It Is MarkIT That Is New: An Italian Treebank of Marked Constructions

Teresa Paccosi, Alessio Palmero Aprosio, Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive 18, Trento (Italy)
[tpaccosi|aprosio|satonelli]@fbk.eu

Abstract

English. In this paper we present MarkIT, a treebank of marked constructions in Italian, containing around 800 sentences with dependency annotation. We detail the process to extract the sentences and manually correct them. The resource covers seven types of marked constructions plus some ambiguous sentences, whose syntax can be wrongly classified as marked. We also present a preliminary evaluation of parsing performance, comparing a model trained on existing Italian treebanks with the model obtained by adding MarkIT to the training set.

Italiano. *In questo lavoro presentiamo MarkIT, un treebank di costruzioni marcate in italiano che contiene circa 800 frasi annotate con strutture a dipendenze. Abbiamo descritto nel dettaglio il processo seguito per estrarre le frasi e correggerne manualmente la struttura sintassi. La risorsa comprende sette tipologie di costruzioni marcate oltre ad alcune costruzioni ambigue che potrebbero essere classificate erroneamente come marcate. Presentiamo inoltre una valutazione preliminare delle performance del parser in cui confrontiamo un modello allenato sui treebank esistenti dell'italiano con il modello ottenuto aggiungendo anche MarkIT.*

1 Introduction

In recent years, the goal to develop robust frameworks for consistent annotation of syntactic dependencies across different human languages has led to the creation of Universal Dependencies (UD), an initiative covering nearly 200 treebanks in more than 100 languages. Since UD treebanks are then used to train syntactic parsers, it is important that they account for as many phenomena as possible that can be found in a language, and not only for canonical expressions typically written in news. The purpose to encompass the variety of use in the Italian language has been pursued by including different genres in the VIT treebank (Delmonte et al., 2007) and in ParTUT (Sanguinetti and Bosco, 2014) and more recently by including syntactically annotated tweets (Cignarella et al., 2019; Sanguinetti et al., 2018) in the UD framework. Overall, seven treebanks are listed under the UD initiative for Italian. In this work, we contribute to this effort by presenting a novel treebank including syntactically annotated marked constructions, which we call MarkIT (MARKed structures Italian Treebank). The samples have been extracted from a corpus of students' essays and to our knowledge represents the first effort to include in UD a repository of marked structures, which are typical of neo-standard language and are therefore more and more frequent in informal settings (D'Achille, 2003). The sentences have been first syntactically parsed and then manually corrected, so that we were also able to analyse which kinds of mistakes are typically done by dependency parsers. The dataset is freely available on Github at https://github.com/dhfbk/markit.

2 Related Work

In the last years, Universal Depedencies (UD) have become the most widely used standard for syntactic annotation (de Marneffe and Manning, 2008) upon which treebanks for other languages have been built, including Italian. The first one has been the Italian Stanford Dependency Treebank or ISDT (Bosco et al., 2013). Other treebanks have been later built with different purposes, covering a rich collection of different usages and genres. In particular, the VIT treebank (Delmonte et al., 2007) is composed of several texts ranging from news to literature, while TWITTIRO (Cignarella et al., 2019) and PoSTWITA (Sanguinetti et al., 2018) are two social-media-based treebanks, composed of tweets. These two Twitter-based treebanks represent an important resource in terms of documentation of the usage of non-standard Italian. We address the same topic in the present work, but instead of considering social-media data, we look at more formal writings, and in particular at the use of marked sentence constructions in students' essays. To our knowledge, a grammatical UD treebank for Italian language does not exist, and also in other languages there are only few examples. A grammatical treebank is a dataset of annotated trees sharing the same type of grammatical constructions, such as the English Pronouns treebank (Munro, 2021), which is the most similar resource to ours. It was created to make independent genitive pronoun's identification more accurate, by annotating only English sentences which display that construction. For what concerns marked structures in Italian, a comparative study on the distribution of the phenomenon of syntactic markedness has been presented in (Pieri et al., 2016), but the different structures were identified using automated tools. Overall, syntactic markedness is a phenomenon poorly analyzed, especially in the field of dependency grammar. However, it is crucial to make parsers more robust to different syntactic structures.

