PaCCSS–IT: A Parallel Corpus of Complex–Simple Sentences for Automatic Text Simplification

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

Abstract

In this paper we present PaCCSS–IT, a Parallel Corpus of Complex–Simple Sentences for ITalian. To build the resource we develop a new method for automatically acquiring a corpus of complex–simple paired sentences able to intercept structural transformations and particularly suitable for text simplification. The method requires a wide amount of texts that can be easily extracted from the web making it suitable also for less–resourced languages. We test it on the Italian language making available the biggest Italian corpus for automatic text simplification.

1 Introduction

The availability of monolingual parallel corpora is a prerequisite for research on automatic text simplification (ATS), i.e. the task of reducing sentence complexity by preserving the original meaning. This has been recently shown for different languages, e.g. English (Zhu et al., 2010; Woodsend and Lapata, 2011; Wubben et al., 2012; Siddharthan and Angrosh, 2014), Spanish (Bott and Saggion, 2011; Bott and Saggion, 2014), French (Brouwers et al., 2014), Portuguese (Caseli et al., 2009), Danish (Klerke and Søgaard, 2012), Italian (Brunato et al., 2015). While English can rely on large datasets like the well-known Parallel Wikipedia Simplification corpus (Coster and Kauchak, 2011; Zhu et al., 2010) and, more recently, the Newsela corpus (Xu et al., 2015), for other languages similar resources are difficult to acquire and tend to be very small, thus preventing the application of data–driven techniques to automatically induce simplification operations. This is true for the language we are considering, i.e. Italian, where the only documented corpus for text simplification contains approximately 1,000 aligned original and manually simplified sentences (Brunato et al., 2015).

In this paper we present *PaCCSS–IT*, a Parallel Corpus of Complex–Simple Aligned Sentences for ITalian. To build the resource we developed a new approach for automatically acquiring a large corpus of paired sentences containing structural transformations which can be used as a developmental resource for text simplification systems. The proposed approach relies on monolingual sentence alignment techniques which have been exploited in different scenarios such as e.g. paraphrase detection (Ganitkevitch et al., 2013; Barzilay and Lee, 2003; Dolan et al., 2004) and evaluation (Chen and Dolan, 2011), question answering (Fader et al., 2013), textual entailment (Bosma and Callison-Burch, 2007), machine translation (Marton et al., 2009), short answer scoring (Koleva et al., 2014), domain adaptation of dependency parsing (Choe and McClosky, 2015). Specifically in ATS, these techniques are typically applied to already existing parallel corpora; in this case the task of aligning the original sentence to its corresponding simple version can be tackled by applying similarity metrics that consider the TF/IDF score of the words in the sentence (Barzilay and Elhadad, 2003; Nelken and Shieber, 2006; Coster and Kauchak, 2011) or methods taking into account also the order in which information is presented (Bott and Saggion, 2011).

Differently from these methods, our approach

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 351–361,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

contains two important novelties: the typology of the starting data and consequently the methodology developed to build the complex–simple aligned corpus. To overcome the scarcity of large parallel corpora of complex and simple texts in less–resourced languages like Italian, we started from a wide amount of texts that can be easily extracted from the web for all languages. This makes our method less expensive since it does not need a manually created corpus of aligned documents.

The proposed alignment method has been strongly shaped by the perspective from which we investigate text simplification, i.e. syntactic rather than lexical simplification. While lexical simplification aims at the substitution of complex words by simpler synonyms, syntactic simplification attempts to reduce complexity at grammatical level (Bott and Saggion, 2014). As shown by comparative analyses of monolingual parallel corpora in many languages, syntactic simplification concerns transformations affecting e.g. verbal features, the order of phrases or the deletion of redundant or unnecessary words (Brunato et al., 2015; Bott and Saggion, 2014; Coster and Kauchak, 2011; Caseli et al., 2009). Following this second perspective we define a method for bootstrapping and pairing sentences that intercepts simplification operations at morpho–syntactic and syntactic level typically used by human experts when simplify real texts.

Section 2 illustrates the approach to automatically acquire the corpus of complex–simple aligned sentences. In Section 3, the approach is tested and tuned on a development corpus. In Section 4, our approach is applied on a large corpus thus obtaining the final corpus of paired sentences, named PaCCSS–IT. In this last section, we also provide a global evaluation of the whole process and a qualitative analysis of the linguistic phenomena related to sentence complexity that we intercepted.

