SIMPITIKI: a Simplification corpus for Italian

Sara Tonelli
Fondazione Bruno Kessler
satonelli@fbk.eu

Alessio Palmero Aprosio
Fondazione Bruno Kessler
aprosio@fbk.eu

Francesca Saltori
Fondazione Bruno Kessler
fsaltori@fbk.eu

Abstract

English. In this work, we analyse whether Wikipedia can be used to leverage simplification pairs instead of Simple Wikipedia, which has proved unreliable for assessing automatic simplification systems, and is available only in English. We focus on sentence pairs in which the target sentence is the outcome of a Wikipedia edit marked as 'simplified', and manually annotate simplification phenomena following an existing scheme proposed for previous simplification corpora in Italian. The outcome of this work is the SIMPITIKI corpus, which we make freely available, with pairs of sentences extracted from Wikipedia edits and annotated with simplification types. The resource contains also another corpus with roughly the same number of simplifications, which was manually created by simplifying documents in the administrative domain.

Italiano. *In questo lavoro si analizza la possibilità di utilizzare Wikipedia per selezionare coppie di frasi semplificate. Si propone questa soluzione come un'alternativa a Simple Wikipedia, che si è dimostrata inattendibile per studiare la semplificazione automatica ed è disponibile solo in inglese. Ci concentriamo soltanto su coppie di frasi in cui la frase target è indicata come il frutto di una modifica in Wikipedia, indicata dagli editor come un caso di semplificazione. Tali coppie sono annotate manualmente secondo una classificazione delle tipologie di semplificazione già utilizzata in altri studi, e vengono rese liberamente disponibili nel corpus SIMPITIKI. La risorsa include anche un secondo corpus, contenente circa lo stesso numero di semplificazioni, realizzato intervenendo manualmente su alcuni documenti nel dominio amministrativo.*

1 Introduction

In recent years, the shift of interest from rule-based to data-driven automated simplification has led to new research related to the creation of simplification corpora. These are parallel monolingual corpora, possibly aligned at sentence level, in which source and target are an original and a simplified version of the same sentence. This kind of corpora is needed both for training automatic simplification systems and for their evaluation. For English, several approaches have been evaluated based on the Parallel Wikipedia Simplification corpus (Zhu et al., 2010), containing around 108,000 automatically aligned sentence pairs from cross-linked articles between Simple and Normal English Wikipedia. Although this resource has boosted research on data-driven simplification, it has some major drawbacks, for example its availability only in English, the fact that automatic alignment between Simple and Normal versions shows poor quality, and that only around 50% of the sentence pairs correspond to real simplifications (according to a sample analysis performed on 200 pairs by Xu et al. (2015)). In this work, we present a study aimed at assessing the possibility to leverage a simplification corpus from Wikipedia in a semi-automated way, starting from Wikipedia edits. The study is inspired by the work presented in Woodsend and Lapata (2011), in which a set of parallel sentences was extracted from Simple Wikipedia revision history. However, the present work is different in that: *(i)* we use the Italian Wikipedia revision history, demonstrating that the approach can be applied also to languages other than English and on edits of Wikipedia that were not created for educational purposes, and *(ii)* we

manually select the actual simplifications and label them following the annotation scheme already applied to other Italian corpora. This makes possible the comparison with other resources for text simplification, and allows a seamless integration between different corpora.

Our methodology can be summarised as follows: we first select the edited sentence pairs which were commented as 'simplified' in Wikipedia edits, filtering out some specific simplification types (Section 3). Then, we manually check the extracted pairs and, in case of simplification, we annotate the types in compliance with the existing annotation scheme for Italian (Section 4). Finally, we analyse the annotated pairs and compare their characteristics with the other corpora available for Italian (Section 5).

2 Related work

Given the increasing relevance of large corpora with parallel simplification pairs, several efforts have been devoted to develop them. The most widely used corpus of this kind is the Parallel Wikipedia Simplification corpus (Zhu et al., 2010), which was automatically leveraged by extracting normal and simple Wikipedia sentence pairs. However, Xu et al. (2015) have recently presented a position paper, in which they describe several shortcomings of this resource and recommend the research community to drop it as the standard benchmark for simplification. Other alternative approaches, suggesting to further refine the selection of normal – Simple parallel sentences to target specific phenomena like lexical simplification, have been also proposed (Yatskar et al., 2010), but have had limited application. The fact that Simple Wikipedia is not available for languages other than English has proved beneficial to the development of alternative resources. Manually or automatically created corpora have been proposed among others for Brazilian Portuguese (Pereira et al., 2009), German (Klaper et al., 2013) and Spanish (Bott and Saggion, 2011). As for Italian, the only available corpus containing parallel pairs of simplified sentences is presented in Brunato et al. (2015). We borrow from this study the annotation scheme for our corpus, so that we can make a comparison between the two resources. We include in the comparison also another novel corpus, made of manually simplified sentences in the administrative domain, which we release together with the Wikipedia-based one.