3 Sentence Collection

Our goal is to build a treebank of marked constructions that reflects actual usage of Italian, in particular of the neo-standard variant (Berruto, 2012). We avoid to manually create sentences ourselves, also to increase linguistic variability. Therefore, we resort to a corpus of students' essays which were collected by Istituto provinciale Trentino per la Ricerca e la Sperimentazione educativa (IPRASE) with the goal to study the evolution of high-school students' writing skills, taking into account essays spanning 15 years (from 2001 to 2016). In particular, the project tracked the presence of expressions and constructions typical of neo-standard Italian, requiring a pool of expert annotators, i.e. high-school teachers, to manually mark in essays a number of linguistic traits (Sprugnoli et al., 2018; Tonelli et al., 2020). Among others, annotators were asked to mark dislocated sentences, cleft sentences and hanging topics (see details in Section 4). These were first automatically identified through the TINT NLP Suite (Aprosio and Moretti, 2018) and then manually revised by annotators to distinguish between the constructions of interest and other types of similar constructions.

The final corpus contains more than 2,500 essays and almost 1.5 million tokens. We extract around 800 sentences labeled with a marked structure and annotate them at syntactic level. Although the essays cannot be released because of copyright issues, the sentences in isolation, with no additional information related to the authors or the textual context, can be freely distributed.

The essays were written in a time span of 15 years by different authors and dealing with a number of different topics, which guarantees a high variability of the sentence content and structure. On the other hand, since they were part of a formal students' examination, they tend to be free from jargon, grammatical errors and abbreviations that may derive from sentences extracted from social media and that may represent an additional challenge for parsers.

4 Marked Structures in IPRASE Corpus

With marked sentences we refer to those constructions which present a non-canonical order of constituents. In Italian, the canonical order of the syntactic structure is *S V+fin V-fin OX*, where *S* is subject, *V+fin* is a finite verb or an auxiliary verb, *V-fin* is a non-finite verb, *O* is the direct object and *X* other complements (Benincà et al., 1988). Marked structures are instead intended to focus on an element of the sentence, by moving the focalized constituent in a different position from the one it occupies in a canonical sentence. The reason for markedness in Italian can be phonotactic or bound to the whole meaning of the sentence. In syntactical terms, we can say that the marked structures operate a modification in the distribution of *topic* and *comment* with respect to the corresponding non marked structure (Cinque, 1990). There are seven possible marked structures in Italian: sentences with postverbal subject, sentences with presentative "there", sentences with left or right dislocation, hanging topic sentences, cleft sentences