2 The Approach

Our approach for automatically acquiring the collection of paired sentences combines three steps. In a first step, we devised an unsupervised methodology *i)* to collect pools of sentences from a large corpus with overlapping lexicon but possible different structures; *ii)* to rank the resulting candidate sentences according to a similarity metric intended to bootstrap lexical–equivalent pairs undergoing structural transformations. In the second step, the top–list of the ranked pairs was manually revised and used to develop a classifier based on lexical, morpho–syntactic and syntactic features to detect the sentences correctly paired. In the third step, the individual sentences of each pair were ordered with respect to linguistic complexity computed by using an automatic readability assessment tool.

This approach has been tuned on PAISÀ[1] (Lyding et al., 2014) and tested on ItWaC (Baroni et al., 2009). The two analysed corpora were automatically POS tagged by the Part–Of–Speech tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi et al., 2009).

PAISÀ is a freely distributed corpus of texts with Creative Commons license automatically harvested from the web. This corpus includes approximately 388,000 documents for a total of 250 millions of tokens and it is a large existing corpus of authentic contemporary texts in Italian which is free of copyright restrictions. ItWaC is the largest existing corpus of authentic contemporary texts in Italian. It is a 2 billion word corpus constructed from the Web limiting the crawl to the .it domain and using medium-frequency words from *La Repubblica* journalistic corpus and Basic Italian Vocabulary lists as seeds.

2.1 Unsupervised Step

The first step is aimed at clustering all sentences contained in a large corpus. To be included in the same cluster, the sentences have to share all lemmas tagged with the Part–Of–Speech (POS) “noun”, “verb”, “numeral”, “personal pronoun” and “negative adverb”. Nouns and verbs were selected because they capture the informational content of a sentence. The other functional categories have also to be shared, otherwise the meaning of the sentence would be altered. For example, the deletion of the negative adverb *non* (not) in one of the two following sentences would convey the opposite meaning: *Non farei mai una cosa del genere!* (I would never do something like that) *Non potevo fare una cosa del genere.* (I could not do something like that). In

[1] http://www.corpusitaliano.it/

the overlapping process we did not take into account the linear order of the considered lemma POS. This was meant to capture lexically–equivalent sentences undergoing potential structural transformations (e.g. passivization, topicalization).

All sentences within the same cluster were paired and the pairs were ranked for similarity by calculating the cosine distance between the sentence vectors. Each vector is constituted by the frequencies in the cluster of all lemma of the sentence. The cosine similarity served to discard different and equal or quasi–equal sentences.

The whole unsupervised step was used to select the set of candidate pairs reducing the number of pairs on which the following supervised step had been applied.

2.2 Supervised Step

The supervised step is meant to classify whether candidate pairs were correctly or incorrectly aligned. To this end, we built a classifier based on Support Vector Machines with a quadratic kernel using LIBSVM (Chang and Lin, 2001) that was trained on a corpus of paired sentences correctly aligned. The classifier used different types of linguistic features, i.e. lexical, morpho–syntactic and syntactic, meant to mainly capture structural transformations occurring in the paired sentences.

These features were extracted both calculating their distribution in each sentence and considering their overlap between the two paired sentences. They can be classified into the following types:
cosine similarity feature: it refers to the cosine value calculated for each pair of sentences;
raw text feature: it refers to the sentence length calculated in terms of i) tokens of each of the two paired sentences and ii) the different number of tokens between the two sentences;
lexical features: they refer to i) the lemma unigrams contained in the two sentences excluding the PoS already considered in the pairing process (i.e. nouns, verbs, numerals, personal pronouns, negative adverbs); ii) the distribution of word unigrams overlapping between the two paired sentences considering all PoS.
morpho–syntactic feature: it refers to the distribution of up to 4–grams of coarse grained Parts–Of–Speech;
syntactic features: they refer to i) the distribution of up to 4–grams of dependency types calculated with respect to the hierarchical parse tree structure and the surface linear ordering of words; ii) the distribution of up to 4–grams of coarse grained Parts–Of–Speech of a dependent (*d*) involved in a dependency relation and the dependency relation type (*t*) with respect to the hierarchical parse tree structure.

