Design and Annotation of the First Italian Corpus for Text Simplification

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

Abstract

In this paper, we present design and construction of the first Italian corpus for automatic and semi–automatic text simplification. In line with current approaches, we propose a new annotation scheme specifically conceived to identify the typology of changes an original sentence undergoes when it is manually simplified. Such a scheme has been applied to two aligned Italian corpora, containing original texts with corresponding simplified versions, selected as representative of two different manual simplification strategies and addressing different target reader populations. Each corpus was annotated with the operations foreseen in the annotation scheme, covering different levels of linguistic description. Annotation results were analysed with the final aim of capturing peculiarities and differences of the different simplification strategies pursued in the two corpora.

1 Introduction and Background

Automatic Text Simplification (ATS) is receiving growing attention over the last few years due to the implications it has for both machine– and human–oriented tasks. ATS has been employed as a preprocessing step to improve the efficiency of e.g. parsing, machine translation and information extraction. Recently, ATS has been used in educational scenarios and assistive technologies; e.g. for the adaptation of texts to particular readers, like children (De Belder et al., 2010), L2 learners (Petersen and Ostendorf, 2007), people with low literacy skills (Aluísio et al., 2008), cognitive disabilities (Bott and Saggion, 2014) or language impairments, e.g. aphasia (Carroll et al., 1998) or deafness (Inui et al., 2003).

The purpose of ATS, within both perspectives, is to reduce lexical and syntactic complexity while preserving the original meaning of the text. To this aim, three main approaches have been followed. The more traditional one relies on the use of hand-crafted rules (Chandrasekar et al., 1996; Siddharthan, 2002; Siddharthan, 2010; Siddharthan, 2011), which typically cover specific phenomena that are symptoms of linguistic complexity, especially at the syntactic level (e.g. passives, relative clauses, appositions). Recently, the availability of larger parallel corpora, i.e. sentence-aligned corpora consisting of both the original and the simplified version of the same text (e.g. English and Simple English Wikipedia, in short EW and SEW), has allowed a consistent use of machine learning techniques for automatically acquiring simplification rules. This is the approach followed by e.g. Woodsend and Lapata (2011), who based their ATS system on a quasi-synchronous grammar, Zhu et al. (2010), who adapted a Statistical Machine Translation (SMT) algorithm to implement simplification operations on the parse tree, and Narayan and Gardent (2014), who similarly adopted SMT techniques but also combined a deep semantic representation of the sentence. Both hand-written and automatically acquired rules have advantages and shortcomings. While the former can potentially account for the maximum linguistic information, they are extremely costly to develop and tend to cover only a few lexical and syntactic constructs; on the other side, data-driven approaches require the least linguistic knowledge but they are not feasible

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 31–41,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

without a large quantity of aligned data. Hybrid approaches seem to offer a good alternative; as shown by Siddharthan and Angrosh (2014), a system that combines automatically harvested lexical rules with hand-crafted syntactic rules outperformed the state of the art. Besides, all these systems exploit the EW/SEW dataset as a training corpus. Such resources are lacking for languages other than English, making it rather impossible to approach ATS as pure machine learning task. For some of these languages, parallel monolingual corpora are annotated with simplification rules corresponding to transformations to perform on a complex sentence. This is the approach followed by Brouwers et al. (2014) for French; Bott and Saggion (2014) for Spanish; Caseli et al. (2009) for Brazilian Portuguese. A different approach is advanced by Specia (2010) for Brazilian Portuguese, who adopted phrase-based machine learning from a parallel corpus. For Basque, Aranzabe et al. (2013) used the output of a readability assessment system for detecting complex sentences, which are simplified by a large set of hand-crafted rules.

Typically, ATS approaches rely on the output of a syntactic parser although the main cause of errors for an ATS system is due to erroneous parses also when state–of–the–art parsers are used (De Belder et al., 2010; Siddharthan, 2011; Drndarević et al., 2013; Brouwers et al., 2014; Siddharthan and Angrosh, 2014). In particular, this concerns relative clause attachments and clause boundary identification (Siddharthan and Angrosh, 2014). According to Drndarević et al. (2013), one third of ATS errors depends on previous parsing errors and Brouwers et al. (2014) revealed that 89% of text simplification (TS) errors are due to preprocessing errors.

ATS is largely underinvestigated for what concerns Italian. The only exception is (Barlacchi and Tonelli, 2013), who devised a rule–based architecture focusing on a limited set of linguistic structures, but no previous study has addressed ATS by using parallel corpora.