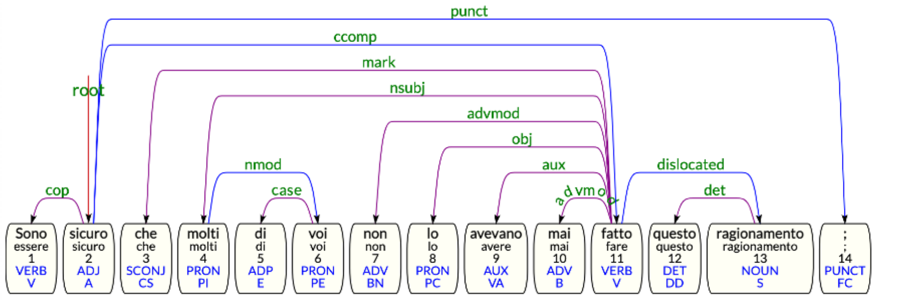

Figure 1: Right dislocated sentence annotated with dislocated relation

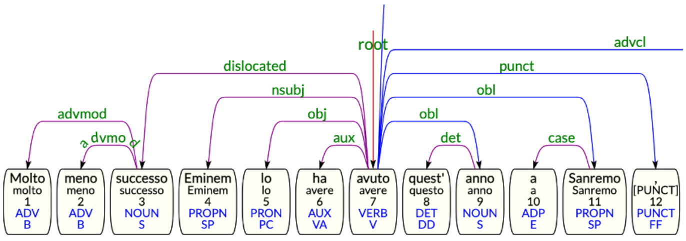

Figure 2: Left dislocated sentence annotated with dislocated relation

and pseudo-cleft sentences (Ferrari and Zampese, 2016). Among the sentences from the IPRASE corpus originally marked as dislocated, cleft and hanging topic, we were able to find other types of marked structures which had been wrongly identified by the annotators, so that in the end all seven phenomena are present. Below we report a brief description of the main marked structures annotated in our treebank.

4.1 Left Dislocated Sentences

Left dislocated sentences entail the displacement or anteposition of a specific syntagm to the left of the sentence. The dislocated element connects with the rest of the sentence thanks to an introductory preposition (1) or a pronominal reprise (2), for which a resumptive clitic pronoun pleonastically co-refers to the displaced nominal element (the topic). The clitic reprise is compulsory whether the displaced element was the direct object, as long as it is in the positive form (Benincà et al., 1988).

(1) A questo evento (ci) partecipano soltanto artisti già noti
To this event (clitic) participate only artists already known

(2) Molto meno successo Eminem lo ha avuto quest'anno
Much less success Eminem it has had this year

4.2 Hanging Topic Sentences

In hanging topic sentences, similarly to left dislocation, the dislocated element is moved to the left, at the beginning of the sentence. However, in this case, the displaced element is isolated at the beginning of the sentence, and it is not syntactically linked to the verb (D'Achille, 2003). The main difference between the two structures is when the dislocated element is the direct object. In fact, since direct objects in Italian exclude prepositional government, only the non-clitic reprise allows the distinction between left dislocated sentences and hanging topics. In hanging topic constructions, the isolated element is always deprived of indicators for its syntactic function, and it is typically reprised in the following phrase by different anaphorical expressions such as atonic pronouns, possessive pronouns, adverbs, and by a whole nominal phrase (3). When there is no reprise of the dislocated element in the subsequent sentence, we refer to that as an example of anacoluthon (Ferrari and Zampese, 2016).

(3) [...] ma il cervello, senza di esso non siamo niente
But the brain, without it we are nothing

4.3 Right Dislocated Sentences

Right dislocated sentences operate a topicalization of the comment and, differently from left dislocated structures, the pronominal reprise is not

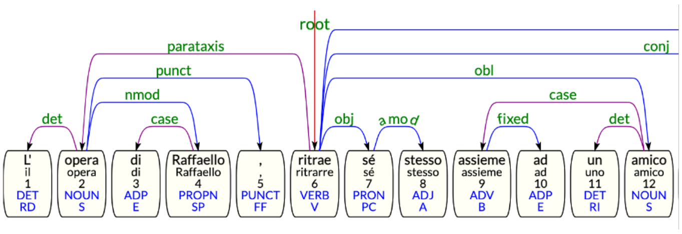

Figure 3: Hanging topic annotated with parataxis relation

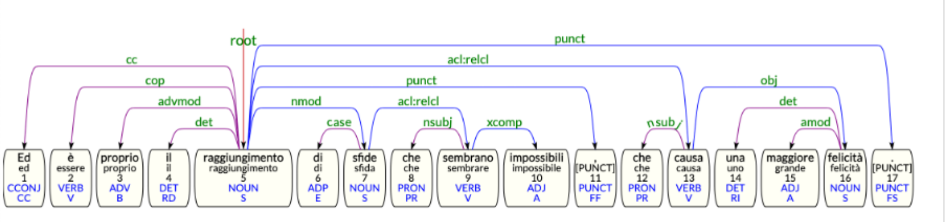

Figure 4: Cleft sentence with relative clause (*acl:relcl*)

compulsory when the dislocated element is the direct object. Nevertheless, since the non-marked position of the right dislocated elements is still in postverbal position (apart from the subject), it makes the presence of the anticipatory clitic (4) or of the comma (5) compulsory.