2.3 Readability Assessment Step

In the third step, the individual sentences of each classified pair were ordered with respect to linguistic complexity computed by using an automatic readability assessment tool. In text simplification research it is widely accepted the use of readability assessment metrics for evaluating the transformations that reduce sentence complexity (Zhu et al., 2010; Woodsend and Lapata, 2011; Vajjala and Meurers, 2016). Since our approach is devoted to building resources for developing ATS systems, we relied on readability assessment techniques to rank the individual sentences of the pair. To this aim, we used READ–IT (Dell'Orletta et al., 2011), the only existing NLP–based readability assessment tool devised for Italian. It operates on syntactically parsed texts and assigns to each sentence a score quantifying its readability. The assigned readability score ranges between 0 (easy–to–read) and 1 (difficult–to–read) referring to the percentage probability for the documents or sentences to belong to the class of difficult–to–read documents. The two poles were defined on two typologies of texts belonging to the same textual genre (i.e. newswire texts) but intended for different users: adults with a rudimentary literacy level or with mild intellectual disabilities for the easy–to–read pole and readers of a national daily newspaper considered of medium difficulty for ∼70% of Italian laymen for the difficult–to–read pole.

3 Tuning Process and Evaluation

In order to tune and evaluate each step of the proposed approach, we tested it on PAISÀ. We first pruned from the corpus the sentences with a number of tokens <5 and >40. The resulting sentences were then grouped with respect to their shared POS (i.e. nouns, verbs, numerals, personal pronouns and negative adverbs) and paired using cosine similarity.

Cosine	nº *correct* pairs	% *correct* pairs
0.92	54	12.03
0.91	151	29.49
0.90	91	26.76
0.89	157	23.36
0.88	331	48.04
0.87	336	68.15
0.86	674	57.36
0.85	107	57.22
0.84	176	61.75
0.83	1096	40.35
0.71	1092	38.97
0.70	62	27.80
	Total: 4,327	

Table 1: Absolute number and % distribution of *correct* extracted pairs for each manually reviewed cosine threshold in PAISÁ.

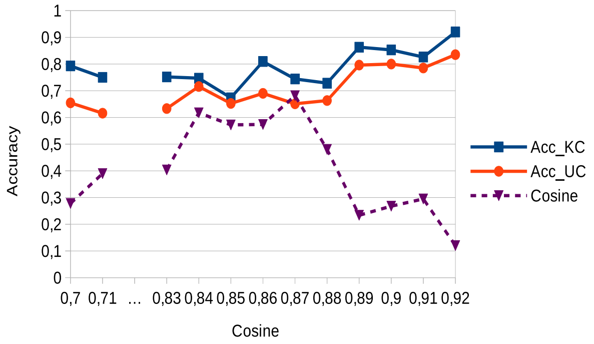

Figure 1: Accuracy of the *KC* and *UC* experiments compared with the distribution of correctly paired sentences at different cosine similarity scores.

We obtained 256,383 clusters containing at least two sentences. In order to discard different and equal or quasi–equal sentences we empirically set two cosine pruning thresholds: we discarded pairs with cosine below 0.4 since they were too lexically different and above 0.93 since they were too identical.

To build the training set for the supervised step we selected a subset of pairs resulting from the unsupervised step at different cosine similarity scores. This subset was manually reviewed by two native–speaker linguists with a background in text simplification. Specifically, they reviewed a subset of 10,543 pairs at different cosine similarity scores, i.e. those comprised between 0.92 and 0.83. In order to evaluate sentence similarity at lower values we also selected cosine scores 0.71 and 0.70. In the end, we obtained 4,327 correct pairs (i.e. about 41% of the whole set of candidate sentence pairs) distributed as in Table 1.

This manually revised set of pairs was then used to test the classifier in two different experimental scenarios. In the first one, named *Known Cosine, KC*, we tested the classifier in a five–fold cross validation process where pairs of sentences belonging to all the cosine scores were contained in each training and test set. In the second experiment, named *Unknown Cosine, UC*, the manually–revised corpus was differently split. In this case, the test set was composed by pairs of sentences with a cosine similarity score not contained in the training set and consequently twelve classification runs were performed.

In order to assess the discriminating power of the linguistic features used in the classification, we carried out an Information Gain analysis. This analysis showed the effectiveness of all the selected features in both experiments (i.e. KC, UC). In particular, we observed that the best ranked features are the morpho–syntactic and syntactic ones. This might suggest that our classification approach is intercepting pairs of sentences undergoing different typologies of structural transformations involving e.g. the use of verbal features or the order of phrases. Sentence length and lexical features play a lower discriminative role with respect to the grammatical features; this follows from the constraints we put on the unsupervised sentence pairing process. As it can be expected, the best ranked features are those providing information about the overlapping characteristics of the paired sentences.