2 Our Contribution

We present the first Italian resource for automatic and semi-automatic text simplification. We collected and hand–aligned two monolingual corpora representative of two different strategies of manual simplification and addressing different target readers. The corpora were annotated with a set of rules designed to capture simplification operations at diverse levels of linguistic description. There are several motivations underlying the proposed approach. As a universal *native simplified–language speaker* does not exist (Siddharthan, 2014), it follows that ATS systems are typically specialized with respect to a specific target user. Hence, we introduce a new annotation scheme able to handle different simplification strategies, at the level of both method and target users.

This is the starting point to develop a flexible automatic or semi–automatic TS system.The proposed resource can be used to train a supervised classifier aimed at carrying out a semi–automatic TS task. In the semi–automatic scenario, the system will be able to identify the areas of linguistic complexity within a sentence and suggest the authors the most appropriate simplification rule for the intended audience and domain. This classifier, using the information extracted from the syntactic tree as one of the features exploited to predict the rules to be applied, is expected to be more robust to syntactic parsing errors than TS systems based on hand–crafted or automatically acquired rules heavily relying on parses transformations. To give an idea of how wrong parses can affect a TS system, let's consider that the accuracy of the state-of-the-art dependency parser for Italian is 87.89% in terms of Labeled Attachment Score corresponding to 293 erroneously parsed sentences out of the total of 376, i.e. 78% of the test sentences contain at least one parsing error.[1]. Moreover, it should be noted that in a TS scenario the parsers are typically tested on domains outside of the data from which they were trained or developed on (i.e. out–domain scenario) and it is widely acknowledged that state–of–the–art statistical parsers have a dramatic drop of accuracy when tested in a out–domain scenario (Gildea, 2001).

In this paper, we also carried out a comparative analysis between different TS strategies addressing different target users: this was possible thanks to

[1] These data are reported in the Evalita evaluation campaign and were obtained by combining four state–of–the–art parsers using a test set with gold (i.e. manually revised) part–of–speech tags (Bosco et al., 2014)

the internal composition of the developed resource, which allowed us to investigate the effects of simplification rules on the linguistic peculiarities of abridged texts with respect to their original versions.

3 Corpora

The annotated resource[2] presented here is made up of two sub–corpora that can be considered representative of two different TS strategies: the "structural" and the "intuitive" strategy, following Allen (2009)'s definition, who addressed TS in the context of L2 learning. The former uses predefined graded lists (covering both word and structural levels) or traditional readability formulas. The latter is dependent on the author's teaching experience and personal judgments about the comprehension ability of learners. Although with main distinctions, this classification can be applied for our purpose.

The first sub–corpus (*Terence*) contains 32 short novels for children and their manually simplified version.[3] 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. To align the corpus, we selected the last two levels of simplification (i.e. local cohesion and lexicon/syntax) which were considered respectively as the original and the simplified version. This was motivated by the need of tackling only those textual simplification aspects with a counterpart at the morpho–syntactic and syntactic level. We hand-aligned the resulting 1036 original sentences to the 1060 simplified ones. The results (Table 1) provide some insights into the typology of human editing operations. 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 underwent a three–fold split (0.19%), i.e. they correspond to three sentences in the simplified version; 15 pairs of original sentences were merged into a single one (2.88%). Finally, the percentage of unaligned sentences is 1%.

The second sub–corpus (*Teacher*) is composed by 24 pairs of original/simplified texts, which were collected by surfing specialized educational websites providing free resources for teachers. They cover different textual genres, such as literature (e.g. extracts from famous Italian novels) and handbooks for high school on diverse subjects (e.g. history, geography), and they are addressed to different targets. Unlike *Terence*, the simplification was performed independently by a teacher, with the aim of adapting the text to the need of audience, typically L2 students with at least a B2 level in Italian. Thus, *Teacher* can be considered as an instance of "intuitive" simplification: while the target is usually the same (i.e. L2 learners), each text was produced by a different author and the interventions made on the text span over different linguistic levels without any predefined distinction or hierarchy. On the contrary, *Terence* exemplifies a "structural" simplification, since: i) it was produced by a pool of experts; ii) it addressed a well–defined target; iii) it was consistent with a predefined guideline tackling the simplification at three separate textual dimensions. This can also explain the higher percentage of texts which were perfectly aligned at sentence level (92.1% see Table 1) with respect to *Teacher* (68.32%).

	1:1	1:2	1:3	2:1	1:0	0:1
Terence	92.1	3.75	0.19	2.88	0.67	0.38
Teacher	68.32	11.45	0.76	13.74	1.15	0.0

Table 1: Percentage of sentence alignments.

