task of corpus-based acquisition of syntactic simplification rules for the French language. We show that the data-driven manual acquisition of simplification rules can be complemented by the semi-automatic detection of syntactic constructions requiring simplification. We provide the first comprehensive set of syntactic simplification rules for French, whose size is comparable to similar resources that exist for English and Brazilian Portuguese. Unlike these manually-built resources, our resource integrates larger lists of lexical cues signaling simplifiable constructions, that are useful for informing practical systems.

Keywords: text simplification, syntax, corpus-driven methods

1. Introduction

In recent years, *text simplification* has emerged as an important area of research in Natural Language Processing (NLP). Text simplification – henceforth, TS – refers to "any process that reduces the syntactic or lexical complexity of a text while attempting to preserve its meaning and information content" (Siddharthan, 2003). TS is closely related to other NLP applications, such as paraphrasing, sentence compression, and text summarisation. The specific motivation behind TS is that the original text is transformed so that it becomes more comprehensible, and, thus, accessible to a broader audience.[1] Increasing the readability of text has great value in education, public health, and safety (Napoles and Dredze, 2010).

Previous work in this area has targeted various categories of readers, from people with aphasia (Carroll et al., 1998; Max, 2006) and hearing-impaired people (Daelemans et al., 2004), to children (De Belder and Moens, 2010), second language learners (Petersen and Ostendorf, 2007), and poor literacy readers (Aluísio et al., 2008; Candido et al., 2009; Aluísio et al., 2010).

In addition to human readers, TS is considered very useful for other NLP applications. Thus, as Chandrasekar et al. (1996) argue, the performance of parsing, machine translation, or information retrieval systems is expected to increase when simplified versions of the input sentences are taken into account. In the same vein, TS has been used by Vickrey and Koller (2008) and by Siddharthan et al. (2004) as a preliminary step in semantic role labelling and text summarisation, respectively, and was shown to lead to better performance.

The bulk of TS work has until now been devoted to English, e.g., Chandrasekar et al. (1996), Carroll et al. (1998), Siddharthan (2003), Max (2006), Petersen and Ostendorf (2007), Napoles and Dredze (2010), De Belder and Moens (2010), Woodsend and Lapata (2011). There are only a few exceptions, dealing with Dutch (Daelemans et al., 2004), Brazilian Portuguese (Aluísio et al., 2008; Candido et al., 2009; Aluísio et al., 2010), and, to some extent, Spanish (Bott and Saggion, 2011). In this paper, we present the first steps we achieved towards the implementation of a TS for French, by acquiring syntactic simplification rules both manually and semi-automatically in a data-driven fashion.

We adopted a rule-based approach, motivated mainly by the lack of suitable parallel corpora in French that would enable the automatic learning of syntactic simplifications, as in previous work for English (Chandrasekar and Srinivas, 1997; Petersen and Ostendorf, 2007) and Dutch (Daelemans et al., 2004). Another motivation originates in the observation that parallel corpora produced with the goal of syntactic simplification in mind cannot reflect the variety of simplification phenomena that can be observed in spontaneous text. In accordance with Siddharthan (2003), we consider that a system that learns rules from a corpus simplified by hand is unlikely to outperform a system that uses hand-crafted rules.

For these reasons stated above, we decided to capitalise on non-parallel resources of naturally occurring text, both more readily available and more informative. We acquire syntactic simplification rules in a data-driven fashion, by applying two complementary methods. The first method consists of a manual corpus analysis aimed at identifying sentences requiring simplification and defining simplification rules. The second method is semi-automatic and consists of the automatic identification of sentences requiring simplification, followed by manual analysis. Its role is to inform and, thus, to support the design of simplification rules.

This strategy is expected to lead to higher rule coverage, as the hand-crafted rules are complemented by automatically acquired information on rule instantiation on a corpus. Thus, instead of providing just a small set of rules involving

[1] In contrast to sentence compression and text summarization, in text simplification there is no information loss, and the length of the text might slightly increase rather than decrease.

few lexical cues[2] as in previous work, we provide a much larger set of rules together with corpus-attested examples which can be used by French text simplification systems.

The article is structured as follows. In Section 2., we discuss the related work on syntactic text simplification, showing how the present work differs from it. In Section 3., we present our methods for data-driven acquisition of syntactic simplification rules for French. We describe both the manual and the semi-automatic method, and provide preliminary experimental results. In Section 4., we make concluding remarks and point out directions for future research.