Figure 1 and 2 report the results for each cosine threshold considered in the manual revision of the two experiments respectively in terms of *i)* Accuracy in the classification of the *correct* and *incorrect* alignments, and of *ii)* Precision, Recall of the classification of the *correct* alignments. As it can be noted in Figure 1, in both experiments the classifier is able to outperform the process of sentence pairing based only on the cosine (i.e. line *Cosine*, that represents the unsupervised step of the pairing process). As we can expect, Precision, Recall and Accuracy of the *KC* experiment are higher than the classification results obtained in the *UC*. The latter represents a more

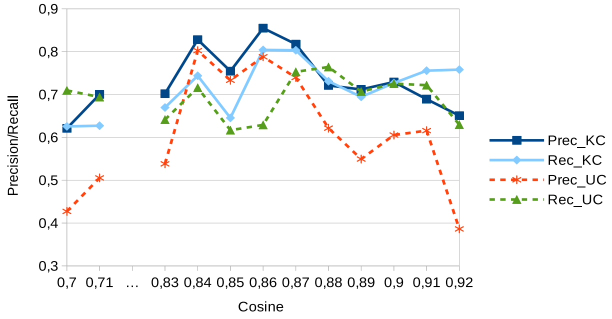

Figure 2: Precision and Recall of the two experiments (*KC* and *UC*) in the classification of the *correct alignments*.

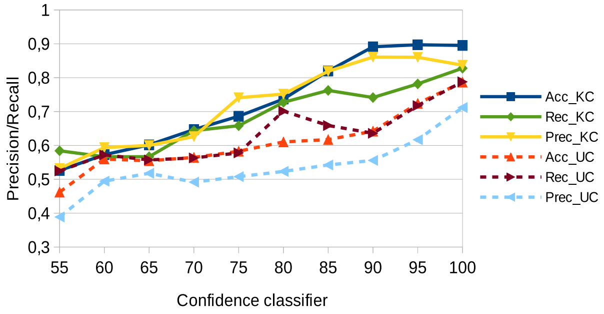

Figure 3: Classifier performance in the *KC* and *UC* experiments with probability intervals reported along the *x* axis.

challenging experimental scenario where the classifier is tested on a cosine threshold unseen in training. The overall results for the *KC* and the *UC* experiments are respectively 73.95% and 58.71% in terms of Precision, 70.3% and 68.1% in terms of Recall; and respectively 77.64% and 67.2% in terms of Accuracy. These results are significantly higher when compared with the accuracy of 41% reported for the unsupervised alignment (i.e. line *Cosine*). Interestingly, in the *KC* experiment, Precision and Recall lines are close and they remain stable with respect to all cosines even if the distribution of correct pairs varies in the different cosine values.

Figure 3 shows the accuracy of our classifier at different confidence thresholds (i.e. the probability assigned by the classifier for the *correct* alignments) for both the *KC* and *UC* experiments. Note that for each confidence intervals we have a different number of total pairs, and of gold-correct alignments and gold-incorrect alignments. As expected, the performance grows as the confidence grows. Interestingly, in the *KC* scenario, the classifier reached up to 90% of accuracy in discriminating the *correct* from the *incorrect* alignments when the classifier has a confidence score ≥0.90. These results look very promising if we consider that 30% of the pairs of the whole test set classified as correct alignments is comprised in the subset for which the classifier is more confident. This is also the case of the *UC* experiment, where, even if with lower accuracies, more than 56% of the *correct* alignments occurs when the classifier has a confidence score ≥0.90.

We carried out a last evaluation to estimate the classifier performance in the *UC* scenario for low cosine ranges not comprised in the manually revised portion of the corpus (from 0.45 to 0.75, excluding cosines 0.70 and 0.71). We considered only the pairs classified as *correct* with a confidence score of ≥85%. As expected, the system performance grows as the cosine grows: only few correct pairs occur at cosine <0.60, at cosine 0.60–0.65 the classifier assigns the *correct* class 237 times with an accuracy of 62.97%, at 0.65–0.69 330 times with 71.82% and at 0.72–0.75 256 times with an accuracy of 87.89%. According to these evaluations, we extracted from PAISÁ those pairs with a confidence score ≥ 85% and cosine similarity between 0.6 and 0.93, resulting in a collection of about 20,000 pairs.