To compare the two different simplification strategies with respect to the effect of the simplification process, we evaluated the two corpora with the readability index existing for the Italian language, i.e. READ–IT (Dell'Orletta et al., 2011). For both the corpora, we calculated the Spearman's correlation between the scores obtained by different READ–IT models (i.e. using different types of linguistic features) on the original and the simplified version. As reported in Table 2, the two simplified corpora are significantly correlated with all READ–IT models. In particular, *Teacher* is especially correlated with the model using a combination of *raw text* and *lexical* features (*READ–IT lexical* model in Table 2). This possibly follows from the "intuitive" simplification process of *Teacher* that mostly concerns lexical substitution operations.

[2] http://www.italianlp.it/software-data/

[3] This corpus was produced within the EU project Terence targeting "poor comprehenders": http://www.terenceproject.eu

Readability index	Terence	Teacher
READ–IT global	**0.77**$*$	**0.47**
READ–IT base	**0.80**$*$	**0.50**
READ–IT lexical	**0.65**$*$	**0.72**$*$
READ–IT syntax	**0.54**$*$	**0.46**

Table 2: Spearman's correlation between different READ-IT models and the simplified corpora. Significant correlations ($p < 0.05$) are bolded; those with $p < 0.001$ are also marked with $*$.

The two corpora were annotated by two undergraduate students in computational linguistics, who received preliminary training lessons on the simplification rules covered by the annotation tagset. Each student annotated a different corpus and all their annotations were verified by a trained linguist.

4 Simplification Annotation Scheme

We defined an annotation scheme covering six macro–categories: *split*, *merge*, *reordering*, *insert*, *delete* and *transformation*. Following Bott and Saggion (2014), we used a two–level structure, i.e. for some categories more specific subclassed have been introduced. In Table 3, we show the tagset of the annotation scheme. In the following examples extracted from the annotated corpus, we bolded the text span marked in the original sentence by each rule-tag and we highlighted in italics the corresponding text span in the simplified version.[4]

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 of these sentences 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.

O: Mamma Gorilla sembrava completamente distrutta per le cure che dava al suo vivace cuccioletto Tito, **che stava giocando vicino alle grosse sbarre di acciaio che circondavano il recinto**. [Mummy Gorilla looked completely worn out from looking after her lively baby, Tod, **who was playing by the thick steel bars that surrounded the enclosure.**]

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.* [Mummy Gorilla looked completely worn out from looking after her lively baby Tod. *Tod was playing by the thick steel bars that surrounded the enclosure.*]

Merge: it is to be taken as the reverse of split, i.e. the operation by which two (or more) original sentences are joined into a unique simplified sentence. This 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 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.

O: **Clara pensò che fosse uno dei cigni. Ma poi si rese conto che stava urlando!** [**Clara thought it was one of the swans. But then she realised it was shouting!**]

S: *In un primo momento, Clara pensò che fosse uno dei cigni, ma poi sentì urlare!* [*At first, Clara thought it was one of the swans, but then she heard it shouting.*]

Reordering: this tag marks word order changes between the original sentence and its simplified counterpart. Clearly, altering the position of the elements in a sentence depends, in turn, 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 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 she believes e.g. that it allows the reader to keep the focus stable over two or more sentences.

O: Il passante gli spiegò che, per arrivare al bidone,**doveva contare ben 5 bidoni a partire dal semaforo**. [The passer-by explained him that, to get to the dustbin, **he had to count exactly 5 dustbins starting from the traffic light.**]

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

Simplification Annotation Scheme					
Classes	Subclasses	Terence		Teacher	
Split		1.71	(43)	2.06	(35)
Merge		0.81	(20)	1.30	(22)
Reordering		8.65	(212)	7.89	(134)
Insert	Verb	4.92	(121)	2.53	(43)
	Subject	1.79	(44)	1.94	(33)
	Other	12.01	(295)	11.19	(190)
Delete	Verb	2.04	(50)	1.88	(32)
	Subject	0.49	(12)	0.24	(4)
	Other	19.41	(477)	23.20	(394)
Transformation	Lexical Substitution (word level)	26.50	(651)	20.73	(352)
	Lexical Substitution (phrase level)	13.39	(329)	11.60	(197)
	Anaphoric replacement	0.61	(15)	3.53	(60)
	Noun_to_Verb	1.59	(39)	0.88	(15)
	Verb_to_Noun (nominalization)	0.61	(15)	0.47	(8)
	Verbal Voice	0.53	(13)	0.77	(13)
	Verbal Features	4.92	(121)	9.78	(166)

Table 3: Simplification tagset and the percentage distribution (with its absolute value) for each rule-tag.