(4) Sono sicuro che molti di voi non lo avevano mai fatto, questo ragionamento
I am sure that many of you do not it have never done, this reasoning

(5) Interessante, in questo senso, la riflessione di Paul
Interesting, in this sense, Paul's thought

4.4 Cleft Sentences

Cleft sentences are typically composed of a main clause without a subject introduced by the verb 'to be' in different forms, followed by the cleft constituent and by a subordinate clause introduced by "che" (*that*), whose function can be of relative pronoun (6) or relative conjunction. Sometimes, the subordinate clause can be introduced by "a" (*to*) + a verb in the infinite form (7), if the subject is the element to put into focus (Berruto and Cerruti, 2011). Besides the subject, cleft structures can focalize on several constituents, such as the object, prepositional constituents, adverbs and also verbs, especially in the infinitive form (Renzi, 2001).

(6) È lo Stato che [...] impone i suoi modelli
It is the State that [. . .] imposes its models

(7) Non è dunque l'ottica dell'utilità e del guadagno a guidare verso la felicità
It is not then the view of utility and profit to guide to happiness

5 MarkIT Annotation

Marked structures, such as the ones described above, are very difficult to parse, since they belong to non-standard Italian constructions. In order to annotate them syntactically, we therefore need to follow a semi-automatic approach, by analysing them first with a dependency parser and then manually correcting them. The selected marked constructions from the IPRASE corpus were processed with the TINT parsing module (Aprosio and Moretti, 2018), which is built following Universal Dependencies guidelines (de Marneffe and Manning, 2008), and trained on the Italian Stanford Dependency Treebank, ISDT (Bosco et al., 2013). ISDT includes mostly standard language with few non-canonical constructions. The dependency trees parsed by TINT are then manually corrected by an expert linguist using the TINTful interface (Frasnelli et al., 2021). They are also marked with one of the categories from Table 1.

Concerning *dislocated sentences*, the main issue with TINT is that it assigns to the pronoun the role of direct object and treats the dislocated element as the subject, as in the example shown in Figure 6. The sentence was manually corrected by marking the dislocated element with the *dislocated* relation and the pronoun of reprise with the core argument relation which it represents (*obj* or *subj*), as we can see in Fig. 1 and Fig. 2.

As previously mentioned, *hanging topics* differ from left dislocated sentences because the element to the left is not syntactically linked to the

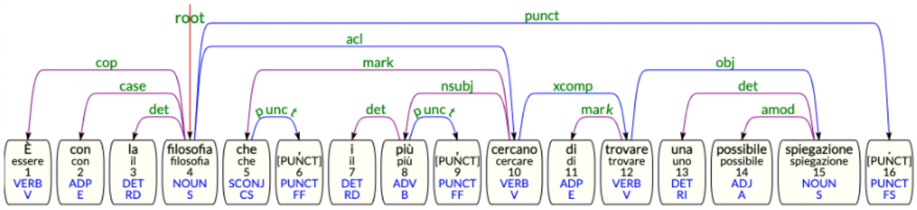

Figure 5: Cleft sentence with adnominal clause (*acl*)

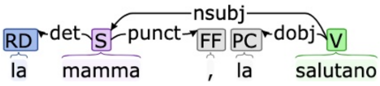

Figure 6: Wrong parsing output of dislocation

verb and there is no clitic reprise of the lexical element. Since there is a sort of isolation of the topicalized element, we choose to use the *parataxis* relation to link it to the head of the sentence, since *parataxis* is defined as a relation between a word and other elements, without any explicit coordination, subordination, or argument relation with the head word, which is usually the verb. An example is reported in Figure 3.

