IJCoL
Italian Journal of Computational Linguistics

9-1 | 2023
Italian Journal of Computational Linguistics vol. 9, n.1 june 2023

Adding a Novel Italian Treebank of Marked Constructions to Universal Dependencies

Teresa Paccosi, Alessio Palmero Aprosio and Sara Tonelli

Electronic version
URL: https://journals.openedition.org/ijcol/1110
ISSN: 2499-4553

Publisher
Accademia University Press

Electronic reference
Teresa Paccosi, Alessio Palmero Aprosio and Sara Tonelli, "Adding a Novel Italian Treebank of Marked Constructions to Universal Dependencies", *IJCoL* [Online], 9-1 | 2023, Online since 01 August 2023, connection on 17 September 2023. URL: http://journals.openedition.org/ijcol/1110

Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International - CC BY-NC-ND 4.0
https://creativecommons.org/licenses/by-nc-nd/4.0/

Adding a Novel Italian Treebank of Marked Constructions to Universal Dependencies

Teresa Paccosi*
Fondazione Bruno Kessler, Trento (Italy)
University of Trento, Trento (Italy)

Alessio Palmero Aprosio**
Fondazione Bruno Kessler, Trento (Italy)

Sara Tonelli†
Fondazione Bruno Kessler, Trento (Italy)

In this paper we present a novel treebank developed to analyse marked constructions in Italian called MarkIT. The resource contains almost 1,300 sentences manually annotated with dependency relations following the Universal Dependencies paradigm. The sentences have been extracted from essays written by high-school students along several years, which accounts for the structure and the topic variability of the sentences. In this work, we detail the process to select the sentences, parse them automatically and then manually correct them. The resource covers seven types of marked constructions (839 sentences overall) plus some sentences, whose syntax can be wrongly classified as marked and which can serve as negative examples of markedness (453 sentences). We also present an evaluation of parsing performance, comparing a model trained on existing Italian treebanks with the model obtained by adding MarkIT to the training set.

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.

As regards Italian, different genres have been included in the VIT treebank (Delmonte, Bristot, and Tonelli 2007) and in ParTUT (Sanguinetti and Bosco 2014), with the purpose to encompass the variety of language use. More recently, also syntactically annotated tweets have been included in the UD framework (Cignarella, Bosco, and Rosso 2019; Sanguinetti et al. 2018). 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 treebank is composed of around 65% of marked sentences and 35% of non-marked ones. The latter have been selected because they present a syntactic

* Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy; University of Trento, Trento (Italy). E-mail: tpaccosi@fbk.eu, teresa.paccosi@unitn.it
** Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy. E-mail: aprosio@fbk.eu
† Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy. E-mail: satonelli@fbk.eu

© 2023 Associazione Italiana di Linguistica Computazionale

structure that may resemble marked ones and that could therefore be misleading for parsers. Both types of samples have been extracted from a corpus of students' essays and to our knowledge represent 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 was accepted under the release of Universal Dependencies 2.10 and it is freely available on Github at `https://github.com/dhfbk/markit`. While a first version of the resource was presented in (Paccosi, Aprosio, and Tonelli 2021), the latest release includes 1,292 sentences (around 400 more than the first release) and some adjustments of the first annotation guidelines have been performed.

2. Background

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, Montemagni, and Simi 2013). Other treebanks have been later built, covering a rich collection of different usages and genres. In particular, the VIT treebank (Delmonte, Bristot, and Tonelli 2007) is composed of several texts ranging from news to literature, while TWITTIRO (Cignarella, Bosco, and Rosso 2019) and PoSTWITA (Sanguinetti et al. 2018) are two treebanks composed of tweets. These 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 UD treebank of grammar examples for Italian does not exist, and also in other languages there are only few examples. A treebank of grammar examples 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, Brunato, and Dell'Orletta 2016), but the different structures were identified using automated tools.