2. Related Work

Text simplification involves changes at both the lexical level, aimed at replacing complex, uncommon words by simpler variants (e.g., *corroborate* → *validate*), and the syntactic level, aimed at replacing complex constructions with simpler ones. The first type of transformation is taken into account by *paraphrasing methods*, reviewed, for instance, by Androutsopoulos and Malakasiotis (2010). Their application to TS consists, most usually, of using lexical thesauri and lists of frequent words in a language to substitute an uncommon word by a frequent synonym, as in the seminal work of Carroll et al. (1998).

Our work is focussed on simplification at the syntactic level, which are relatively less explored and technically more challenging than lexical simplification methods. As most of the TS work, our method operates at the sentence level. In contrast, methods operating at the discourse level are normally considered as belonging to the field of text summarization. A review of text summarization methods can be found, for instance, in Das and Martins (2007).

Our work can further be distinguished from sentence compression methods, which deal with transformations that are often clause-internal, such as deletion of certain constituents, reordering of constituents, or flattening of syntactic trees as a result of lexical substitution (e.g., *had a noticeable effect on* → *noticeably affected*).[3] For an overview of work on sentence compression, see, for instance, Woodsend et al. (2010).

We are specifically interested in syntactic transformations that change the macro-structure of a complex sentence, i.e., splitting clauses, extracting clauses and dis-embedding non-finite clauses, while leaving (most of) the internal structure of the clauses intact. Unlike sentence compression methods, our syntactic simplification method yields text that is expanded, rather than compressed. Although the text produced is longer, this expansion is arguably helpful for the reader as the syntactic complexity is drastically reduced.

Also, unlike most previous work, our work does not target language-impaired readers – in particular, we do not perform passive to active voice transformations, as does the work motivated by neurolinguistic experiments (Carroll et al., 1998; Max, 2006). Instead, our work is motivated by aspects inherent to the French language and, to a certain extent, to the specificity of the domain considered, namely, newspaper articles: the syntactic structure of French sentences is generally very complex,[4] particularly in the news article domain. It goes without saying that our work is still helpful for the language-impaired reader, and it can be used as well to inform sentence compression.

The transformations we consider are, in principle, most similar in goal to those used by Dras (1999) to perform reluctant paraphrasing (altering a text so that it satisfies some predefined constraints, while remaining as close as possible to the original version). Dras (1999) uses Tree-Adjoining Grammars as a formalism for describing manually-defined transformation rules. Instead, for practical reasons, we adopt a shallow approach based on pattern matching over POS-tagged text, complemented with chunking. This has been proven feasible and reliable for English (Siddharthan, 2002; Siddharthan, 2006), and has also been applied to Brazilian Portuguese (Aluísio et al., 2008; Candido et al., 2009; Aluísio et al., 2010) and Dutch (Daelemans et al., 2004). Like Daelemans et al. (2004), we also make use of information on basic syntactic relations, such as subject-verb and verb-object.

The main difference compared with previous work lies in the manner in which syntactic simplification rules are acquired. Most previous work use only a small set of hand-crafted rules (Carroll et al., 1998; Siddharthan, 2006; Max, 2006; Daelemans et al., 2004). Larger sets of rules have been described in Dras (1999) for English and Specia et al. (2008) for Brazilian Portuguese.[5] Our work covers all these rules with the exception of voice transformation rules (e.g., *The detective was murdered by the butler* → *The butler murdered the detective*) and lexical paraphrasing rules (e.g., *had a noticeable effect on* → *noticeably affected*), that are clause-internal.[6]

It is not really possible to compare our work against work on automatic rule learning (Chandrasekar and Srinivas, 1997; Petersen and Ostendorf, 2007; Daelemans et al., 2004), as the acquired rules are not explicitly listed. The coverage of these rules is heavily dependent on the characteristics of the parallel data used. Chandrasekar and Srinivas (1997) used a manually-built corpus, which arguably leads to limited rule coverage (see comments in Section 1.). Daelemans et al. (2004) used a subtitle corpus, more suitable for sentence compression than for macro-structural simplification. Finally, Petersen and Ostendorf (2007) used a corpus of simplifications in which original sentences can be dropped, whereas our aim is to preserve the information content as much as possible.

The originality of our method consists in using a strategy in which an initial set of hand-crafted rules manually acquired from corpora is augmented and refined, by semi-automatic means, through the detection of constructions that can be simplified. This approach has the advantage of overcoming the data bottleneck of learning-based approaches (as

[2] For instance, splitting clauses joined by *and*, *if - then, which, although, since, when*, etc. (Siddharthan, 2003; Max, 2006).

