Italian and English Sentence Simplification: How Many Differences?

Martina Fieromonte•, Dominique Brunato◇, Felice Dell'Orletta◇, Giulia Venturi◇
• University of Pavia
m.fieromonte@gmail.com
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato,felice.dellorletta,giulia.venturi}@ilc.cnr.it

Abstract

The paper proposes a cross-linguistic analysis of two parallel monolingual corpora conceived for automatic text simplification in two languages, Italian and English. The aim is to find similarities and differences in the process of simplification in two typologically different languages. To carry out the comparison, 1,000 sentences were extracted from the two corpora and annotated with a scheme previously used to annotate simplification phenomena.[1]

1 Introduction

In recent years, the availability of parallel monolingual corpora has boosted the adoption of data-driven techniques for the task of automatic text simplification (ATS). These corpora are in general aligned at sentence level and consist of complex sentences paired with their simple version. However, except for English which can rely on two large parallel corpora, i.e. the Parallel Wikipedia Corpus[2] (Coster and Kauchak, 2011)(ParWik) and the Newsela corpus[3] (Xu et al., 2015), these corpora are scarce or rather small in other languages. To reduce time and effort required for the construction of parallel corpora, some works tried new approaches to automatically or semi-automatically collect such resources, e.g. Coster and Kauchak (2011),Yatskar et al. (2010), Brunato et al. (2016), Tonelli et al. (2016). Moreover to take advantage of empirical data, most of these resources were annotated with rules aimed at identifying the typologies of modifications an original sentence goes through during the process of simplification. The inspection can be considered useful for several reasons: it permits i) to detect and classify a set of necessary transformations in TS, ii) to assess if a given corpus complies with user requirements and simplification tasks and iii) to evaluate the impact of simplification operations on target populations. If the corpus investigation also encompasses a cross-linguistic comparison, it might also shed light on peculiarities and similarities underlying the process of simplification across languages. However, so far this last issue has been rather ignored with the exception of Gonzalez-Dios et al. (2018), who compared how macro-simplification operations derived from different annotation schemes are distributed in Italian, Basque and Spanish parallel corpora. This paper intends to explore this under-investigated perspective and proposes a cross-linguistic analysis of two parallel monolingual corpora, i.e. the Italian corpus PaCCSS-IT (Parallel Corpus of Complex–Simple Aligned Sentences for ITalian) (Brunato et al., 2016) and the English Parallel Wikipedia Corpus (Coster and Kauchak, 2011). Through this comparison, the paper tries to answer the following three questions:

1. To what extent can an annotation scheme conceived for the annotation of simplification in one language be used to annotate simplifications in other language?

2. Are there any differences or similarities in the distribution and nature of simplification operations in the two languages?

3. If we find differences, to what extent do they depend on language only, or on the type of corpora?

To answer these questions, 1,000 paired sentences were extracted from the two corpora and annotated with the scheme described in Brunato et al. (2016). This allows us to carry out a quantitative and qualitative analysis focused on understanding the nature of the modifications occurring in the datasets.

[2]http://www.cs.pomona.edu/∼dkauchak/simplification/

[3]https://newsela.com/data/

2 Related work

Given the relevance of parallel monolingual corpora in ATS, many projects have driven their attention on the development of these resources. The main approaches in the literature vary from the manual simplification of original texts carried out by experts (see e.g. Xu et al. (2015) in English, Bott and Saggion (2014) in Spanish, Brunato et al. (2015) in Italian), to the alignment of already existing text collections, containing same-topic documents written in two different styles, a complex and a simple one. It is the case of e.g. Coster and Kauchak (2011) and Tonelli et al. (2016), both relying on the Wikipedia corpus but in a different way. The first is based on the alignment between articles extracted from the standard and the Simple English Wikipedia, a project started in 2001 containing English Wikipedia pages written in basic English; the latter relies on the edits that users had made on the Italian Wikipedia and explicitly marked as instances of simplification. A further strategy was envisaged by Brunato et al. (2016), who first collected a corpus of sentences sharing the same meaning from a large web corpus, and then ranked the most similar pairs according to their linguistic complexity assigned by an automatic readability assessment system.