According to (Haspelmath 2006), the intuitive shared sense of "markedness" is unusual or uncommon and the author provides twelve different meanings which have been assigned to the word "markedness" throughout time and in different linguistic subfields (such as phonetics, syntax or semantics). In the present paper, we adopt a notion of syntactical markedness which is situated between two of the senses proposed by (Haspelmath 2006): a deviation from default parameter setting and a specifically defined distribution opposed to a default one. More precisely, with marked sentences we refer here to those constructions which present a non-canonical order of constituents, assuming that 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. Marked structures are 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, for reasons that are phonotactic or bound to the whole meaning of the sentence (Benincà, Gianpaolo, and Lorenza

1988). In syntactical terms, we can say that marked structures operate a modification in the distribution of *topic* and *comment* with respect to the corresponding non marked structure (Cinque 1990). The notion of markedness has been quite discussed over time, especially in quantitative studies on word order (Merlo 2016; Futrell, Mahowald, and Gibson 2015), but we are not interested here in contributing to the debate on specific definitions of markedness. Instead, we adopt the notion that considers as marked those sentences presenting a specific deviation from the SVO order in Italian, distinguishing among different types of marked structures (see Section 4). (Delmonte 2016) already demonstrated that non-canonical sentences are very difficult to parse. The author tested in fact four parsers on a small dataset containing several examples of non-canonical structures, yielding a low accuracy in all the cases. For an example of wrong parsing output of a marked structure (dislocation), see Figure 1, where the object *"La mamma"* ("the mum") is labeled as nsubj. In the light of such findings, we argue therefore that it is crucial to make parsers more robust to this kind of constructions and, in general, to different syntactic structures. It is also important to agree on a standard way to annotate different marked structures in UD, which is part of our contribution presented in this work.

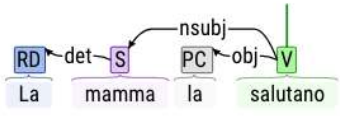

Figure 1
Wrong parsing output of dislocation (En: *The mum, they greet her*)

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, given that a large number of marked structures could be found in the IPRASE corpus of written Italian (Sprugnoli et al. 2018). This corpus is comprised 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). The corpus contains more than 2,500 essays and almost 1.5 million tokens. 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 (Tonelli et al. 2020). Among others, annotators were asked to mark dislocated sentences, cleft sentences and hanging topics (see details in Section 4). So, we start from these annotated sentences and then revise them, adding also some types of markedness that were not initially foreseen. In this respect, students' essays are a very interesting textual source to analyse: on the one hand, they were written in a formal educational setting by students who tried their best to explain their thoughts clearly and correctly, therefore the essays tend to be free from grammatical errors. On the other hand, they may reflect a still incomplete mastery of writing skills and include expressions that are admissible only in informal spoken language.

We extract from the IPRASE corpus around 1,300 sentences and annotate them at syntactic level. The sentences are first automatically parsed using the TINT NLP Suite (Palmero Aprosio 2021) and then manually revised by a linguist to distinguish between the constructions of interest and other types of similar constructions. 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. Also different writing skills can be detected among the students.

4. Overview of Marked Structures

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 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, so that in the end all seven phenomena are present. We collect from the IPRASE corpus left and right dislocations, cleft, and some hanging topic sentences. As we said above, during the annotation process we found instances of other marked structures erroneously identified as cleft or dislocated. Below we report a brief description of the 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 if the displaced element was the direct object, as long as it is in the positive form (Benincà, Gianpaolo, and Lorenza 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

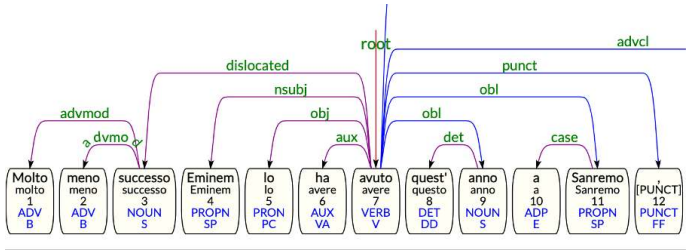

Figure 2
Left dislocated sentence annotated with *dislocated* relation

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 (4). When there is no reprise of the dislocated element in the subsequent sentence (3), we refer to that as an example of anacoluthon (Ferrari and Zampese 2016).

(3) ma il cervello, senza non siamo niente
But the brain, without (it) we are nothing

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

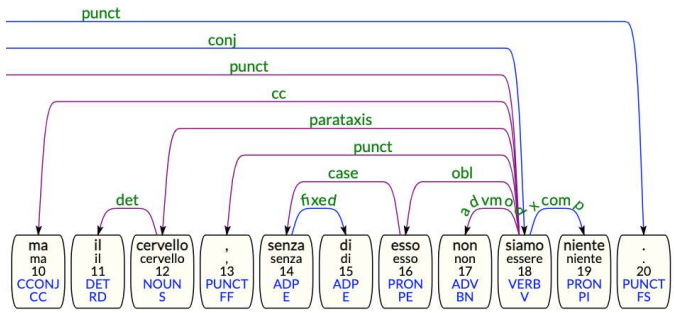

Figure 3
Hanging topic annotated with *parataxis* relation

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 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 (5) or of the comma compulsory.

