10.12871/CLICIT2014118

Defining an annotation scheme with a view to automatic text simplification

Dominique Brunato Felice Dell'Orletta Giulia Venturi Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR)
ItaliaNLP Lab – *www.italianlp.it*
Via G. Moruzzi,1 – Pisa (Italy)
{name.surname}@ilc.cnr.it

Abstract

English. This paper presents the preliminary steps of ongoing research in the field of automatic text simplification. In line with current approaches, we propose here a new annotation scheme specifically conceived to identify the typologies of changes an original sentence undergoes when it is manually simplified. Such a scheme has been tested on a parallel corpus available for Italian, which we have first aligned at sentence level and then annotated with simplification rules.

Italiano. *In questo contributo presentiamo i primi passi delle ricerche attuali sulla semplificazione automatica del testo. In linea con gli approcci più recenti, proponiamo qui un nuovo schema di annotazione teso specificamente a identificare le tipologie di cambiamenti che una frase originale subisce quando viene semplificata manualmente. Questo schema è stato testato su un corpus parallelo disponibile per l'italiano, che abbiamo precedentemente allineato a livello di frase e successivamente annotato con le regole di semplificazione.*

1 Introduction

Automatic Text Simplification (ATS) as a field of research in NLP is receiving growing attention over the last few years due to the implications it has for both machine- and human-oriented tasks. For what concerns the former, ATS has been employed as a pre-processing step, which provides an input that is easier to be analyzed by NLP modules, so that to improve the efficiency of, e.g., parsing, machine translation and information extraction. For what concerns the latter, ATS can also play a crucial role in educational and assistive technologies; e.g., it is used for the creation of texts adapted to the needs of particular readers, like children (De Belder and Moens, 2010), L2 learners (Petersen and Ostendorf, 2007), people with low literacy skills (Aluìsio et al., 2008), cognitive disabilities (Bott and Saggion, 2014) or language impairments, such as aphasia (Carroll et al., 1998) or deafness (Inui et al., 2003).

From the methodological point of view, while the first attempts were mainly developed on a set of predefined rules based on linguistic intuitions (Chandrasekar et al., 1996; Siddharthan, 2002), current ones are much more prone to adopt data-driven approaches. Within the latter paradigm, the availability of monolingual parallel corpora (i.e. corpora of authentic texts and their manually simplified versions) turned out to be a necessary prerequisite, as they allow for investigating the actual editing operations human experts perform on a text in the attempt to make it more comprehensible for their target readership. This is the case of Brouwers et al. (2014) for French; Bott and Saggion (2014) for Spanish; Klerke and Søgaard (2012) for Danish and Caseli et al. (2009) for Brazilian Portuguese. To our knowledge, only a parallel corpus exists for Italian which was developed within the EU project Terence, aimed at the creation of suitable reading materials for poor comprehenders (both hearing and deaf, aged 7-11)[1]. An excerpt of this corpus was used for testing purposes by Barlacchi and Tonelli (2013), who devised the first rule-based system for ATS in Italian focusing on a limited set of linguistic structures.

The approach proposed in this paper is inspired to the recent work of Bott and Saggion (2014) for Spanish and differs from the work of Barlacchi and Tonelli (2013) since it aims at learning from a parallel corpus the variety of text adaptations that characterize manual simplification. In particular, we focus on the design and development of a new annotation scheme for the Italian language intended to cover a wide set of linguistic phenomena implied in text simplification.

[1] More details can be found in the project website: http://www.terenceproject.eu/

2 Corpus alignment

The Terence corpus is a collection of 32 authentic texts and their manually simplified counterpart, all covering short novels for children. The simplification was carried out in a cumulative fashion with the aim of improving the comprehension of the original text at three different levels: global coherence, local cohesion and lexicon/syntax.

Given its highly structured approach and the clearly focused target, we believe the Terence corpus represents a very useful resource to investigate the manual simplification process with a view to its computational treatment. In particular, we proceeded as follows. First, we selected the outcomes of the last two levels of simplification (i.e. local cohesion and lexicon/syntax) which were considered respectively as the original and the simplified version of the corpus. This choice was motivated by the need of tackling only those textual simplification aspects with a counterpart at the linguistic structure level. We then hand-aligned the resulting 1036 original sentences to the 1060 simplified ones. The alignment results (table 1) provide some insights into the typology of human editing operations. As we can see, in 90% of the cases a 1:1 alignment is reported; 39 original sentences (3.75%) have a correspondence 1:2, thus suggesting an occurred split; 2 original sentences have undergone a three-fold split (0.19%), i.e. they correspond to three sentences in the simplified version; 15 pairs of original sentences have been merged into a single one (2.88%). Finally, the percentage of misaligned sentences is 1% (7 sentences were completely deleted after the simplification, whereas 4 novel ones have been introduced in the simplified corpus).