In many cases, existing ATS corpora were also annotated with rules to make explicit the most frequent operations occurring in the process of sentence simplification and distinguishing different typologies of linguistic phenomena involved in sentence transformation. The classification of simplification operations is typically two-level based, i.e. it contains a few macro-level operations and for some of them a more specific subclass which can depend on the size of the unit affected (e.g. sentence, phrase or word) or the linguistic level at which the operation applies (i.e. lexical, syntactic, discourse). Comparing ParWik with the manually simplified corpus Newsela, Xu et al. (2015) also noticed that the approach adopted to construct ATS resources has an impact on the type of simplification phenomena. For instance, there are more differences between paired sentences before and after simplification in Newsela, suggesting that complex linguistic structures are often retained in ParWik. Simple sentences in ParWik contains also longer words, together with a greater number of function words and punctuation. Similar differences related to the approach underlying the construction of parallel corpora were also observed in Italian. For example, the comparison reported in Tonelli et al. (2016) between a corpus of Wikipedia edit stories and two corpora of heterogeneous texts for young readers manually simplified according to different strategies (i.e. a structural and an intuitive one) proved the existence of differences in terms of the linguistic level affected by simplification. They concern for instance the distribution of some simplification operations and the average of operations per sentence. As regards the first aspect, in manually simplified corpora, editors opted for a word-level lexical substitution, while Wikipedia editors for a phrase-level substitution. As regards the second aspect, the Wikipedia edit story corpus contains an average lower distribution of simplification per sentence. Though related to these works, our contribution differs in that it adds a cross-linguistic level of comparison and also tries to provide an overview of possible factors affecting the distribution and the nature of simplification operations in ATS corpora.

3 Corpora and annotation scheme

Corpora. The corpora used in the analysis are the Italian corpus PaCCSS-IT and the English Parallel Wikipedia Corpus (ParWik). PaCCSS-IT is a parallel corpus composed of about 63,000 paired sentences, obtained crawling the web. The corpus is the result of a three-step approach strongly shaped by the level of simplification under investigation, i.e. syntactic simplification, consisting in: i) an unsupervised step in which a great amount of sentences with overlapping lexicon and different syntactic structure was clustered according to a similarity metric and automatically aligned [4]; ii) a supervised step aimed to train a classifier to predict the sentence alignment and iii) a readability assessment step aimed at assigning a readability score to the sentences in each pair. ParWik instead was obtained aligning two already existing text collections: the English Wikipedia and the Simple English Wikipedia. The authors aligned paragraphs whose TF*IDF cosine similarity was over a threshold of 0.5. The final corpus consists of 167,000 aligned sentence pairs.

To summarize, the two corpora differ in the fol-

[4] To be part of a cluster a sentence had to share all lemmas with PoS 'noun', 'verb', 'numeral', 'personal pronoun' and 'negative adverb'.

lowing aspects: i) language; ii) corpus collection approach iii) domain of texts; iv) level of simplification under investigation.

	PaCCSS-IT	ParWik
i)	Italian	English
ii)	Web crawling	Wiki-based alignment
iii)	Web corpus	Encyclopedic
iv)	Mainly syntax	Lexicon+Syntax

Table 1: Corpora design criteria.

Annotation of simplification operations. The comparison was conducted on 1,000 sentence pairs randomly extracted from the two corpora. To make possible the comparison, the sentences were annotated with the scheme in Table 2, previously conceived to annotate PaCCSS-IT.

Simplification operations
Deletion
Insertion
Verbal Features
Lexical Substitution
Reordering
Sentence Type
Residual

Table 2: Annotated simplification operations.

The manual annotation was carried out by one of the authors using the web-based annotation tool *Brat*[5]. As reported in the next section, the results of the manual annotation process provide an answer to the first question. The adopted schema originally designed to identify simplification operations within different typologies of parallel corpora in another language is able to cover almost all transformations in ParWik. The main limit is that the scheme does not take into account one of the more typical simplification operations, that is splitting long and complex sentences into one or more shorter ones (Narayan et al., 2017). This is because it was conceived to make explicit the transformations occurring in the PaCCSS-IT corpus, which only includes 1:1 pairs, i.e. for each 'complex' sentence only one 'simple' version exists. To annotate this operation in ParWik, we used the tag residual.