S: Il signore spiegò a Ugolino che *doveva contare 5 bidoni a partire dal semaforo*, per arrivare al bidone della carta. [The man explained Little Hugh that *he had to count 5 dustbins starting from the traffic light* to get to the wastepaper dustbin.]

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 reducing the inference load of a text, especially with less skilled or low–knowledge readers (Ozuru et al., 2009), it is difficult to predict what an author will actually add to the original 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 into a lexical noun phrase[5].

O: Essendo da poco andata in pensione dal suo lavoro, disse che le mancavano i suoi studenti [...] [Having just retired from her job, she said that she missed her students]

S: Essendo da poco andata in pensione dal suo lavoro *come insegnante*, disse che le mancavano i suoi studenti [...] [Having just retired from her job *as a school teacher*, she said that she missed her students]

Delete: dropping redundant information is also a strategy for simplifying a text. 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). Such occurrences have been marked with the underspecified *delete* rule; 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).

O: **Sembrò veramente che** il fiume stesse per straripare. [**It really seemed that** the river was going to burst.]

S: Il fiume stava per straripare. [The river was going to burst.]

Transformation: this label covers six typologies of transformations that a sentence may undergo to become more comprehensible for the intended reader. Such modifications can affect the sentence at the lexical, morpho–syntactic and syntactic level, also giving rise to overlapping phenomena. Our annotation

[5] The covert/overt subject realization is an option available in null–subject languages like Italian.

scheme has intended to cover the following phenomena.

– *Lexical substitution (word level)*: when a single word is replaced by another word (or more than one), which is usually a more common synonym or a less specific term.

O: Il **passante** gli spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni a partire dal semaforo. [The **passer-by** explained him that, to get to the dustbin, he had to count exactly 5 dustbins starting from the traffic light.]

S: Il *signore* spiegò a Ugolino che doveva contare 5 bidoni a partire dal semaforo, per arrivare al bidone della carta. [The *man* explained to Little Hug that he had to count 5 dustbins starting from the traffic light, to get to the dustbin.]

Given the relevance of lexical changes in TS, 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). However, synonyms or hypernyms replacements do not cover all the editing options, since we observed that an author might also restate the meaning of the complex word with a multi-word pharaphrase.

O: Tutti si **precipitarono** verso il tendone. [Everyone **rashed** outside the tent.]

S: Tutti si *misero a correre* verso la tenda. [Everyone *came running* outside the tent.]

– *Lexical substitution (phrase level)*: it differs from the previous rule with respect to the "size" of the original unit involved in the substitution, which in this case consists of a phrase. But, similarly to the previous one, the simplified unit can be either a single word or a phrase itself.

O: Persino il tempo era **di buon umore**. [Even the weather was **in a party mood**.]

S: Persino il tempo era *buono*. [Even the weather was *good*.]

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

O: Il passante **gli** spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni [...] [The passer-by explained **him** that, to get to the dustbin, he had to count exactly 5 dustbins]

S: Il signore spiegò *a Ugolino* che doveva contare 5 bidoni a partire dal semaforo [...] [The man explained *to Little Hug* that he had to count 5 dustbins starting from the traffic light]

– *Noun_to_Verb*: when a nominalization or a support verb construction is replaced by a simple verb. In this case, the correspondence between the noun and the verb involved in the transformation had to be suggested by the presence of a similar morphological root.

O: Il giorno **della partenza**, i bambini salutarono i loro genitori durante la colazione. [On the day of their **parents' departure**, the children said their goodbyes to their parents over breakfast.]

S: Il giorno *in cui i genitori partirono*, i bambini li salutarono durante la colazione. [The day *that their parents left*, the children said them goodbye over breakfast.]

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

O: Benedetto era molto arrabbiato e voleva **vendicare** sua sorella. [Ben was very angry and he wanted **to avenge** his sister.]

S: Benedetto era molto arrabbiato e voleva *ottenere vendetta* per sua sorella. [Ben was very angry and he wanted *to get revenge* for his sister.]

– *Verbal voice*: to signal the transformation of a passive sentence into an active or vice versa. Within both the corpora very few examples of the latter were found; this result 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 textual typologies, 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.

O: Solo il papà di Luisa, "Crispino mangia cracker" era dispiaciuto, perché **era stato battuto da Tonio Battaglia**. [Only Louise's Dad, "Cream

Cracker Craig", was disappointed, because **he'd been beaten by Tod Baxter**.]