	1:1	**1:2**	**1:3**	**2:1**	**1:0**	**0:1**
N°sentences	958	39	2	30	7	4
%	92.1	3.75	0.19	2.88	0.67	0.38

Table 1: Corpus alignment results

3 Simplification annotation scheme

For the specific concerns of our study, we have defined the following annotation scheme, covering six macro-categories: split, merge, reordering, insert, delete and transformation. For some of them, a more specific subclass has been introduced, while for others (e.g. reordering) we are providing a finer internal distinction and a qualitative analysis focused on some selected constructs. Such a two-leveled structure has been similarly proposed by Bott and Saggion (2014) and we believe it is highly flexible and reusable, i.e. functional to capture similarities and variations across paired corpora from diverse domains and for different categories of readers. In table 2 we report the typology of rules covered by the annotation scheme. For each rule we also provide the frequency distribution within the Terence corpus.

Simplification Annotation Scheme		
Classes	***Sub-classes***	***Freq. %***
Split		1.75
Merge		0.57
Reordering		8.65
Insert	Verb	4.93
	Subject	1.79
	Other	12.03
Delete	Verb	2.04
	Subject	0.49
	Other	19.45
Transformation	Lexical Substitution	40.01
	Anaphoric replacement	0.61
	Noun_to_Verb	1.59
	Verb_to_Noun (nominalization)	0.61
	Verbal Voice	0.53
	Verbal Features	4.93

Table 2: Simplification annotation scheme

Split: it is the most investigated operation in ATS, for both human- and machine-oriented applications. Typically, a split affects coordinate clauses (introduced by coordinate conjunctions, colons or semicolons), subordinate clauses (e.g., non-restrictive relative clauses), appositive and adverbial phrases. Nevertheless, we do not expect that each sentence of this kind undergoes a split, as the human expert may prefer not to detach two clauses, for instance when a subordinate clause provides the necessary background information to understand the matrix clause. In (1) we give an example of split from the corpus[2].

(1) O: *Mamma Gorilla sembrava completamente distrutta per le cure che dava al suo vivace cucciolетto Tito,* ***che stava giocando vicino alle grosse sbarre di acciaio che circondavano il recinto***.

[2] In all the examples of aligned sentences from the corpus, O stands for original and S for simplified.

S: *Mamma Gorilla sembrava proprio distrutta per le cure che dava al suo vivace cuccioletto Tito.* ***Tito stava giocando vicino alle grosse sbarre di acciaio che erano intorno alla loro area***.

Merge: it has to be intended as the reverse of split, i.e. the operation by which two (or more) original sentences are joined into a unique simplified sentence. Such a kind of transformation is less likely to be adopted, as it creates semantically denser sentences, more difficult to process (Kintsh and Keenan, 1973). Yet, to some extent (see the alignment results), this is a choice the expert can make (ex. 2) and it can be interesting to verify whether the sentences susceptible to be merged display any regular pattern of linguistic features that can be automatically captured.

(2) O: *Clara pensò che fosse uno dei cigni. Ma poi si rese conto che stava urlando!*

S: *In un primo momento, Clara pensò che fosse uno dei cigni, ma poi sentì urlare!*

Reordering: this tag marks rearrangements of words between the original sentence and its simplified counterpart (3). Clearly, changing the position of the elements in a sentence is not an isolated event but it depends upon modifications at lexicon or syntax; e.g., replacing an object clitic pronoun (which is preverbal with finite verbs in Italian) with its full lexical antecedent[3] yields the unmarked order SVO, associated with easier comprehension and earlier acquisition (Slobin and Bever, 1982). Conversely, the author of the simplified text may sometimes prefer a non-canonical order, when s/he believes, e.g., that it allows the reader to keep the focus stable over two or more sentences.

(3) O: *Il passante gli spiegò che, per arrivare al bidone,* ***doveva contare ben 5 bidoni a partire dal semaforo***.
S: *Il signore spiegò a Ugolino che* ***doveva contare 5 bidoni a partire dal semaforo****, per arrivare al bidone della carta*.

Insert: the process of simplification may even result in a longer sentence, because of the insertion of words or phrases that provide supportive information to the original sentence. Despite the cognitive literature suggests to reduce the inference load of a text, especially with less skilled or low-knowledge readers (Ozuru et al., 2009), it is difficult to predict what the author of a simple text will actually add to the sentence to make it clearer. It can happen that the sentence is elliptical, i.e. syntactically compressed, and the difficulty depends on the ability to retrieve the missing arguments, which are then made explicit as a result of the simplification. Our annotation scheme has introduced two more specific tags to mark insertions: one for verbs and one for subject. The latter signals the transformation of a covert subject in a lexical noun phrase[4].