4 Corpora analysis

4.1 Distribution of simplification operations

Figure 1 reports the average distribution of simplification operations in the two corpora. As we can see, the first three most frequent operations in PaCCSS-IT are: 'deletion', 'verbal features' and 'insertion' and in ParWik 'deletion', 'lexical substitution' and 'insertion'. Excluding deletion, the differences resulted to be statistically significant for all operations, according to the Chi-squared test (p value <0.05).

At first glance, these results seem to suggest that language-specific factors affect the process of simplification. However, it is interesting to note that a qualitative analysis of these findings partially rules out this hypothesis, suggesting instead to interpret the differences also in view of the other criteria reported in Table 1. Specifically, the impact of language is limited to the different distribution of the 'verbal feature' operation. In PaCCSS-IT, it represents 29% of the total number of annotated operations while it is much less frequent in ParWik (<5 %). In particular, the distribution of this operation in the Italian corpus is mainly due the higher number of verbs at the conditional mood, which are transformed into indicative in the simplified sentence. As expected, verbs in ParWik are mostly at the indicative in both versions of the sentence. However, this different distribution has to be read also in view of another factor, i.e. the domain of texts in the corpora. Since it has been crawled from the web, PaCCSS-IT contains heterogeneous domains and many complex sentences belong to a 'written to be spoken' style, which implies the use of polite forms, expressed in Italian with the conditional mood. As a consequence of the different domain of texts contained in the two corpora, we can also observe a gap concerning the frequency of 'insertion'. Specifically, the encyclopedic nature of texts in ParWik may require the insertion of glosses and explanations to improve the understanding of complex terms. The lower frequency of lexical substitution operations in the Italian corpus (8.9% vs 23.9%) is easily explained if one considers the main purpose for which the corpus was designed, i.e. the investigation of syntactic simplification. On the contrary, editors of Simple Wikipedia are explicitly recommended "to write using Basic English words"[6].

4.2 Linguistic analysis

The diversity between the two corpora affects also the nature of the linguistic phenomena subjected

[5]https://brat.nlplab.org/

[6]https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

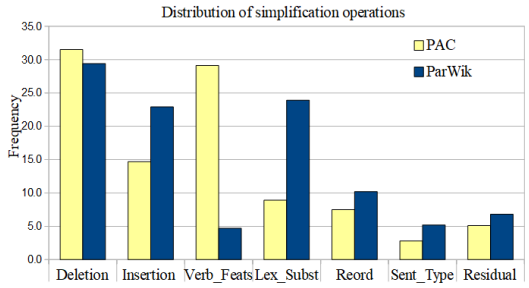

Figure 1: Distribution of simplification operations.

to simplification. This means that the type of linguistic elements which are, for example, deleted, inserted or substituted might be different. Again this variance is poorly attributable to the proprieties of the languages at play. In the following, we will try to outline a categorization of the linguistic elements subjected to modifications in the two corpora, providing an example for each case.

Deletion. This operation involves the deletion of single words or clauses. In particular, we observe a similar trend in the two corpora with the deletion of functional words, modal adverbs and adjectives alone or entire clauses containing these parts of speech.

- C: The main bar at King's is far older **and is the site of more informal meeting between students**. [ParWik]
- S: The main bar at King's is far older. [ParWik]
- C: **Probabilmente sospetto che** non sarebbe **comunque** una buona idea. (Probably, I suspect that it would not be however a good idea.) [PaCCSS-IT]
- S: Non fu una buona idea. (It was not a good idea.) [PaCCSS-IT]

Insertion. In both corpora auxiliaries and full verbs are inserted. Moreover in ParWik also nouns and pronouns are inserted as subjects of the new sentence, typically as a consequence of a split. As said, this does not occur in PaCCSS-IT, where however, implicit-explicit clause transformation implies the insertion of explicit elements, such as articles and verbs.