In the last step, the sentences in each pair were ranked according to the readability score automatically assigned by READ–IT making a collection of complex–simple aligned sentences. However, the average difference of the readability score between the complex and simple sentences is only 0.13, making this collection not so useful for ATS. For this reason, we selected only pairs with a difference of readability score higher than a significant threshold set at 0.2. We defined this threshold on the basis of previous empirical experiments carried out using READ-IT on different typologies of texts. Lower variations of READ-IT score are scarcely perceived by human subjects. Since the construction of this resource has been specifically designed to develop ATS systems for human target, this READ-IT variation is a fundamental parameter of PaCCSS-IT. We thus obtained about 4,450 pairs.

4 PaCCSS–IT

The unsupervised step applied to ItWaC resulted in ∼28 million of clusters with overlapping lexicon for a total of ∼35 million of single sentences. From this initial set we pruned sentences with a number of tokens <5 and >40 and clusters containing less than two sentences. We obtained 419,252 clusters for a total of ∼8,5 million of single sentences and an average number of pairs in each cluster of 1,613. Filtering the pairs according to the cosine similarity range defined in the development step, we obtained a subset of 73,142 clusters with an average of ∼112 pairs for each. The classifier with the same model tested on PAISÁ recognised about 1 million of *correct* aligned pairs. Excluding pairs below the confidence score $\geq$ 85% we obtained ∼284,000 pairs. This collection was further pruned selecting only those pairs with at least 0.2 points in terms of variation in readability score. PaCCSS–IT is the resulting resource. It is a freely available resource [2] composed of ∼63,000 pairs of sentences (∼126k sentences) ranked with respect to the readability score of the two sentences. For each pair the cosine similarity, the probability score of the classifier and the readability level of the sentences are provided.

The following sections report the evaluation and the qualitative analysis we carried out on PaCCSS–IT. The evaluation was performed to assess the reliability of sentence alignment and of the sentence ranking with respect to readability level. The qualitative analysis was focused on studying which linguistic phenomena typically related to text simplification are successfully intercepted by our approach in order to show the applicability of the resource in a ATS scenario.

4.1 Evaluation

The evaluation process was intended to calculate the accuracy of i) the automatic classification process in predicting correct sentence alignments and ii) the automatic readability ranking of each pair.

The alignment evaluation was carried out by two trained linguists who manually revised 40 pairs of randomly selected sentences for each cosine score (1,088 paired sentences). It resulted that 85% of pairs were correctly classified (i.e. 921 pairs) and precision increases as cosine grows (from 73.2% at cosine 0.65-0.69 to 90.8% at cosine 0.90-0.92).

The subset of 921 pairs correctly classified was further investigated with respect to the readability level automatically assigned. To this aim we elicited human judgements through the crowdsourcing platform CrowdFlower[3]. We collected judgements from 7 workers that were asked to rate for each pair which of the two individual sentences was simpler. We considered the majority label to be true label for each pair. Comparing the score obtained by our system with the human judgements we obtained an accuracy of 74%. Restricting the evaluation only to pairs with the same label assigned by at least five out seven annotators (i.e. 79% of the whole pairs), the system achieved an accuracy of 78%.

4.2 Qualitative Analysis

Two qualitative analyses were carried out on PaCCSS–IT. The first analysis took into account the subset of 921 revised pairs with the aim of manually investigating what kinds of sentence transformations previously observed in the literature on text simplification were intercepted by our approach. In the second one, the whole resource was automatically investigated to study how the alignment process impacts on the distribution of multi–level linguistic features correlated to sentence complexity.

4.2.1 Analysis of Simplification Operations

Following the classification of simplification operations proposed in the literature (Brunato et al., 2015; Bott and Saggion, 2014; Coster and Kauchak, 2011; Caseli et al., 2009), we identified the major types of operations occurring in the subset of revised pairs [4], namely:

Deletion: the second sentence (S) does not contain one or more than two words occurring in the first one (C):

- C: *Ma c'è un altro problema, ancora più grave*. [Lit: But there is another problem, even more serious.]

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

[3] www.crowdflower.com

[4] In each of the following examples the first sentence (C) is the complex sentence and the second (S) the simple one. We underlined the text span affected by the operation.