(5) 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

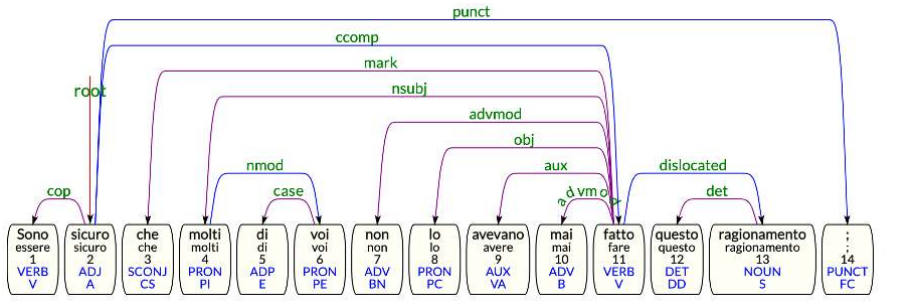

Figure 4
Right dislocated sentence annotated with *dislocated* relation

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 (8). 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). The nature of the clause following "che" in cleft sentences is controversial. It has always been interpreted as a relative clause, but this interpretation holds true only when the dislocated element is the subject or the direct object. When the dislocated element is an adverbial element or a whole subordinate clause it is not easy to individuate the antecedent, and then the clause cannot be considered a relative (Quirk et al., 1985). We are not interested here in defining the nature of this structure, but we only want to put the accent on differentiating these two structures, as we have annotated them in two different ways (see Fig. 5 and Fig. 6).

(6) Ed è proprio il raggiungimento di sfide che sembrano impossibili che causa una maggiore felicità
And it really is the achievement of challenges which seem impossible that provides a greater happiness

(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

(8) È con questo spirito che si sta progettando una nuova missione
It is with this spirit that we are planning a new mission

4.5 Pseudo-cleft sentences

Similar to the cleft structure, the pseudo-cleft sentence is characterized by two core units: one usually introduced by the relative pronoun "chi" (who) or by the "demonstrative+that" construction which contains the verb, and the other introduced by the copula

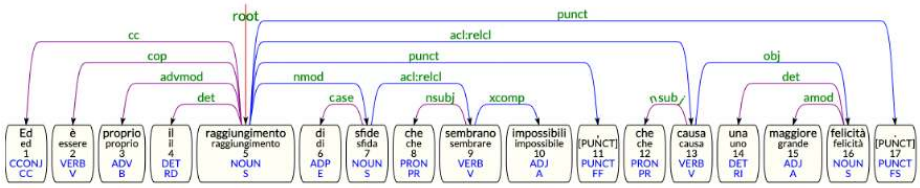

Figure 5
Cleft sentence with relative clause (*acl:relcl*)

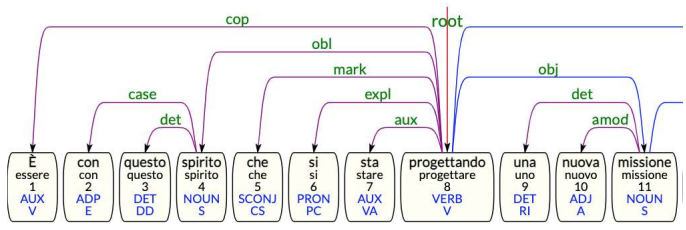

Figure 6
Cleft sentence with *obl* relation (oblique nominal)

and containing the new element we want to emphasize and put into focus (De Cesare 2005).