[3] Example from Dras (1999).

[4] French has "longer" syntactic structures (Jacquemin et al., 1997).

[5] The practical work on Brazilian Portuguese (Aluísio et al., 2008; Candido et al., 2009; Aluísio et al., 2010) is based on the description provided by Specia et al. (2008).

[6] Both examples are taken from Dras (1999).

it only relies on the original complex text), while allowing for more flexibility and coverage with respect to purely manually-based approaches.

3. Rule Acquisition

In this section, we describe our methods for data-driven acquisition of syntactic simplification rules for French.

3.1. Manual Acquisition

The domain of newspaper articles has been selected as the working domain of our simplification system. This domain is particularly challenging for human and automatic text comprehension, given the particularly high structural complexity of sentences from French news articles. Moreover, this is a strategic domain, as the automatic simplification of news articles may impact a very large public.

Consider, as an illustration of a typical sentence from French news articles, the sentence in Example (1a) below.[7]

(1) a. Franco Frattini a par ailleurs indiqué que la vente de pétrole par les rebelles, *pour leur permettre de financer leur lutte contre les forces de Kadhafi,* figurerait à l'ordre du jour de la prochaine réunion du groupe de contact sur la Libye, *qui se tiendra «la première semaine de mai» à Rome*.
"Franco Frattini also said that the sale of oil by the rebels, *in order to allow them to finance their fight against the forces of Gaddafi*, was on the agenda of the next meeting of the Contact Group on Libya, *which will be held 'the first week of May' in Rome.*"

b. Franco Frattini a par ailleurs indiqué que la vente de pétrole par les rebelles figurerait à l'ordre du jour de la prochaine réunion du groupe de contact sur la Libye. Le but est de leur permettre de financer leur lutte contre les forces de Kadhafi. La réunion se tiendra «la première semaine de mai» à Rome.
"Franco Frattini also said that the sale of oil by the rebels was on the agenda of the next meeting of the Contact Group on Libya. The goal is to enable them to finance their fight against the forces of Gaddafi. The meeting will be held 'the first week of May' in Rome."

The two subordinate clauses shown in (1a) in italics are a final clause and a relative clause, which add a considerable comprehension burden for the conveyed message. These clauses can be detached from the initial sentence in order to increase its readability. This transformation yields the three-sentence text in (1b). Therefore, starting from a sentence like (1a), two simplification rules can be inferred: one for extracting subordinate final clauses (clauses expressing purpose), and another one for extracting relative clauses. Each time a new rule is designed, it is informally described in linguistic terms, and the transformation that takes place is illustrated with a sentence that occurs in the corpus.[8]

The process of manual rule acquisition consists of the analysis of a number of newspaper articles, then, for each article, the annotation of clause boundaries in the complex sentences for which a simplification can be proposed, and, finally, the definition of simplification rules. The level of generality of rules is not fixed beforehand; the annotator goes from specific to general, a classification of rules emerging as more examples of rule instantiations are found in the corpus and a certain degree of similarity is found with the rules already proposed. For example, the relative clause extraction rule is actually a group of similar rules, which includes the extraction of subject relative clauses, the extraction of object relative clauses, and the extraction of indirect relative clauses.

Many rules are quite specific, as they refer to constructions involving particular words. The acquired rules are therefore not necessarily expressed by means of syntactic abstraction, but they remain relatively lexicalised. An example is the '*Split* si + *adjective* + que' rule, illustrated in Example (2) below:

(2) a. Nous sommes dans des véhicules sécurisés et *si* confortables *que* rouler à 130 km/h ne donne pas l'impression de vitesse.
"We are in secure vehicles, *so* comfortable *that* driving at 130 km/h does not give the impression of speed."

b. Nous sommes dans des véhicules sécurisés et confortables. Rouler à 130 km/h ne donne pas l'impression de vitesse.
"We are in secure and comfortable vehicles. Driving at 130 km/h does not give the impression of speed."

A preliminary manual acquisition process has been performed on a small corpus of 30 articles from the on-line version of the French newspaper *Le Figaro*,[9] totalling 607 sentences (titles excluded). It has led to the definition of about 40 simplification rules, grouped into the main categories displayed in Table 1.