3 Corpus extraction

The extraction of the pairs has been performed using the dump for the Italian Wikipedia available on a dedicated website.[1] This huge XML file (more than 1 TB uncompressed) contains the history of every operation of editing in every page in Wikipedia since it has been published for the first time. In particular, the Italian edition of Wikipedia contains 1.3M pages and is maintained by around 2.500 active editors, who made more than 60M edits in 15 years of activity. The Italian language is spoken by 70M people, therefore there are on average 35 active editors per million speakers, giving to the Italian Wikipedia the highest ratio among the 25 most spoken languages around the world.

We parse the 60M edits using a tool in Java developed internally and freely available on the SIMPITIKI website.[2] The user who edits a Wikipedia page can insert a text giving information on why he or she has modified a particular part of the article. This action is not mandatory, but it is included most of the times. We first select the edits which description includes word such as "semplificato" (simplified), "semplice" (simple), "semplificazione" (simplification), and similar. Then, the obtained set is further filtered by removing edits marked with technical tags such as "Template", "Protected page", "New page". This eliminates, for instance, simplifications involving the page template and not the textual content. The text in the Wikipedia pages is written using the Wiki Markup Language, therefore it needs to be cleaned. We use the Bliki engine[3] for this task. Finally, the obtained list of cleaned text passages is parsed using the Diff Match and Patch library,[4] identifies the parts of each article where the text was modified. With this process, we obtain a list of 4,356 sentence pairs, where the differences between source and target sentence are marked with *deletion* and *insertion* tags (see Figure 1).

4 Corpus annotation

We manually annotate pairs of sentences through a web interface developed for the purpose and freely available for download.[2] Differently from

[1] https://dumps.wikimedia.org/
[2] https://github.com/dhfbk/simpitiki
[3] http://bit.ly/bliki
[4] http://bit.ly/diffmatchpatch

Figure 1: Annotation interface used to mark simplification phenomena in the SIMPITIKI corpus.

corpora specifically created for text simplification, in which modifications are almost always simplifications, annotating Wikipedia edits is challenging because the source sentence may undergo several modifications, being partly simplifications and partly other types of changes. Therefore, the interface includes the possibility to select only the text segments in the source and in the target sentence that correspond to simplification pairs, and assign a label only to these specific segments. It also gives the possibility to skip the pair if it does not contain any simplification.

A screenshot of the annotation tool is displayed in Figure 1. On the left, the source sentence(s) are reported, with the modified parts marked in red (as given by the Diff Match and Patch library). On the right, the target sentence(s) were displayed, with segments marked in green to show which parts were introduced during editing. A tickbox next to each red/green segment could be selected to align the source and target segments that correspond to a modification. The annotation interface provides the possibility to choose one of the simplification types proposed in a dropdown menu ('Conferma'), or to skip the pair ('Vai Avanti'). The second option was given to mark the sentences where a modification did not correspond to a proper simplification. For example the last edit shown in Fig. 1 reports in the original version 'Contando *esclusivamente* sulla capacità del mare', which was modified into 'Contando *soprattutto* sulla capacità del mare'. Since this change affects the meaning of the sentence, turning *exclusively* into *mainly*, but not its readability, the pair was not annotated.

In order to develop a corpus which is compliant with the annotation scheme already used in previous works on simplification, we followed the simplification types described in (Brunato et al., 2015). The tagset is reported in Table 1 and comprises 6 main classes (Split, Merge, Reordering, Insert, Delete and Transformation) and some subclasses to better specify the Insert, Delete and Transformation operations. The labels are available in the dropdown menu on the annotation interface and can be used to tag selected pairs of sentences.

Class	**Subclass**
Split	
Merge	
Reordering	
Insert	Verb Subject Other
Delete	Verb Subject Other
Transformation	Lexical substitution (word) Lexical substitution (phrase) Anaphoric replacement Verb to Noun (nominalization) Noun to Verb Verbal voice Verbal features

Table 1: Simplification classes and subclasses. For details see Brunato et al. (2015).