- S: *Poi c'è un altro problema.* [Lit: Then there is another problem.]

Verbal Features: the two sentences differ with respect to verbal mood and tense:

- C: *I suoi libri sono stati tradotti in molte lingue.* [Lit: His books have been translated in many languages.]

- S: *I suoi libri sono tradotti in diverse lingue.* [Lit: His books have been translated in different languages.]

Lexical Substitution: the two sentences contain synonyms of words tagged with POS which were not considered in the clustering step based on POS overlapping, e.g. adjectives, adverbs:

- C: *Il colore è un rosso rubino fittissimo, quasi impenetrabile, limpido.* [Lit: The color is a rubyred very dense, almost impenetrable, clear.]

- S: *Il colore è un rosso rubino vivo quasi impenetrabile.* [Lit: The color is a bright red ruby almost impenetrable.]

Reordering: the two sentences contain a different word order both at single word (e.g. subject in pre- vs. post-verbal position) and phrase level (e.g a subordinate clause proceeds vs. follows the main clause):

- C: *Ringraziandola per la sua cortese attenzione, resto in attesa di risposta.* [Lit: Thanking you for your kind attention, I look forward to your answer.]

- S: *Resto in attesa di una risposta e ringrazio vivamente per l'attenzione.* [Lit: I look forward to your answer and I thank you greatly for your attention.]

Insertion: the second sentence contains one or *n*-words more than the first one:

- C: *In attesa di un sollecito riscontro, distinti saluti.* [Lit: Waiting for an early reply, yours faithfully.]

- S: *In attesa di un riscontro porgiamo distinti saluti.* [Lit: Waiting for a reply, we offer our regards.]

Sentence Type: the two sentences differ with respect to their form (i.e. affirmative vs. interrogative):

- C: *Quale consiglio darebbe ai genitori?* [Lit: Which advice would you give to parents?]

- S: *Diamo un consiglio ai genitori.* [Lit: Let's give an advice to parents.]

For each operation there can be different degrees of sentence transformation. For example, focusing on *Verbal Feature*, the example reported above represents a "light" transformation while a "stronger" transformation can occur when the verb changes from the conditional to the indicative mood (or vice versa), as in the following pair:

- C: *Sarebbe un grave un errore.* [Lit: It would be a serious error.]

- S: *Ma è un grave errore.* [Lit: But it is a serious error.]

Figure 4 reports the distribution of these sentence operations in the manually revised portion of the corpus. The distribution in *All cosines* shows that the two most frequent operations are deletion (30.74%) and changes affecting verbal features (26.30%). According to the literature, the deletion of redundant information (e.g. adjectives, adverbs) is one of the main phenomena typically related to reduction of complexity. Also transformations of verbal features are likely to intercept simplification operations in a language like Italian with a rich inflectional paradigm. The third most frequent operation is lexical substitution (15.52%). According to the POS filter used in the unsupervised step of sentence alignment, this operation affects morpho–syntactic categories such as e.g. adverbs, adjectives, conjunctions or prepositions which are substituted with a simpler synonym. Reordering and insertion of words or phrases are respectively the fourth and the fifth types of transformation. Reordering can be expected as a simplification strategy especially when it yields a more canonical word order. The distribution of reordering here reported, i.e. 14.24%,

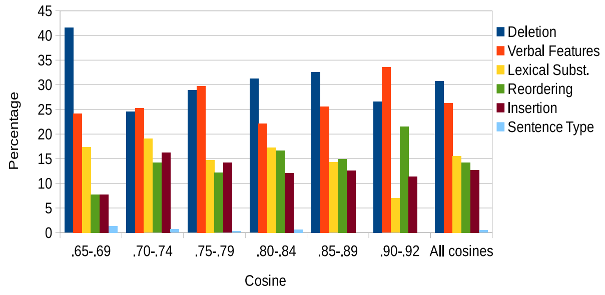

Figure 4: Distribution of sentence operations at different cosine ranges.