As we have seen above, the “che” (*that*) before a subordinate clause can be a relative pronoun introducing a relative clause or a relative conjunction, followed by a structure whose nature is controversial. A relative clause is an instance of clausal modifier *acl*, which takes the specific name of *acl:relcl*, where the noun can be omitted or substituted by a relative pronoun, relative conjunction, or an adverb. As regards *cleft sentences*, we choose to use the same relation in two different ways, in order to distinguish between the case in which the cleft sentence comprehends a relative clause or an unspecified subordinate clause. When the dislocated element is the subject or the direct object (substituted by a relative pronoun) we use the *acl:relcl* relation, selecting the role of “che” (see Fig. 4). Instead, if there is no dislocation of the subject or the object, we use the *acl* relation but we do not select the function of “che”. Indeed, “che” is treated as a mere introducer for the subordinate clause with the *mark* relation (Fig. 5).

Table 1 shows the eight types of constructions in MarkIT. Beside the four types of marked structures described above, we found several other structures coming from the erroneous identification of cleft sentences and right dislocated sentences in the original IPRASE corpus. “Presentative there” (i.e. “there” + verb to be + nominal element + that)[1] and “pseudo-cleft” structures (“what” clause + verb to be) [2] were wrongly identified as cleft sentences, while the structures with a postverbal subject were originally labeled as right dislocated. Furthermore, we include in the “Other” category the structures which resulted challenging to tag for the annotators and which are not “presentative there” nor pseudo-cleft constructions. “Other” structures are namely those which usually present an explicit subject in the main clause and are erroneously identified as cleft, for example *La capacità di concepire un insieme di diritti è una facoltà che distingue l’uomo dagli altri esseri viventi* (EN: The ability to conceive a set of rights is a faculty that distinguishes humans from other living beings). “Other” structures include also passive clauses, which were originally tagged as right dislocated because of the postverbal position of the subject. Sentences in this last category are particularly challenging both for parsers and for human annotators, since they were wrongly classified even by IPRASE experts (i.e. high-school teachers) and have been assigned the correct label only after our revision.

6 Parsing Evaluation

As already mentioned in the Introduction, the lack of marked structures in treebanks used to train syntactic parsers may affect the system robustness, since structures which are not represented in the training data tend to be poorly analysed. In order to measure the impact of our novel treebank on the dependency analysis of marked structures, we compare the performance of the parser included in TINT, part of Stanford CoreNLP (Manning et al., 2014) by testing it on the new annotated sentences and training on different datasets. In particular,

[1] e.g. *C’è Michela che ti cerca* (EN: There is Michela that is looking for you)

[2] e.g. *Ciò che voglio davvero è che tu te ne vada* (EN: What I really want is that you go)

Type	Sents
Cleft sentences	309
Left dislocated	121
Right dislocated	49
Presentative “there”	25
Postverbal subject	16
Pseudo-clefts	11
Hanging topic	7
Other	275
Total	813
Total (tokens)	24,623

Table 1: Number of examples in the dataset.

we first split our novel treebank into training, dev, and test, respectively 80%, 10%, and 10%, proportionally with respect to the categories listed in Table 1. When the number of examples is tiny, we include a minimum of two examples for each class in each split, therefore test and dev set contain two examples of *hanging topic* each, leaving the three sentences for the training set.

We then compare two models: the original neural transition-based parser model used by TINT, which is trained using ISDT, VIT, and ParTUT (see Section 2), and the model obtained by adding to the above training data also the training set of MarkIT. We choose not to include the other Italian datasets available from Universal Dependencies (such as the ones derived from Twitter) because of their particularly informal language, which is very different from MarkIT sentences taken from students' essays. In both cases, we use the concatenation of the development sets of the four datasets as development set during the training phase. Following the standard evaluation used in dependency parsing, we compute unlabeled attachment score (UAS) and labeled attachment score (LAS) in the two tests.

Training set	UAS	LAS
ISDT+VIT+ParTut	82.53	76.62
ISDT+VIT+ParTut+MarkIT	82.74	77.41

Table 2: Evaluation of the dependency parsing.