S: Solo il papà di Luisa era triste, perché *Tonio Battaglia lo aveva battuto*. [Only Louise's Dad was sad, because *Tod Baxter had beaten him*.]

– *Verbal features*: Italian is a language with a rich inflectional paradigm and changes affecting verbal features (mood, tense) have proven useful in discriminating between easy– and difficult–to–read texts in readability assessment task (Dell'Orletta et al., 2011). Poor comprehenders also find it difficult to properly master verb inflectional morphology; the same holds for other categories of atypical readers, e.g. dyslexics (Fiorin, 2009), but also for L2 learners (Sorace, 1993); thus, the simplification, according to the intended target, will probably alter the distribution of verbal features.

O: Non capisco e non **potrei** parlare con nessuno. [I can't understand and I **could** not talk to anybody.]

S: Non capisco e non *posso* parlare di queste cose con nessuno. [I can't understand and I *can* not speak of such things to anybody.]

5 Simplification Rules and Linguistic Features

The analysis of the frequency distribution of each rule within the two annotated corpora (Table 3) allows us capturing similarities and variations across corpora representing two different TS strategies and addressed to diverse categories of readers. The majority of rules are similarly distributed across the two corpora showing that a number of simplification choices are shared by a team of experts and independent teachers. This is an interesting finding as it might suggest the existence of an "independent" simplification process shared by approaches targeting different audience and based on different simplification methods. Exceptions are represented by some rules involving verbs (i.e. transformation of verbal features and insert verb) and anaphoric replacements. For what concerns the latter, it should be noted that the *Terence* original version here adopted inherits previous sentence transformations covering, among others, anaphoric replacements. The different distribution of rules involving verbs might reflect both the different simplification choices related to the *structural* and *intuitive* simplification strategies and the different textual genres included in *Teacher* and *Terence*.

For a more in-depth analysis of the impact and the significance of each simplification rule, we focused on the most frequently applied rules and we chose a set of features which are typically involved in automatic readability assessment and also express language–specific peculiarities. For each linguistic feature, we calculated the Spearman's correlation between the feature values extracted from the original text and from the simplified version with respect to the selected rules.

5.1 Linguistic Features

The set of linguistic features spans across different levels of linguistic analysis and are broadly classifiable into four main classes: raw text, lexical, morpho–syntactic and syntactic features, shortly described below. They were extracted from the corpora automatically tagged by the part–of–speech tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi, 2006).

Raw text features (Features [1–2] in Table 4) are typically used within traditional readability metrics and include *sentence length* (average number of words per sentence), and *word length* (average number of characters per words).

Feature [3] refers to the percentage of all unique words (types) on the *Basic Italian Vocabulary* (*BIV*) by De Mauro (2000) in the sentence. The *BIV* includes a list of 7,000 words highly familiar to Italian native speakers.

The set of morpho–syntactic features [4–19] ranges from the probability distribution of part–of–speech types, to the lexical density of the text, calculated as the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. It also includes verbal mood and tense distributions, a language–specific feature related to Italian rich verbal morphology.

The set of syntactic features [20–35] captures different aspects of the syntactic structure, such as:
– **parse tree depth features**, going from the *depth of the whole parse tree* [26], calculated in terms of the longest path from the root of the dependency tree to some leaf, to a more specific feature referring to the *average depth of embedded complement 'chains'* [23] governed by a nominal head and including ei-

ther prepositional complements or nominal and adjectival modifiers;
– **verbal predicate features**, going from the *arity of verbs* [27], meant as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers), to the *distribution of verbal roots with explicit subject* [28] with respect to all sentence roots occurring in a text and the *relative ordering of subject and object with respect to the verbal head* [29–32].
– **subordination features** include the *distribution of subordinate vs. main clauses* [20–21]; for subordinates, their *relative ordering with respect to the main clause* [33–34] and the *average depth of 'chains' of embedded subordinate clauses* [22];
– the **length of dependency links** is calculated in terms of the words occurring between the syntactic head and the dependent: the feature includes the *length of all dependency links* [24] and the *maximum dependency links* [25];
– **clause length** [35] is measured as the number of tokens occurring within a clause.