5 Corpus analysis

So far, annotators viewed 2,671 sentence pairs, 2,326 of which were skipped because the target sentence was not a simplified version of the source one. 345 sentence pairs with 575 annotations are currently part of the SIMPITIKI corpus, and all

Class	Subclass	# wiki	# PA	Total
Split		20	18	38
Merge		22	0	22
Reordering		14	20	34
Insert	Verb	11	5	16
Insert	Subject	5	1	6
Insert	Other	58	21	79
Delete	Verb	12	1	13
Delete	Subject	17	1	18
Delete	Other	146	31	177
Transformation	Lexical Substitution (word level)	96	253	349
Transformation	Lexical Substitution (phrase level)	143	184	327
Transformation	Anaphoric replacement	14	3	17
Transformation	Noun to Verb	3	32	35
Transformation	Verb to Noun (nominalization)	2	0	2
Transformation	Verbal Voice	2	1	3
Transformation	Verbal Features	10	20	30
Total		575	591	1166

Table 2: Number of simplification phenomena annotated in the Wikipedia-based and the public administration (PA) corpus

phenomena presented in the annotation scheme proposed by (Brunato et al., 2015) are currently covered.

As a comparison, we analyse also the content of the annotated corpora described in (Brunato et al., 2015), which represent the only existing corpora for Italian simplification. These include the *Terence corpus* of children stories, which was specifically created to address the needs of poor comprehenders, and contains 1,036 parallel sentence pairs, and the *Teacher corpus*, a set of documents simplified by teachers for educational purposes, containing 357 sentence pairs. Besides, we include in the comparison also another corpus, which we manually created by simplifying documents issued by the Trento Municipality to rule building permits and kindergarten admittance. This corpus was simplified following the instructions in (Brunato et al., 2015) but pertains to a different domain, i.e. public administration (PA). The wikipedia-based and the PA corpus have a comparable size (575 vs. 591 pairs), but the simplification phenomena have a different frequency, as shown in Table 2.

In Fig. 2 we compare the distribution of the different simplification types across the four corpora. The graph shows that the same phenomena such as subject deletion, nominalizations, transfer of verbal voice tend to be rare across the four datasets. Similarly, the three top-frequent simplification types, i.e. delete-other, word transformation and phrase transformation, are the same across the four datasets. However, in the Wikipedia-based corpus, word transformation is less frequent than in the other document types, while phrase transformation is much more present. This may show that the 'controlled' setting, in which the Terence and the Teacher corpora were created, may lead educators to put more emphasis on word-based transformations to teach synonyms, while in a more 'ecological' setting like Wikipedia the performed simplifications are not guided or constrained, and phrase-based transformations may sound more natural. As for the PA documents, transformation phenomena are probably very frequent because of the technical language characterised by domain-specific words, which tend to be replaced by more common ones during manual simplification. In this corpus, noun-to-verb transformations are particularly frequent, since nominalizations are typical phenomena of the administrative language affecting its readability (Cortelazzo and Pellegrino, 2003).

While the *Terence* corpus contains on average 2.1 annotated phenomena per sentence pair, *Teacher* 2.8 and the *PA corpus* 2.9 , the Wikipedia-based corpus includes only 1.6 simplifications for each parallel pair. As expected, corpora that were explicitly created for simplification tend to have a higher concentration of simplification phenomena than corpora developed in less controlled settings.