(4) O: *Essendo da poco andata in pensione dal suo lavoro, disse che le mancavano i suoi studenti* […]

S: *Essendo da poco andata in pensione dal suo lavoro* ***come insegnante****, disse che le mancavano i suoi studenti* […]

Delete: a text should be made easier by eliminating redundant information. As for the *insert* tag, also deletion is largely unpredictable, although we can imagine that simplified sentences would contain less adjunct phrases (e.g. adverbs or adjectives) than the authentic ones. Such occurrences have been marked with the underspecified *delete* rule (ex. 5); two more restricted tags, *delete_verb* and *delete_subj*, have been introduced to signal, respectively, the deletion of a verb and of an overt subject (made implicit and recoverable through verb agreement morphology).

(5) O: ***Sembrava veramente che*** *il fiume stesse per straripare*.
S: *Il fiume stava per straripare*.

Transformation: under this label we have included six main typologies of transformations that a sentence may be subject to, in order to become more comprehensible for the intended reader. Such modifications can affect the lexical, morpho-syntactic and syntactic levels of sentence representation, also giving rise to overlapping phenomena. Our annotation scheme has intended to cover the following phenomena:

- *Lexical substitution*: that is when a word (or a multi-word expression) is replaced with another (or more than one), which is usually a more common synonym or a less specific term. Given the relevance of lexical changes in text simplification, which is also confirmed by our results, previous works have proposed feasible ways to automatize lexical simplification, e.g. by relying on electronic resources, such as WordNet (De Belder et al., 2010) or word frequency lists (Drndarevic et al., 2012). In our annotation scheme this rule has been conceived to be quite generic, as synonyms or hypernyms replacements do not

[3] This is also a case of anaphora resolution, for which a dedicated tag has been conceived.

[4] The covert/overt realization of the subject is an option available in null-subject languages like Italian.

cover all the strategies an author can adopt to reduce the vocabulary burden of a text. A finer characterization will be part of a qualitative analysis.

(6) O: *Il **passante** gli spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni a partire dal semaforo*.

S: *Il **signore** spiegò a Ugolino che doveva contare 5 bidoni a partire dal semaforo, per arrivare al bidone della carta*.

- *Anaphoric replacement*: the substitution of a referent pronoun with its full lexical antecedent (a definite noun phrase or a proper noun);

(7) O: *Il passante **gli** spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni* […].

S: *Il signore spiegò **a Ugolino** che doveva contare 5 bidoni a partire dal semaforo*[…]

- *Noun_to_verb*: when a nominalization or a support verb construction is replaced with a simple verb.

(8) O: *Il giorno **della partenza**, i bambini salutarono i loro genitori durante la colazione.*

*S: Il giorno **in cui i genitori partirono**, i bambini li salutarono durante la colazione.*

- *Verb_to_noun*: to mark the presence of a nominalization or of a support verb construction instead of an original simple verb.

(9) O: *Benedetto era molto arrabbiato e voleva **vendicare** sua sorella*.

S: *Benedetto era molto arrabbiato e voleva **ottenere vendetta** per sua sorella*.

- V*erbal voice*: to signal the transformation of a passive sentence into an active (ex. 10) or vice versa. In our corpus we found only one application of the latter; this finding was expected since passive sentences represent an instance of non-canonical order: they are acquired later by typically developing children (Maratsos, 1974, Bever, 1970; for Italian, Cipriani et al., 1993; Ciccarelli, 1998) and have been reported as problematic for atypical populations, e.g. deaf children (Volpato, 2010). Yet, the "passivization" rule may still be productive in other typologies of texts, where it can happen that the author of the simplification prefers not only to keep, but even to insert, a passive, in order to avoid more unusual syntactic constructs in Italian (such as impersonal sentences). This is also in line with what Bott and Saggion (2014) observed for passives in Spanish text simplification.

(10) O: *Solo il papà di Luisa, "Crispino mangia cracker" era dispiaciuto, perché **era stato battuto da Tonio Battaglia***.

S: *Solo il papà di Luisa era triste, perché **Tonio Battaglia lo aveva battuto***.