Note that, as far as reporting clauses are concerned (category C), the direct speech, surrounded by quotes, is left intact. Also, regarding relative clauses (category E), note that restrictive clauses are, obviously, not extracted: *une étude de 2007 qui dénonce les mêmes abus* "a 2007 study that denounces the same abuses". Short relatives are also left unextracted for the ease of the reading flow: *a affirmé mardi à Rome le ministre italien des Affaires étrangères, Franco Frattini, citant le chef du CNT, Moustapha Abdeljalil, qu'il venait de recevoir* "said Tuesday in Rome, Italian Foreign Minister Franco Frattini, quoting the head of the CNT, Moustapha Abdeljalil, whom he had just met".

The third column of Table 1 shows the distribution of the total 175 rule instantiations across these categories. In the corpus considered in our analysis, the most represented category is E - subordination (42.9%), divided among relatives (15.4%), gerundial clauses (10.9%) and other subordination types (16.6%). For most of the individual rules within each category, there are only very few instantiations in the corpus considered. To augment the coverage of our set of simplification rules, we therefore designed a semi-automatic method for assisting the rule acquisition process.

3.2. Semi-automatic Acquisition

Defining syntactic simplification rules by manual inspection of a corpus is a task that is time-consuming and in-

[7] This sentence actually occurs in an article from our corpus (details provided later in this section).

[8] Part of the rules are formally described and implemented.

[9] http://www.lefigaro.fr/

ID	Category	Ratio	Examples (Appendix A)
A	Deletion of hedging phrases (i.e., phrases with little semantic context) such as *par ailleurs* "also", *dans ce contexte* "in this context", de son côté "as for him", *ce qu'il convient d'appeler* "so-called" (lit., *what one may conviently call*)	6.3%	1
B	Extraction of appositive phrases and clauses, including clauses surrounded by parentheses and dashes	9.7%	2, 3
C	Extraction of reporting clauses (i.e., clauses that introduce direct speech), such as *a précisé, a affirmé, avance, estime* "stated/states", *selon* "according to"	10.9%	4, 5
D	Splitting coordinated clauses, including clauses separated by commas, colons and semi-colons	11.4%	6
E	Extraction of subordinated clauses, including relative clauses and gerundial clauses	42.9%	7, 8, 9
F	Extraction of long adjuncts and long post-nominal modifiers, including participial modifiers	12%	10, 11, 12
G	Extraction of small clauses	5.7%	13
H	De-clefting, i.e., transformation of cleft sentences into regular non-cleft sentences	1.1%	14

Table 1: Main categories of simplification rules acquired manually and their corpus distribution.

feasible at a large scale. It is not realistic to assume that the inspection of a limited amount of language data can spot all (or most) of the linguistic phenomena pertaining to simplification; a language in general, and French in particular, exhibits a high richness and variability of linguistic expression from this point of view. Moreover, the rules that are manually acquired from data are generally instantiated only once, or just a few times, in the inspected corpus (see Section 3.1.). Although the annotator may come up with generalisations (such as on the extraction of subordinated clauses), the acquired rules would still lack a comprehensive list of lexical triggers, required by practical systems (for instance, subordinating conjunctions like *avant de* "before", *alors que* "while", *tandis que* "whereas", *puisque* "because", *parce que* "because" , *même si* "even if", etc).

To assist the manual acquisition work, we devised a method that leverages linguistic tools to provide annotators with statistically salient linguistic expressions that are potentially interesting from a rule inference point of view. That is, our method detects frequent sentence parts that may contain lexical triggers, or, more generally, that may constitute the beginning of a rule argument (i.e., a phrase or clause that has to be promoted to sentence status). Such sentence parts may be, for instance, conjunctions or adverbs preceded by a comma and introducing coordinated or subordinated clauses; definite noun phrases constituting appositions; pronouns – possibly preceded by prepositions – introducing relatives; past participles introducing long post-nominal modifiers, etc. (see also the rule classification in Table 1).

This method is applied to corpora from the same domain as the domain we considered for the simplification task, namely, newspaper articles. Given such a corpus, the method consists of linguistically pre-processing of the original sentences by POS-tagging them, collecting punctuation-free sequences of words following a comma, and gathering lexical and POS statistics on the resulting collection.

We consider various lengths for the collected sequences, from 1 to 5. The most frequent sequences are computed by taking into account lexical information (i.e., the word proper), POS information (either the lexical category – conjunction, verb, etc. – or more detailed morphological information – e.g., subordinating conjunction, coordinating conjunction, infinitive verb), both separately and in combination. The combination of lexical and POS information would help to identify, for instance, sequences like *sans + infinitive verb* (the French equivalent of the English construction "without + gerundive verb").