- C: Spese del presente grado di giudizio compensate tra le parti costituite. (Expense of the present level of justice compensated among the parts) [PaCCSS-IT]
- S: **Le** spese del presente grado di giudizio **possono essere** compensate tra le parti. (The expense of the present level of justice can be compensated among the parts). [PaCCSS-IT]

As said before, ParWik editors tend to insert explanations of complex terms and concepts. The contribute to simplicity of this type of insertion is quite clear in:

- C: According to the Armenian tradition, Saint Jude suffered martyrdom about 65 AD in Beirut, in the Roman province of Syria, together with the apostle Simon the Zealot, with whom he is usually connected.
- S: St. Jude was martyred, **killed for his beliefs**, with another apostle, Simon the Zealot in Beirut, Lebanon, around AD 65.

Instead, it is debatable in:

- C: Velvet Revolver is an American hard rock supergroup consisting of former Guns N' Roses members Slash, Duff McKagan, and Matt Sorum, alongside Dave Kushner formerly of punk band Wasted Youth.
- S: Velvet Revolver, **VR**, is a Grammy Award-winning rock supergroup. The members of the band are Slash **guitarist**, Duff McKagan **bassist, backing vocals**, Matt Sorum **drums of Guns N' Roses, Scott Weiland lead vocals of Stone Temple Pilots** and Dave Kushner **guitarist** of Wasted Youth.

Lexical substitution the more striking difference between the two corpora concerns this operation, not only in terms of frequency but also in respect of the type of substitution. In PaCCSS-IT, the operation affects only the substitution of words whose PoS was not considered in the clustering step, e.g. adjectives, adverbs and articles, etc. Moreover the substitution does not always contribute to the simplification of the sentence: this means that in some cases the complex term may be not replaced with a simpler synonym. In ParWik instead the operation affects phrase and sentence level, yielding to real paraphrases.

- C: Il concorrente è preventivamente stato avvertito **per** assistere all'operazione (The concurrent had been informed in advance to assist to the operation [PaCCSS-IT]
- S: Il concorrente è stato avvertito preventivamente, **affinché** possa assistere all'operazione. (The concurrent had been informed in advance in order to assist to the operation) [PaCCSS-IT]
- C: Sporting venues in the city include the Millennium Stadium the national stadium for the Wales national rugby union team and the Wales national football team, SWALEC Stadium the home of Glamorgan County Cricket Club, Cardiff City Stadium the home of Cardiff City football team and Cardiff Blues rugby union team, Cardiff International Sports Stadium the home of Cardiff Amateur Athletic Club and Cardiff Arms Park the home of Cardiff Rugby Club. [ParWik]
- S: Cardiff has one of the largest stadiums in the United Kingdom, the Millennium Stadium, where important world sports matches and concerts happen. Other big stadiums in the city are the Cardiff City Stadium, where the main football and rugby teams play, and the SWALEC Stadium where cricket is played. [ParWik]

Verbal features As said before, the Italian 'conditional→indicative' transformation does not occur in the English corpus, where instead the tag 'verbal features' was assigned to mark voice modification and 'indefinite→finite' mood transformations.

- C: Salve, **avrei bisogno** di una informazione piuttosto urgente. (Good morning, I would need a rather urgent information.) [PaCCSS-IT]
- S: Ho bisogno di una informazione urgente. (I need a urgent information.) [PaCCSS-IT]
- C: It is most often black but can come in a variety of colors including clear, **allowing** the top of the deck to be decorated. [ParWik]
- S: However, it can come in many different colors like clear. Clear allows the top of the deck to be decorated. [ParWik]

Reordering In general, in PaCCSS-IT, reordering implies the resetting of the canonical word order, while in ParWik there is a tendency to transform noun pre-modifiers in appositive phrases. As regards the position of subordinate clauses, neither of the two corpora assign to them a fixed position, i.e. before or after the main clause, although in ParWik embeddings are often extracted to form a new sentence.