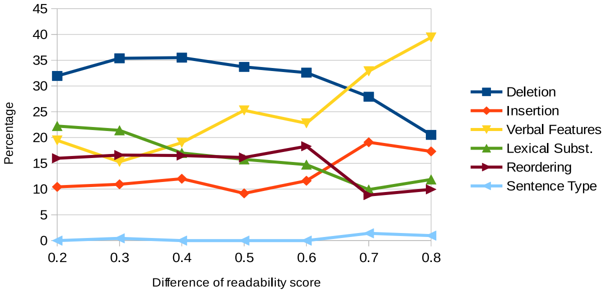

Figure 5: Distribution of sentence operations at different readability scores.

is quite high if compared to the distribution of the same operation found in hand-crafted simplified corpora where it represents about 8% of sentence operations (Brunato et al., 2015). This result gives evidence that our approach succeeds in automatically intercepting this kind of syntactic transformation. In the manually revised portion of PaCCSS–IT insertion represents 12.72% of the whole operations. Despite inserting words or phrases could make more complex a sentence, this operation is used in the simplification process e.g. when it makes explicit missing arguments in elliptical clauses more frequently used in non–standard language varieties or sublanguages such as legal language. This is the case of the heterogeneous nature of the corpus from which PaCCSS–IT derives, where documents characterized by non–canonical languages (e.g. blogs, e–mails) or domain–specific documents (e.g. administrative acts) are mixed to texts representative of more standard varieties, e.g. newspapers, novels.

Let us consider the relation between simplification operations, cosine values and readability levels. For what concerns the distribution of the operations at different cosines (Figure 4), we observe that deletion is the most frequent operation at all cosines, in particular at lower cosines i.e. <.70. At high cosines, i.e. >.90, operations affecting word order and verbal features increase. The relation between readability score and sentence operations is shown in Figure 5. Specifically, we calculated how the distribution of operations changes with increasing differences between the readability score assigned to the complex and the simple sentence of each pair. Although it is difficult to study the effect of each single operation on the readability score variation since these operations are usually applied in combination, we observe some clear tendencies. In particular, operations concerning deletion and verbal features are the most frequent ones both at lower and higher readability scores differences. However, they have an opposite distribution: transformations of verbal features increase at higher readability differences (>0.6) while deletions decrease. For what concerns the other operations, the trend is quite homogeneous along with the different readability scores. In particular, this is the case of reordering thus showing the proposed approach is able to intercept syntactic transformations which impact at different readability variations.

4.2.2 Analysis of Linguistic Phenomena

The second qualitative analysis focused on the whole resource which was searched for linguistic phenomena correlating with the process of sentence alignment. To this end, we compared the distribution of a set of different linguistic features, i.e. raw text, lexical, morpho–syntactic and syntactic, automatically extracted from the set of complex and simple sentences of PaCCSS–IT, which was previously tagged and dependency–parsed. In Table 2 we report a selection of the features with a statistically significant variation[5] between the complex and the simple sentences. As expected, the average sentence length (feature [1]) of the *Simple* sentence is lower than the *Complex* one. The higher distribution of adjectives [2], adverbs [3] and determiners [4] might be

[5] Wilcoxon's signed rank test was used to evaluate statistical significance.

related to the insertion of simple lexicon belonging to the Basic Italian Vocabulary (De Mauro, 2000). The distribution of verbal moods is also significantly correlated to a higher readability level: simple sentences have a higher percentage of indicatives [6] (a simple mood indicating a state of being or reality) and less participles [7] and gerundives [8] which are non finite moods and thus can be more ambiguous with respect to the reference. In addition, sentences classified as complex have higher parse trees [13], longer dependency links [14] and longer embedded complement chains modifying a noun [15], all features correlated with syntactic complexity (Gibson, 1998; Lin, 1986; Frazier, 1985). On the contrary, sentences classified as simple are characterised by a more canonical word order (Subject–Verb–Object in Italian) i.e. a lower distribution of post-verbal subjects [16] and of pre-verbal objects [17]. These sentences also contain a higher distribution of subordinate clauses following the main clause [18], an order easier to process.

Since syntactic features intercepting the structure of the sentence (e.g. parse tree depth and dependency length) heavily depend on the overall sentence length, we carried out an analysis only on pairs of sentences where the complex and the simple sentence have the same number of tokens (i.e. 15,958 pairs in PaCCSS–IT) and we compared how linguistic features vary between the complex and the simple sentences of these pairs. We observed that simple sentences have a more canonical position of the subject (i.e. a lower percentage distribution of post-verbal subjects: *C*: 18.14%, *S*: 15.72%) and of the object (i.e. a lower percentage distribution of pre-verbal objects: *C*: 1.52%, *S*: 1.18%). Simple sentences have also lower parsed trees (*C*: 2.42, *S*: 2.37) and shorter embedded complement chains modifying a noun (*C*: 0.27, *S*: 0.26). Since these variations cannot be due to sentence shortening, they rather follow from reordering phenomena e.g. changing from active to passive voice.