As for non simplifications discarded during the

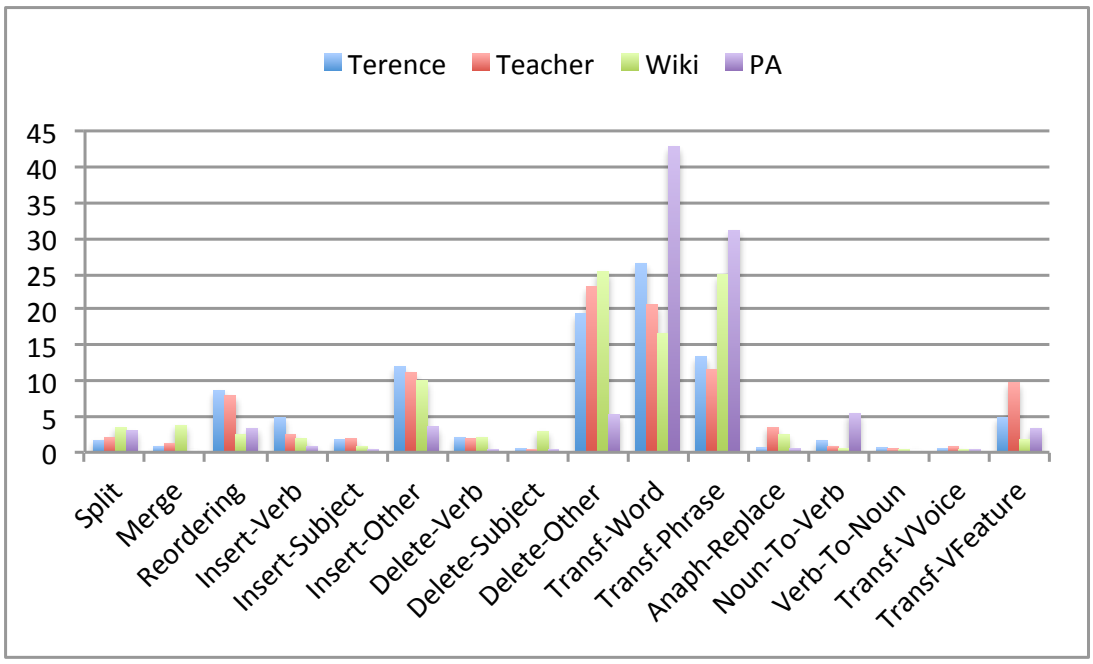

Figure 2: Distribution of the simplification phenomena covered in the Terence, Teacher and Wikipedia-based and Public Administration corpora.

1. Lo psicodramma è *stato il precursore di tutte le* forme di psicoterapia di gruppo 2. Lo psicodramma è *in relazione con altre* forme di psicoterapia di gruppo
1. Partigiani non comunisti e giornalisti *democratici* furono uccisi per il loro coraggio 2. Partigiani non comunisti e giornalisti furono uccisi per il loro coraggio
1. Il dispositivo di memoria di massa utilizza memoria allo stato solido, *ovvero basata su un semiconduttore* 2. Il dispositivo di memoria di massa *basata su semiconduttore* utilizza memoria allo stato solido

Table 3: Examples of parallel pairs which were not annotated as simplifications.

creation of the Wikipedia-based corpus, they include generalizations, specifications, entailments, deletions, edits changing the meaning, error corrections, capitalizations, etc. (see some examples in Table 3). These types of modifications are very important because they may represent negative examples for training machine learning systems that recognize simplification pairs.

6 Conclusions and Future work

We presented a study aimed at the extraction and annotation of a corpus for Italian text simplification based on Wikipedia. The work has highlighted the challenges and the advantages related to the use of Wikipedia edits. Our goal is to propose this resource as a testbed for the evaluation of Italian simplification systems, as an alternative to other existing corpora created in a more 'controlled' setting. The corpus is made available to the research community together with the tools used to create it. The SIMPITIKI resource contains also a second corpus, of comparable size, which was created by manually simplifying a set of documents in the administrative domain. This allows cross-domain comparisons of simplification phenomena.

In the future, this work can be extended in several directions. We plan to use the simplification pairs in this corpus to train a classifier with the goal of distinguishing between simplified and not-simplified pairs. This could extend the gold standard with a larger set of "silver" data by labelling all the remaining candidate pairs extracted from Wikipedia. Besides, the SIMPITIKI methodology is currently being used to create a similar corpus for Spanish, using the same annotation interface. The outcome of this effort will allow multilingual studies on simplification.

Finally, we plan to evaluate the Ernesta system for Italian simplification (Barlacchi and Tonelli, 2013) using this corpus. Specifically, since different simplification phenomena are annotated, it would be interesting to perform a separate evaluation on each class, as suggested in (Xu et al.,

2015).

Acknowledgments

The research leading to this paper was partially supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819).

References

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, pages 476–487, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA, June. Association for Computational Linguistics.

M. Cortelazzo and F. Pellegrino. 2003. *Guida alla scrittura istituzionale*. Laterza, New York, US.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, Sofia, Bulgaria*, pages 11–19.

Tiago F. Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Ra M. Aluisio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *In: 10th Conference on Intelligent Text Processing and Computational Linguistics, Mexico City*, pages 59–70.

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

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.