A frequency threshold of 5 is applied to the sequences and these are displayed in decreasing order of frequency, together with the full sentence in which they were detected. In addition to sentence parts preceded by commas, we also collect sentence-initial sequences of length 2, which may be informative of phenomena like small clauses or inversion, relevant to syntactic simplification.

Preliminary experiments have been performed on news articles from the *Le Monde* corpus distributed by the Linguistic Data Consortium. These total about 400'000 words, for a repoted number of 505'907 tokens, including punctuation. There are 22'182 sentences in the corpus, with an average length of 22.8 tokens.

The POS tagger used is a part of a lexicalist symbolic parser based on generative grammar concepts (Wehrli, 2007). The precision of the French version of the tagger is very high, 99.3% (Ruch and Gaudinat, 2000). The tagger provides a detailed morphosyntactic analysis of the words in a sentence, showing only the reading that is compatible with the syntactic analysis built by the underlying parser. The output also includes information on the grammatical function of words, indicating the predicate-argument structure of the sentence.

Table 2 displays some of the most frequent sequences – of increasing length – collected using the method we described above. As explained, such sequences are deemed to be interesting from the point of view of human annotators that perform the task of rule acquisition for syntactic simplification. They help annotators to explore the richness of linguistic expressions characterising the complex journalistic sentences. Although – as for any linguistic phenomenon – there is a high dispersion of the data which is potentially relevant for simplification, some patterns emerge and the statistics provide help in discovering them. The information gathered may be queried and visualised so that the annotators can discover specific aspects of the data, helping them to validate their rule hypotheses and to obtain further

Len.	POS Sequence	Freq.	Sample Lexical Sequence
1	PRE	1644	*de* "of", *dans* "in", *à* "to", *avec* "with", *comme* "as", *en* "in", *pour* "for", *dont* "of whom/which", *selon* "according to", *sur* "on", *par* "by", *après* "after", *sous* "under", *malgré* "despite"
	DET-DEF-SIN-MAS	1393	*le* "the"
	DET-DEF-SIN-FEM	1120	*la* "the"
	COJ-SUB	1077	*pour* "to", *comme* "as", *si* "if", *car* "since", *alors que* "while", *que* "that", *parce que* "because", *même si* "even if", *puisque* "since", *tout en* "while", *quand* "when", *sans* "without"
	ADV	1039	*mais* "but", *où* "where", *notamment* "especially", *plus* "more", *alors* "then", *surtout* "especially", *pas* "non", *c'est-à-dire* "that is to say", *encore* "yet", *peut-être* "maybe"
	COJ-COO	829	*et* "and", *mais* "but", *ou* "or", *soit* "or", *ainsi que* "and", *puis* "then", *ni* "neither"
	DET-DEF-PLU-MAS	487	*les* "they"
	NOM-COM-SIN-MAS	380	*directeur* "director", *chef* "head", *président* "president", *professeur* "professor", *auteur* "author", *membre* "member", *conseiller* "advisor", *patron* "boss"
	VER-IND-PRE-3-SIN	380	*dit* "says", *doit* "must", *peut* "can", *écrit* "writes", *explique* "explains", *raconte* "tells", *conclut* "concludes"
	PRO-REL-INN-ING	355	*qui* "which$_{Nominative}$", *que* "which$_{Accusative}$"
2	DET-DEF-SIN-MAS, NOM-COM-SIN-MAS	1119	*le président* "the president", *le chef* "the head", *le temps* "the time"
	COJ-SUB, VER-INF	217	*pour éviter* "to avoid", *à commencer* "to begin", *sans accepter* "without accepting"
3	PRE, DET-DEF-SIN-FEM, NOM-COM-SIN-FEM	248	*à l'image* "just as", *à la veille* "the eve", *selon l'expression* "according to the expression"
	PRE, DET-DEF-SIN-MAS, NOM-COM-SIN-MAS	155	*dans le nord* "in the north", *dans le domaine* "in the field", *de l'avis* "on the opinion" *sous le nom* "under the name"
4	COJ-SUB, PRO-DEM-SIN-MAS, ADV, VER-IND-PRE-3-SIN	10	*si ce n'est* "except"
5	PRO-DEM-INN-MAS, PRO-REL-INN-MAS, ADV, VER-IND-PRE-3-SIN, ADV	6	*ce qui n'est pas* "which is not"

Table 2: Sample output of the method, showing sequences likely to trigger simplification rules. Acronyms used: 3rd person, ADVerb, COnJunction, COMmon, COOrdinating, DEFinite, DEMonstrative, DETerminer, FEMinine, INDefinite, INdefinite Gender, INFinitive, INdefinite Number, MASculine, NOMinal, PREposition, PROnoun, PLUral, RELative, SINgular, SUBordinating, VERb.

information related to a given rule.