The distribution of linguistic features here reported has already been observed in hand–crafted corpora of complex and simple sentences for Italian (Brunato et al., 2015). This is a further evidence of the reliability of our method for automatically creating corpora of complex–simple sentences.

Feature	*Complex*	*Simple*	*Variation*
[1]	8.98	7.80	0.97
[2]	4.10	7.90	-3.80
[3]	9.10	10.0	-0.85
[4]	0.34	1.43	-1.10
[5]	10.70	20.30	-9.61
[6]	5.20	2.72	2.49
[7]	0.47	0.04	0.42
[8]	2.89	4.29	-1.40
[9]	79.18	80.91	-1.73
[10]	1.33	1.57	-0.24
[11]	2.03	2.14	-0.10
[12]	8.35	7.33	0.5
[13]	2.88	2.70	0.18
[14]	1.76	1.63	0.12
[15]	0.44	0.41	0.02
[16]	15.37	14.37	1.00
[17]	2.03	1.38	0.65
[18]	3.29	4.17	-0.90

Table 2: Distribution of a subset of linguistic features with statistically significant variation between the complex and simple sentences. Features [1],[13],[14],[15] are absolute values, the others are percentage distributions. All differences are significant at p <0.001.

5 Conclusion

In this paper we have presented PaCCSS–IT, a corpus of complex–simple aligned sentences for Italian containing ∼63,000 paired sentences. To our knowledge, PaCCSS–IT is the biggest corpus of complex–simple aligned sentences, with the exception of English. It resulted from a new method for automatically acquiring corpora of parallel sentences able to capture structural transformations and particularly suitable for text simplification systems. A comparative analysis of the multi–level linguistic features in the complex and simple sentences showed that this method intercepts linguistic phenomena characterising simplification operations previously observed in manually–created complex–simple corpora. A main novelty of the proposed approach is that it does not rely on a large pre-existing corpus of aligned complex–simple documents like e.g. the English and Simple English Wikipedia. This makes it very appropriate for less–resourced languages. In addition, since the method does not need parallel corpora, the dimension of the web is the only limitation to the size of the corpus that could be created.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of the 2nd Workshop of Evalita 2009 - Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, Italy.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Regina Barzilay and Noemi Elhadad. 2003. Sentence alignment for monolingual comparable corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wauter Bosma and Chris Callison-Burch. 2007. Paraphrase substitution for recognizing textual entailment. *Proceedings of the 7th International Conference on Cross-Language Evaluation Forum (CLEF)*.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. *Proceedings of the Workshop on Monolingual Text-To-Text Generation, co-located with ACL 2011*, Porland, Oregon.

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

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. *Proceedings of the 9th Linguistic Annotation Workshop (LAW'15)*, Denver, Colorado, USA.

Helena M. Caseli, Tiago F. P. Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra 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*.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm*.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. *Proceedings of the Annual Meetings of the Association for Computational Linguistics (ACL)*.

Do Kook Choe and David McClosky. 2015. Parsing paraphrases with joint inference. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

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*.

Tullio De Mauro. 2000. *Il dizionario della lingua italiana*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: assessing readability of italian texts with a view to text simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Edinburgh, UK.

Felice 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.

William Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. *Proceedings of the Annual Meetings of the Association for Computational Linguistics (ACL)*.

Lyn Frazier. 1985. Syntactic complexity. *Natural Language Parsing*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish parallel corpus for text simplification. *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Nikolina Koleva, Andrea Horbach, Alexis Palmer, Simon Ostermann, and Manfred Pinkal. 2014. Paraphrase detection for short answer scoring. *Proceedings of the third workshop on NLP for computer-assisted language learning*.

Dekan Lin. 1986. On the structural complexity of natural language sentences. *Proceedings of COLING 1996*.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Dittmann Henrik, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISA corpus of Italian web texts. *Proceedings of the 9th Web as Corpus Workshop (WAC-9) EACL*.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 3–7 April.

Advaith Siddharthan and Mandya 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)*.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv*.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. *Proceedings of the 50th Annual Meeting of the 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.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. *Proceedings of the 23rd international conference on computational linguistics*.