5.2 Correlation

Table 4 illustrates the correlations between the linguistic features and the most frequently applied simplification rules. It can be noted that all the rules are strongly correlated with the linguistic features. This reveals that these rules have a great impact on the linguistic structure of the simplified text. It also shows the effectiveness of such features to capture simplification operations at varying degrees of linguistic description. Interestingly, if we examine more in-depth the significance value, we can observe a distinction between the two corpora. *Terence* reports a higher number of stronger correlations (i.e. $p < 0.001$) with respect to *Teacher*. These results seem to provide an evidence to the existence of different simplification strategies, which vary according to the person (i.e. expert vs. non-expert), textual genres and intended target. Specifically, the teachers prefer a more vocabulary-oriented simplification approach, as testified by *a)* the highest significant correlations reported by the rules dealing with lexical replacements (i.e. *LexSub_word* and *LexSub_phrase*) and *b)* the fact that the majority of significant correlations at > 0.5 affects linguistic features from [1] to [19], i.e. features not dealing with the syntactic structure. This might suggest that, independently from the simplification rule adopted, the resulting sentence has not undergone a strong modification in its grammatical structure. This is not the case of the "structural" simplification, in which all the rules significantly correlate with both lexical/morpho–syntactic features (set [1-19]) and syntactic features (set [20-35]). On the other side, the correlation results reported by the *Delete*, *LexSub_word* and *LexSub_phrase* rules reveal the existence of a common approach to simplification. In the two corpora these rules are correlated with mainly the same linguistic features.

For what concerns the evaluation of the overall significance of each rule, we observe that a wide number of correlations at ≥ 0.6 occurs especially when *Split* and *LexSub_word* were applied. Both these simplification operations are expected to greatly redefine the structure of the sentence; a split e.g. not only correlates with sentence length, but it also reduces prepositional chains [23]. Split might be triggered by long noun phrases with a deverbal noun; to simplify them the author could have chosen to turn them into an autonomous sentence, by also adding a verb (see the high correlation between [23] and *InsertVerb*).

6 Conclusion

We have presented the first Italian corpus for text simplification. This annotated resource is composed by two monolingual parallel corpora, representing two different strategies of simplification: "structural" and "intuitive". We have defined an annotation scheme able to capture manual simplifications at different levels of linguistic structure as well as to handle the different strategies of simplification. We have carried out an in-depth analysis of the impact of each simplification rule with respect to a set of linguistic features related to text complexity. This study has highlighted the existence of an "independent" simplification process shared by the two considered simplification approaches targeting different audience. We are currently using this finding in the development of a semi–automatic supervised TS system trained on the two corpora able to handle these shared simplification phenomena. Current developments are also devoted to refining the anno-