For example, annotators may look for evidence on subordinate clauses beginning a sentence. By appropriately visualising the sequence data collected they can easily discover phrases which signal the phenomenon they are interested in (such as those shown in Example (3a)).

(3) a. alors que "while", après "after", avant de "before", avec la "with the", au cours de "during", bien que "although", comme "as", depuis "ever since", grâce à "thanks to", lors de "on the occasion of", malgré "despite", parce que "because", plus que "more than", plutôt que "rather than", quand "when", si "if", tandis que "whereas", puisque "since", même si "even if";

b. affirmer "state", assurer "assure", commenter "comment", declarer "state", estimer "say", expliquer "explain", exposer "exposed", indiquer "indicate", lancer "launch", relever "emphasise", souligner "stress".

Likewise, annotators may discover a wide range of verbs beginning reportive clauses, which trigger the rules in the category C from the classification provided in Table 1. Such verbs are shown in Example (3b).

The longer expressions detected from the sequence data may signal turns of phrases that are somewhat specific to the journalistic domain: e.g., *à en croire* "if we are to believe", *à l'image de* "just as", *le temps pour la justice de* "time for the justice to", *dans la ligne de* "in line with", *ne serait-ce que parce que* "even if only because", *il en va autrement pour* "it is different for". Arguably, these are also useful as syntactic simplification cues and may be integrated as information useful for simplification systems.

4. Conclusion

In this paper, we have focussed on the syntactic simplification of text, as a means to achieve increased readability for French newspaper articles so that they become accessible to a broader audience. In contrast to most previous work, we have focussed mainly on macro-structural operations, which transform a complex sentence into a sequence of simpler sentences by promoting clauses to sentence status. The coverage of the acquired set of rules is comparable to that of other sets proposed by Dras (1999) and Specia et al. (2008), for English and Brazilian Portuguese, respectively. To our knowledge, this is the most comprehensive syntactic simplification rule set proposed for the French language.

Another distinguishing feature of our work is that the manual data-driven acquisition of rules is complemented by a semi-automatic method which detects typical lexical/POS sequences that signal constructions requiring simplification. Preliminary results on the *Le Monde* corpus show the potential of the method to provide information which is useful for designing simplification rules. Our approach has the advantage of overcoming the data bottleneck of learning-based approaches, since it does not require parallel resources; at the same time, it allows for increased flexibility and coverage with respect to purely manually-based approaches.

The rules designed are being implemented in a text simplification system currently under development. Future work will focus on evaluating the semi-automatic acquisition method, and on studying to what extent this method can be applied to text from a different domain. In particular, we will explore the user-generated content domain, and use syntactic simplification as a means of pre-editing for statistical machine translation in the ACCEPT European project.

Acknowledgments

This work has been done during the author's stay at the University of Edinburgh. It has received the support of the Swiss National Science Foundation (grant no. PA00P1_131512). Thanks to Mirella Lapata for inspiring discussions, and to Manny Rayner for a thorough reading.

5. References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proc. of the 8th ACM Symposium on Document Engineering*, pages 240–248, Sao Paulo, Brazil.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proc. of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California, June.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, May.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *ACL Workshop on Monolingual Text-to-Text Generation*, Portland, USA.

Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluísio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proc. of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proc. of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, Madison, USA.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10:183–190.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceeding of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 1041–1044, Copenhagen, Denmark.

Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proc. of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048, Lisbon, Portugal.

Dipanjan Das and André F. T. Martins. 2007. A Survey on Automatic Text Summarization. `http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf`.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proc. of SIGIR 2010 Workshop Towards Accessible Search Systems*, pages 19–26, Geneva, Switzerland.

Mark Dras. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proc. of the 35th annual meeting of ACL*, pages 24–31, Morristown, NJ, USA.

Aurélien Max. 2006. Writing for language-impaired readers. In *Proc. of Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006*, pages 567–570, Mexico City, Mexico.

Courtney Napoles and Mark Dredze. 2010. Learning Simple Wikipedia: A cogitation in ascertaining abecedarian language. In *Proc. of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50, Los Angeles, CA, USA, June.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology for Education*, pages 69–72, Pennsylvania, USA.