- C: **Un'unica cosa** vorrei aggiungere. (Only a thing I would like to add.) [PaCCSS-IT]
- S: Volevo aggiungere solo una cosa. (I wanted to add only a thing.) [PaCCSS-IT]
- C: The United States presidential election of 1992 had three major candidates: Incumbent **Republican** President George H. W. Bush; **Democratic Arkansas Governor** Bill Clinton, and **independent** Texas businessman Ross Perot. [ParWik]
- S: The United States presidential election of 1992 was on November 3, 1992 in the United States. The three main people running were: George H. W. Bush, a Republican from Texas and the President; Bill Clinton, who was a Democrat and Governor of Arkansas; and Ross Perot an Independent candidate. [ParWik]

Sentence type. Three main phenomena fall under this tag: i) passive-active modification, ii) implicit-explicit clause modification and iii) verbalization-nominalization modification. While the first two modifications occur in both corpora, the third was found only in ParWik. Again, this difference is partly affected by language-dependent factors but it also depends on specific corpus-dependent constraints.

- C: Il presidente, **ricordato che nella seduta di ieri si è svolta la relazione,** dichiara aperta la discussione generale. (The president, reminded that the reporting was held in the yesterday part-session, declares open the general discussion.) [PaCCSS-IT]
- S: Il presidente ricorda che nella seduta di ieri è stata svolta la relazione introduttiva e dichiara quindi aperta la discussione generale. (The president reminds that in the yesterday part-session was held the introductory reporting and declares open the general discussion.) [PaCCSS-IT]
- C: **Findings of** coins indicate that the Romans were in Buxton throughout their occupation. [ParWik]
- S: Roman coins have been found in Buxton. [ParWik]

5 Conclusions and future works

The paper proposed a cross-linguistic comparison between two monolingual parallel corpora for ATS. The comparison tried to answer three main questions. As regards question 1, the annotation stage proved the possibility to use, except few modifications, a language-specific annotation scheme for another language. More than language-specific factors, an in-depth analysis of the annotated pairs of sentences highlighted that the observed differences are due to linguistic phenomena characterizing different textual genres. This is the case for example of modifications due to the insertion of glosses, which is driven by the encyclopedic nature of Wikipedia pages rather than to the specific language. Similarly, textual genre has an impact on the linguistic level involved in the lexical substitution. The higher occurrence of substitutions at phrase level, rather than at word-level, reflects the attempt of Wikipedia editors to make scientific contents clearer and simpler for a wide target population. Corpus-design differences, especially those occurring between manually and automatically derived corpora, may affect the distribution of the simplification operations also within the same genre. This is one of the possible directions that we want to explore in the near future.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Stefan Bott and Horacio Saggion. 2014. Text Simplication Resources for Spanish. Language Resources and Evaluation. *Language Resources and Evaluation*, 48(1): 93–120.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni. 2015. *Design and Annotation of the Frst Italian Corpus for Text Simplication*. Proceedings of the 9th Linguistic Annotation Workshop (LAW15), Denver, Colorado, USA.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta and Giulia Venturi. 2016. PaCCSS–IT: A Parallel Corpus of ComplexSimple Sentences for Automatic Text Simplication. *Methods in Natural Language Processing (EMNLP 2016)*, pages 1018.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

I. Gonzalez-Dios, M. J. Aranzabe, and A. Díaz de Ilarraza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, 52 (1) 217–47.

Shashi Narayan and Claire Gardent and Shay B. Cohen and Anastasia Shimorina. 2017. Split and Rephrase. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.*

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplication Research: New Data can Help. *Transactions of the Association for Computational Linguistics*, 3:283–29.

Sara Tonelli, Alessio Palmero Aprosio, Francesca Saltori. 2016. *SIMPITIKI: a Simplification corpus for Italian*. Proceedings of the Third Italian Conference on Computational Linguistics, Naples, Italy.

Mark Yatskar, Bo Pang, Cristian Danescu-NiculescuMizil, and Lillian Lee. 2010. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplications 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 Linguistic.