Results in Table 2 show that on the one hand adding MarkIT to the training set improves the classification of marked structures, but on the other hand performance gain is limited. This may be due to the fact that, compared to the other treebanks (more than 23k sentences in total), the number of training instances coming from MarkIT is small (around 650 sentences). More generally, the presence of both marked and not marked sentences (the “Other” category) in the test set represents a challenge for parsers, since very similar constructions are labeled differently, see for example the presence of comma to mark right dislocated elements. Indeed, if the first model is tested only on sentences taken from ISDT+VIT, it achieves 84.47 UAS and 80.69 LAS.

7 Release

MarkIT is released under CC BY 4.0 license,[3] and can be downloaded from Github.[4] The annotation of the treebank will be soon completed with all marked sentences in the essays dataset (see Section 8) and proposed for publication on the Universal Dependencies website.[5] Since the treebank is still being extended with new sentences, it may be that the content of the last version available online exceeds the size of the resource described in this paper.

8 Conclusions

In this work we present MarkIT, a novel treebank composed of 800 sentences with syntactic annotation of marked structures. The resource covers seven types of marked sentences, plus around 200 sentences whose structure is not marked but that may be misleading both for parsers and for human annotators. The treebank is made available to the community and is meant to make dependency parsers more robust to the different syntactic structures present in Italian, in particular in the neo-standard variant. The work is still in progress, since we plan to add to the resource other sentences from the IPRASE corpus. Our goal is to include all marked sentences present in the essays, so to analyse also the distribution of the different sentence structures in this type of texts.

References

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an all-inclusive suite for nlp in italian. In

[3] https://bit.ly/cc-by-40-intl

[4] https://github.com/dhfbk/markit

[5] https://universaldependencies.org/

Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it, volume 10, page 12.

P. Benincà, Salvi G., and Frison L. 1988. L'ordine degli elementi della frase e le costruzioni marcate. In L. Renzi, editor, *Grande grammatica italiana di consultazione. I. La frase. I sintagmi nominale e preposizionale*, pages 115–225. Il Mulino.

G. Berruto and M. Cerruti. 2011. *La linguistica: un corso introduttivo*. UTET Università.

Gateano Berruto. 2012. *Sociolinguistica dell'italiano contemporaneo*. Carocci.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. Association for Computational Linguistics.

Guglielmo Cinque. 1990. *Types of A' Dependencies*. MIT Press.

Paolo D'Achille. 2003. *L'italiano contemporaneo*. Il mulino Bologna.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August. Coling 2008 Organizing Committee.

R Delmonte, A Bristot, and Sara Tonelli. 2007. Vitvenice italian treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technol.

A. Ferrari and L. Zampese. 2016. *Grammatica: parole, frasi, testi dell'italiano*. Carocci editore.

Valentino Frasnelli, Lorenzo Bocchi, and Alessio Palmero Aprosio. 2021. Erase and rewind: Manual correction of NLP output through a web interface. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 107–113, Online, August. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Robert M. Munro. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Giulia Pieri, Dominique Brunato, and Felice Dell'Orletta. 2016. Studio sull'ordine dei costituenti nel confronto tra generi e complessità (analysis of constituents order across textual genres and complexity). In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

L. Renzi. 2001. Le tendenze dell'italiano contemporaneo. note sul cambiamento linguistico nel breve periodo. *Studi di lessicografia italiana*, pages 279–319.

Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank partut in universal stanford dependencies. *Converting the parallel treebank ParTUT in Universal Stanford Dependencies*, pages 316–321.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Rachele Sprugnoli, Tonelli Sara, Alessio Palmero Aprosio, and Moretti Giovanni. 2018. Analysing the evolution of students' writing skills and the impact of neo-standard italian with the help of computational linguistics. In *CLiC-it 2018 Italian Conference on Computational Linguistics*, pages 354–359. aAccademia University Press.

Sara Tonelli, Rachele Sprugnoli, Alessio Palmero Aprosio, Moretti Giovanni, and Menini Stefano. 2020. Gli strumenti informatici. sviluppo e risultati. In Michele Ruele and Elvira Zuin, editors, *Come cambia la scrittura a scuola: Rapporto di ricerca*, chapter 4, pages 113–130. IPRASE.