Feature	Insert		Delete		Reord		LexSub_word		LexSub_phrase		Split		InsertVerb	
[1] Sentence length	**.796∗**	**.342**	**.772∗**	**.345∗**	**.820∗**	**.451∗**	**.818∗**	**.463∗**	**.787∗**	**.433∗**	**.799∗**	**.501**	**.714∗**	**.573∗**
[2] Word length	**.595∗**	**.431∗**	**.593∗**	**.518∗**	**.627∗**	**.637∗**	**.636∗**	**.559∗**	**.512∗**	**.449∗**	**.700∗**	**.581**	**.612∗**	**.375**
[3] Word types in the *BIV*	**.663∗**	**.315**	**.707∗**	**.382∗**	**.699∗**	**.456∗**	**.735∗**	**.580∗**	**.654∗**	**.472∗**	**.630∗**	**.865∗**	**.690∗**	**.413**
[4] Lexical density	**.639∗**	**.246**	**.685∗**	**.416∗**	**.704∗**	**.410∗**	**.757∗**	**.400∗**	**.617∗**	**.402∗**	**.646∗**	**.696∗**	**.566∗**	.082
[5] Adjective	**.693∗**	**.450∗**	**.689∗**	**.406∗**	**.752∗**	**.564∗**	**.724∗**	**.585∗**	**.726∗**	**.527∗**	**.779∗**	**.662**	**.787∗**	.245
[6] Adverb	**.546∗**	**.324**	**.652∗**	**.424∗**	**.667∗**	**.311**	**.729∗**	**.445∗**	**.581∗**	**.245**	**.670∗**	.292	**.716∗**	**.351**
[7] Coord Conjunction	**.609∗**	**.345**	**.707∗**	**.454∗**	**.735∗**	**.588∗**	**.765∗**	**.554∗**	**.746∗**	**.494∗**	**.474**	**.662**	**.667∗**	.306
[8] Subord Conjunction	**.510∗**	**.532∗**	**.611∗**	**.478∗**	**.564∗**	**.606∗**	**.700∗**	**.483∗**	**.716∗**	**.414∗**	**.726∗**	**.554**	**.641∗**	**.441**
[9] Preposition	**.687∗**	**.492∗**	**.678∗**	**.404∗**	**.690∗**	**.354**	**.794∗**	**.498∗**	**.680∗**	**.447∗**	**.688∗**	**.491**	**.743∗**	**.480**
[10] Pronoun	**.619∗**	.179	**.629∗**	**.277**	**.550∗**	**.304**	**.716∗**	**.317∗**	**.594∗**	**.338∗**	**.552∗**	**.578**	**.368∗**	-.030
[11] Noun	**.707∗**	**.566∗**	**.702∗**	**.586∗**	**.708∗**	**.474∗**	**.761∗**	**.601∗**	**.721∗**	**.548∗**	**.666∗**	**.544**	**.728∗**	**.490**
[12] Verb	**.703∗**	**.401∗**	**.634∗**	**.464∗**	**.655∗**	**.435∗**	**.722∗**	**.506∗**	**.653∗**	**.468∗**	**.743∗**	**.679**	**.656∗**	.268
[13] Verb infinitive mood	**.718∗**	**.488∗**	**.644∗**	**.481∗**	**.649∗**	**.440∗**	**.752∗**	**.528∗**	**.720∗**	**.459∗**	**.554∗**	**.753∗**	**.395∗**	**.405**
[14] Verb gerundive mood	**.574∗**	nan	**.585∗**	nan	**.554∗**	nan	**.691∗**	-.038	**.677∗**	nan	**.499∗**	nan	**.519∗**	**.558∗**
[15] Verb participle mood	**.530∗**	.210	**.439∗**	**.395∗**	**.380∗**	**.323**	**.554∗**	**.335∗**	**.349∗**	**.368∗**	**.527∗**	.204	**.371∗**	.148
[16] Verb indicative mood	**.584∗**	.223	**.630∗**	**.422∗**	**.581∗**	.100	**.697∗**	**.344∗**	**.675∗**	**.323**	**.686∗**	**.495**	**.491∗**	.156
[17] Verb present tense	**.573∗**	**.254**	**.622∗**	**.307**	**.574∗**	**.275**	**.683∗**	**.394∗**	**.558∗**	**.296**	**.599∗**	**.568**	**.727∗**	**.527**
[18] Verb imperfect tense	**.741∗**	**.638∗**	**.786∗**	**.533∗**	**.768∗**	**.635∗**	**.849∗**	**.542∗**	**.771∗**	**.479∗**	**.813∗**	**.884∗**	**.777∗**	**.432**
[19] Verb past tense	**.703∗**	.214	**.832∗**	.088	**.787∗**	.080	**.840∗**	**.260∗**	**.811∗**	.187	**.902∗**	nan	**.801∗**	**.504**
[20] Main clauses	**.492∗**	.215	**.395∗**	**.198**	**.495∗**	.046	**.520∗**	**.215**	**.518∗**	.191	**.337**	.000	**.277**	.097
[21] Subord clauses	**.492∗**	.215	**.395∗**	**.204**	**.495∗**	.151	**.520∗**	**.209**	**.518∗**	**.254**	**.337**	.145	**.277**	.238
[22] Embedded subord clauses	**.356∗**	**.303**	**.478∗**	**.351∗**	**.369∗**	**.323**	**.529∗**	**.415∗**	**.463∗**	**.404∗**	**.422**	**.472**	**.499∗**	.173
[23] Prepositional 'chains'	**.647∗**	**.352**	**.547∗**	**.305**	**.679∗**	.225	**.740∗**	**.424∗**	**.627∗**	**.514∗**	**.724∗**	**.712∗**	**.664∗**	**.507**
[24] Length of dependency links	**.608∗**	**.403∗**	**.567∗**	**.431∗**	**.457∗**	**.278**	**.619∗**	**.433∗**	**.571∗**	**.468∗**	**.498∗**	.215	**.512∗**	**.562∗**
[25] Longest dependency links	**.643∗**	**.321**	**.582∗**	**.345∗**	**.523∗**	**.307**	**.621∗**	**.428∗**	**.599∗**	**.493∗**	**.514∗**	.160	**.578∗**	**.596∗**
[26] Parse tree depth	**.559∗**	.166	**.586∗**	**.275**	**.506∗**	**.280**	**.671∗**	**.379∗**	**.602∗**	**.405∗**	**.509∗**	.376	**.499∗**	.294
[27] Verb arity	**.630∗**	.231	**.518∗**	**.236**	**.417∗**	.191	**.588∗**	**.365∗**	**.548∗**	**.321**	**.494**	.019	**.511∗**	.003
[28] Verbal roots with subj	**.469∗**	.182	**.583∗**	**.324∗**	**.438∗**	**.331**	**.585∗**	**.347∗**	**.473∗**	**.365∗**	.017	.439	**.614∗**	.216
[29] Post-verbal obj	**.566∗**	.224	**.570∗**	.178	**.471∗**	**.288**	**.634∗**	**.389∗**	**.575∗**	**.228**	**.573∗**	.162	**.511∗**	.082
[30] Pre-verbal obj	**.416∗**	**.340**	**.524∗**	**.227**	**.380∗**	**.605∗**	**.616∗**	**.307∗**	**.519∗**	**.315**	**.670∗**	-.076	**.619∗**	-.065
[31] Post-verbal subj	**.363∗**	.204	**.381∗**	**.294**	**.207**	**.500∗**	**.521∗**	**.349∗**	**.266∗**	**.228**	**.615∗**	**.570**	**.344∗**	.343
[32] Pre-verbal subj	**.476∗**	.141	**.498∗**	.163	**.220**	.076	**.568∗**	**.326∗**	**.328∗**	**.324**	**.441**	.089	**.572∗**	-.024
[33] Post-verbal subord clauses	**.552∗**	**.337**	**.534∗**	**.336∗**	**.488∗**	**.260**	**.647∗**	**.469∗**	**.528∗**	**.388∗**	**.505∗**	**.556**	**.385∗**	.052
[34] Pre-verbal subord clauses	**.299∗**	.155	**.378∗**	**.233**	**.445∗**	.105	**.495∗**	**.159**	**.308∗**	.085	**.315**	**.444**	**.424∗**	-.100
[35] Clause length	**.707∗**	**.485∗**	**.592∗**	**.481∗**	**.635∗**	**.388**	**.711∗**	**.513∗**	**.659∗**	**.450∗**	**.637∗**	**.514**	**.622∗**	**.462**