- V*erbal features:* Italian is a language with a rich inflectional paradigm and changes affecting verbal features (mood, tense, aspect) have proven useful in discriminating between easy- and difficult-to-read texts in readability assessment task (Dell'Orletta et al., 2011). The easy-to-read texts examined there were also written by experts in text simplification, but their target were adults with limited cognitive skills or a low literacy level. Poor comprehenders also find it difficult to properly master verbal inflectional morphology, and the same has been noticed for other categories of atypical readers, e.g. dyslexics (Fiorin, 2009); thus, there is a probability that the simplification, according to the intended target, will alter the distribution of verbal features over paired sentences, as occurred in (11).

(11) O: *Sembrava veramente che il fiume **stesse** per straripare.*

S: *Il fiume **stava** per straripare*.

4 Conclusions and Perspectives

We have illustrated the first annotation scheme for Italian that includes a wide set of simplification rules spanning across different levels of linguistic description. The scheme was used to annotate the only existing Italian parallel corpus. We believe such a resource will give valuable insights into human text simplification and create the prerequisites for automatic text simplification. Current developments are devoted to refine the annotation scheme, on the basis of a qualitative and quantitative analysis of the annotation results; we are also testing the suitability of the annotation scheme with respect to other corpora we are also gathering in a parallel fashion. Based on the statistical findings on the productivity of each rule, we will investigate whether and in which way certain combinations of rules affect the distribution of multi-leveled linguistic features between the original and the simplified texts. In addition, we intend to explore the relation between text simplification and a related task, i.e. readability assessment, with the aim of comparing the effects of such combinations of rules on the readability scores.

Acknowledgments

The research reported in this paper has been partly supported by a grant from the project "intelligent Semantic Liquid eBook - iSLe", POR CReO 2007 – 2013, Regione Toscana, Italy.

References

S. M. Aluísio, L. Specia, T.A. Pardo, E.G. Maziero, and R. P. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceeding of the eighth ACM symposium on Document engineering*, pp. 240-248.

G. Barlacchi and S. Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing*-2013), pp. 476-487.

T.G. Bever. 1970. The cognitive basis for linguistic structures. In *J.R.Hayes (ed.) Cognition and the development of Language*. New York, Wiley.

S. Bott and H. Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation*, Volume 48, Issue 1, pp. 93-120, Springer Netherlands.

L. Brouwers, D. Bernhard, A-L. Ligozat, and T. François. 2014. Syntactic Sentence Simplification for French. In *The 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*. Gothenburg, Sweden, 27 April.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Association for the Advancement of Artificial Intelligence (AAAI).

H. Caseli, T. Pereira, L. Specia, T. Pardo, C. Gasperin, and S. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009)*, March 01–07, Mexico City.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the international conference on computational Linguistics*, pp. 1041–1044.

L. Ciccarelli. 1998. *Comprensione del linguaggio, dei processi di elaborazione e memoria di lavoro: uno studio in età prescolare*, PhD dissertation, University of Padua.

P. Cipriani, A. M. Chilosi, P. Bottari, and L. Pfanner. 1993. *L'acquisizione della morfosintassi in italiano*: *fasi e processi*. Padova: Unipress.

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

J. De Belder, K. Deschacht, and M-F Moens. 2010. Lexical simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *SLPAT 2011 - Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* (Edimburgo, UK, July 2011), pp. 73-83. Association for Computational Linguistics Stroudsburg, PA, USA.

B. Drndarevic, S. Stajner, and H. Saggion. 2012. Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of "Easy to read on the web" online symposium*.

G. Fiorin. 2009. The Interpretation of Imperfective Aspect in Developmental Dyslexia. In *Proceedings of the 2nd International Clinical Linguistics Conference*, Universidad Autónoma de Madrid, Universidad Nacional de Educación a Distancia, and Euphonia Ediciones.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing*, ACL 2003.

W. Kintsch and J. Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, pp. 257-274.

S. Klerke and A. Søgaard. 2012. Danish parallel corpus for text simplification. In *Proceedings of Language Resources and Evaluation Conference (LREC 2012)*.

M. Maratsos. 1974. Children who get worse at understanding the passive: A replication to Bever. *Journal of Psycholinguistic Research*, 3, pp. 65-74.

Y. Ozuru, K. Dempsey, and D. McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19, pp. 228-242.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Speech and Language Technology for Education (SLaTE)*.

A. Siddharthan. 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC 2002)*.

D. I. Slobin and R. G. Bever. 1982. Children use canonical sentence schemas. A cross-linguistic study of word order and inflections. In *Cognition*, 12(3), pp. 229-265.

F. Volpato. 2010. *The acquisition of relative clauses and phi-features: evidence from hearing and hearing-impaired populations*. PhD dissertation. Ca' Foscari University of Venice.