Patrick Ruch and Arnaud Gaudinat. 2000. Comparing corpora and lexical ambiguity. In *Proc. of the workshop on Comparing corpora - Volume 9*, WCC '00, pages 14–19, Hong Kong.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proc. of the 20th International Conference on Computational Linguistics (COLING '04)*, Geneva, Switzerland.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *In Proc. of the Language Engineering Conference (LEC'02*, pages 64–71, Hyderabad, India.

Advaith Siddharthan. 2003. *Syntactic simplification and text cohesion*. Ph.D. thesis, University of Cambridge.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4:77–109.

Lúcia Specia, Sandra Maria Aluísio, and Thiago A. Salgueiro Pardo. 2008. Manual de simplificação sintática para o portuguěs. Technical report, Universidade de São Paulo.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proc. of ACL-08: HLT*, pages 344–352, Columbus, Ohio, June.

Eric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July.

Kristian Woodsend, Yansong Feng, and Mirella Lapata. 2010. Title generation with Quasi-Synchronous Grammar. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pages 513–523.

Appendix A Syntactic Simplification Rules for French – Representative Examples

No.	Description	Example	Translation
1	Delete hedging phrases	"*Dans ce contexte*, le chef de la diplomatie italienne a promis ... → Le chef de la diplomatie italienne a promis ...	"*In this context*, the head of Italian diplomacy promised ..." → "The head of Italian diplomacy promised ..."
2	Extract appositive phrase	L'organisation a mené mardi matin des raids sur Tripoli, Aziziyeh et Syrte, *ville natale du dirigeant libyen Mouammar Kadhafi* → L'organisation a mené mardi matin des raids sur Tripoli, Aziziyeh et Syrte. Il s'agit de la ville natale du dirigeant libyen Mouammar Kadhafi	"The organization conducted Tuesday morning raids on Tripoli, Aziziyeh and Sirte, *hometown of Libyan leader Muammar Gaddafi*" → "The organization conducted Tuesday morning raids on Tripoli, Aziziyeh and Sirte. This is the hometown of Libyan leader Muammar Gaddafi"
3	Extract appositive clause enclosed in dashes	Protégée de Prince, qui n'hésite pas à voyager pour assister à ses concerts *– on l'a vu admirer la jeune femme à La Cigale en mai 2009 –* , elle a aussi collaboré avec les Rolling Stones. → Elle est protégée de Prince, qui n'hésite pas à voyager pour assister à ses concerts. On l'a vu admirer la jeune femme à La Cigale en mai 2009. Elle a aussi collaboré avec les Rolling Stones.	"Protected by Prince, who does not hesitate to travel to attend her concerts *– we saw him admiring the young woman at La Cigale in May 2009 –* she has also worked with the Rolling Stones." → "She is protected by Prince, who does not hesitate to travel to attend her concerts. We saw him admiring the young woman at La Cigale in May 2009. She has also worked with the Rolling Stones."
4	Extract reporting phrase	*Selon la présidente du groupe nucléaire français, Anne Lauvergeon,* Tepco espère entamer ces opérations avant la fin du mois de mai. → Tepco espère entamer ces opérations avant la fin du mois de mai. C'est ce qu'affirme la présidente du groupe nucléaire français, Anne Lauvergeon.	"*According to the President of the French nuclear group, Anne Lauvergeon,* Tepco hopes to begin these operations before the end of May." → "Tepco hopes to begin these operations before the end of May. This is what says the president of the French nuclear group, Anne Lauvergeon."
5	Extract reporting clause	En deux mois, le conflit en Libye a déjà fait quelque 10.000 morts et 55.000 blessés, *a affirmé mardi à Rome le ministre italien des Affaires étrangères, Franco Frattini,* (...) → En deux mois, le conflit en Libye a déjà fait quelque 10.000 morts et 55.000 blessés. C'est ce qu'a affirmé mardi à Rome le ministre italien des Affaires étrangères, Franco Frattini, (...)	"In two months, the conflict in Libya has claimed some 10,000 dead and 55,000 wounded, *said Tuesday in Rome, Italian Foreign Minister Franco Frattini* (...)" → "In two months, the conflict in Libya has claimed some 10,000 dead and 55,000 wounded. This is what said Tuesday in Rome, Italian Foreign Minister Franco Frattini (...)"
6	Split coordinated clauses	Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques, *mais aussi développer la sécurisation des réseaux routiers* (...) → Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques. Mais il faut aussi développer la sécurisation des réseaux routiers (...)	"We must encourage the education of children and adults for an awareness of the risks, *but also develop the road network security* (...)" → "We must encourage the education of children and adults for an awareness of the risks. But we must also develop the road network security (...)"
7	Extract subordinated clauses	*Après s'être procurés les coordonnées bancaires de l'ANCV (...)*, ces escrocs auraient adressé un faux ordre de virement sur un compte chinois. → Ces escrocs se sont procurés les coordonnées bancaires de l'ANCV (...). Ensuite, ils auraient adressé un faux ordre de virement sur un compte chinois.	''After having obtained the bank details of ANCV (...), these crooks would have issued a forged transfer order to an account in China." → "These crooks have obtained the bank details of the ANCV. Then they would have issued a forged payment order to an account in China."