Table 4: Spearman's correlation between the most frequent rules and a subset of linguistic features. Significant correlations ($p < 0.05$) are bolded; those with $p < 0.001$ are also marked with $*$. For each column, the left value refers to *Terence*, the right value to *Teacher*.

tation scheme, also by testing the suitability of this scheme for other corpora.

References

D. Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4): 585–599.

S. M. Aluísio, L. Specia, T. A. Pardo, E. G. Maziero and R. P. de Mattos Fortes. 2008. Towards Brazilian Portoguese automatic text simplification systems. *Proceedings of the eighth ACM symposium on Document engineering*, 240–248.

M. J. Aranzabe, A. D. De Ilarraza, I. Gonzalez-Dios 2013. Transforming complex sentences using dependency trees for automatic text simplification in Basque. *Procesamiento del lenguaje natural*, 50, 61–68.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, 166–170.

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

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

C. Bosco, F. Dell'Orletta, S. Montemagni, M. Sanguinetti and M. Simi. 2014. The Evalita 2014 Dependency Parsing Task. *Proceedings of Evalita'14, Evaluation of NLP and Speech Tools for Italian*, Pisa, December.

S. Bott and H. Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation*, 48(1): 93–120.

L. Brouwers, D. Bernhard, A.-L. Ligozat and T. François. 2014. Syntactic Sentence Simplification for French. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 47–56.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. *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*. Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. *Proceedings of the international conference on computational Linguistics*, 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. *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*.

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

T. De Mauro. 2000. *Il dizionario della lingua italiana*. Torino, Paravia.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplificatio. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, 73–83.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

B. Drndarević, S. Štajner, S. Bott, S. Bautista and H. Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. A. Gelbukh (ed.) *14th Conference on Computational Linguistics and Natural Language Processing (CICLing'14)*, LNCS 7817 (2):488–500.

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

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

D. Gildea. 2001. Corpus variation and parser performance. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA.

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

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

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

S. Narayan and C. Gardent. 2014. Hybrid Simplification using Deep Semantics and Machine Translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 435–445.

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): 228–242.

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

A. Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2): 259–298.

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

A. Siddharthan and M.A. Angrosh. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*.

A. Siddharthan. 2010. Complex lexico-syntactic reformulation of sentences using typed dependency representations. *Proceedings of the 6th International Natural Language Generation Conference*, 125-133.

A. Siddharthan. 2011. Text Simplification Using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG'11)*, Nancy, France: 2–11.

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

A. Sorace. 1993. Incomplete vs. divergent representations of unaccusativity in non native grammars of Italian. *Second Language Research*, 9(1), 22–47.

L. Specia. *Translating from complex to simplified sentences.* Computational Processing of the Portoguese Language, 6001:30–39.

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.

K.Woodsend and M. Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 409–420.

Z. Zhu, D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. *Proceedings of the 23rd international conference on computational linguistics*, 1353–1361.