No.	Description	Example	Translation
8	Extract object relative clauses	Beauté formelle et science-fiction cohabitent dans cette fable *que l'auteur définit comme «un film catastrophe psychologique»*. → Beauté formelle et science-fiction cohabitent dans cette fable. L'auteur définit cette-dernière comme «un film catastrophe psychologique».	"Formal beauty and science fiction coexist in this fable *that the author defines as 'psychological disaster movie'*." → "Formal beauty and science fiction together in this fable. The author defines this as 'a psychological disaster movie'."
9	Extract gerundial clauses	Areva va aider les Japonais à décontaminer la centrale *en installant une station d'épuration qui permettra de réduire le niveau de radioactivité des eaux traitées*. → Areva va aider les Japonais à décontaminer la centrale. Elle va installer une station d'épuration qui permettra de réduire le niveau de radioactivité des eaux traitées.	"Areva will help the Japanese to decontaminate the plant *by installing a treatment plant that will reduce the level of radioactivity treated water*." → "Areva will help the Japanese to decontaminate the plant. It will install a treatment plant that will reduce the level of radioactivity of treated water."
10	Extract long adjuncts	Le vice-président américain Joe Biden estime, *dans une interview au Financial Times*, que l'Otan peut se passer des Etats-Unis en Libye (...) → Le vice-président américain Joe Biden estime que l'Otan peut se passer des Etats-Unis en Libye (...). Cette affirmation a été faite dans une interview au Financial Times.	"The Vice-President Joe Biden says, *in an interview with the Financial Times*, that NATO can do without the United States in Libya (...)" → "The Vice-President Joe Biden says that NATO can do without the United States in Libya (...). This statement was made in an interview with the Financial Times. "
11	Extract long post-nominal modifiers	(...) l'interrogeait sur la demande du président de la commission des Affaires étrangères de l'Assemblée nationale, Axel Poniatowski (UMP), *d'envoyer en Libye 200 à 300 membres de forces spéciales de pays de l'Otan pour aider la rébellion* → (...) l'interrogeait sur la demande du président de la commission des Affaires étrangères de l'Assemblée nationale, Axel Poniatowski (UMP). Cette demande consistait à envoyer en Libye 200 à 300 membres de forces spéciales de pays de l'Otan pour aider la rébellion	"(...) asked him about the request of the president of the Foreign Affairs Committee of the National Assembly, Axel Poniatowski (UMP), to send in Libya 200 to 300 members of special NATO forces to aid the rebellion" → (...) asked him about the request of the president of the Foreign Affairs Committee of the National Assembly, Axel Poniatowski (UMP). This request was to send in Libya 200 to 300 members of special NATO forces to aid the rebellion"
12	Extract long participial modifiers	(...) du plan de prévention des risques psychosociaux, *lancé en octobre 2009 par Xavier Darcos*. → (...) du plan de prévention des risques psychosociaux. Ce plan a été lancé en octobre 2009 par Xavier Darcos.	"(...) of the psychosocial risks prevention plan, launched in October 2009 by Xavier Darcos." → "(...) of the psychosocial risks prevention plan. This plan was launched in October 2009 by Xavier Darcos."'
13	Promote small clauses	*Intrigué*, le banquier a téléphoné pour vérification à l'élysée → Le banquier a été intrigué. Il a téléphoné pour vérification à l'élysée	"*Intrigued*, the banker called for verification at the Elysee" → "The banker was intrigued. He called for verification at the Elysee"
14	De-cleft	*C'est alors que le Portugal traverse une crise économique et politique majeure* que sa musique vit une période de grande créativité. → Le Portugal traverse une crise économique et politique majeure. C'est à ce moment que sa musique vit une période de grande créativité.	"*It is when Portugal is going through a major political and economic crisis that* his music is experiencing a period of great creativity." → "Portugal is experiencing a major economic and political crisis. That's when his music is experiencing a period of great creativity."