(9) Ciò/quello che distingue gli ultimi due episodi rispetto ai primi è il silenzio
What (that) differentiates the last two episodes from the early ones is the silence

(10) Chi ha mangiato la torta è Giorgia
Who ate the cake is Giorgia

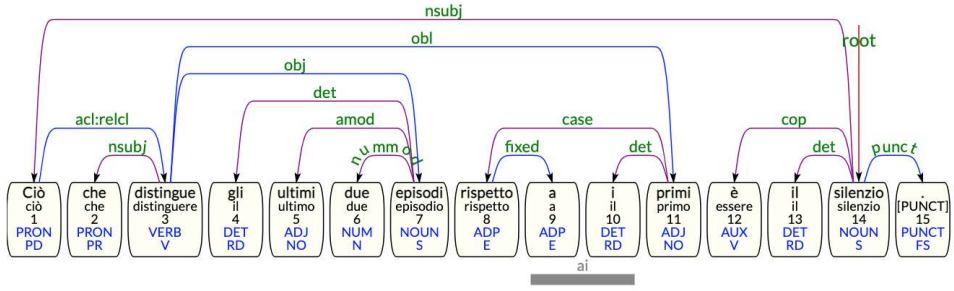

Figure 7
Pseudo-cleft sentence

4.6 Sentences with presentative there

The "presentative there" is a structure which is composed by "there+verb to be", introducing a nominal element followed by a pseudorelative (which is a fundamental component of the presentative structure). This structure is different from other constructions having "there" at the beginning of the clause because of its focalizing property. In fact, several constructions with the structure "there+verb to be" exist but only the "presentative there" puts into focus the nominal element of the clause and it is then considered marked. Sentences such as "C'è Lucia" ("There is Lucia") or "Ci sono i pinguini al polo sud" ("There are penguins at the South Pole") have respectively an existential and a locative function. Even if they are very similar to the "presentative there"'s structure, the difference is that the element in these cases is introduced but not focalized. We choose to use the "presentative there" label instead of the one used in (De Cesare and Ferrari 2007) ("focalizing there") because it is traditionally more recognizable than this new label but we consider only the case in which this structure is marked, as in (Berruto 1986).

(11) C'è una canzone che mi ha fatto particolarmente riflettere da quando l'ho sentita
There is a song that made me particularly reflect since I heard it

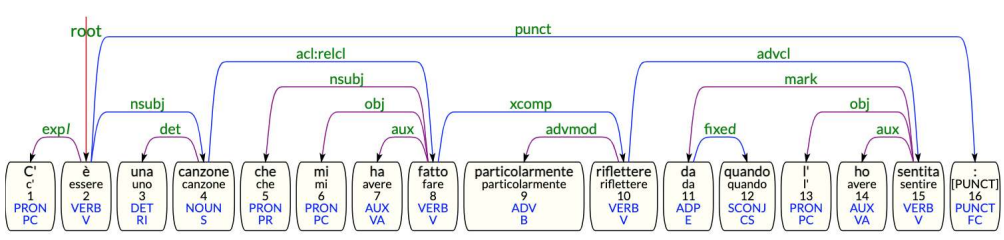

Figure 8
Presentative there sentence

4.7 Dislocation of the subject

The dislocation of subject presents some issues because it is difficult to be recognized in the written form. In fact, while prosodic information marks the topicalization of the dislocated element in the oral form, in the written form intonation and pauses, which usually allow the reader to recognize whether it is dislocated or not, are not present. The literature on subject dislocation is very rich, and it crosses different frameworks (Cardinaletti 2018; De Cesare 2014), but we are interested here in selecting the more intuitive methodology to determine if a subject is dislocated or not. For what concerns the left dislocation of the subject, the difficulty is given by the fact that the subject is already placed before the verb in the canonical form. In order to recognise whether a subject is dislocated or not, we therefore choose to follow the intuition in (Ferrari and Zampese 2016), according to which a subject can be generally dislocated to the left when it is separated from the verb by other syntactic material, such as another dislocated element ("Giulio, la macchina, me l'ha prestata"), or some other elements in incidental position. In some cases, it is possible to focus on the subject through a tonic pronominal

reprise ("Lucia, **te** proprio non sai dire di no")[1] or through a demonstrative ("È di certo un progetto ambizioso, **quello** di andare alla ricerca dei meccanismi del cervello")[2].

In Italian, dislocated subject on the right does not present explicit traits of dislocation, which makes it difficult to distinguish between a dislocation or a postverbal explicit subject. The postverbal position of subject does not entail that this is the result of a dislocation. Therefore, we adopt the syntactic distinction criterion, for which if the subject is separated by the verb by a comma or by other constituents, which must be possibly placed in the first left position (Benincà, Gianpaolo, and Lorenza 1988), the subject is right dislocated (12), otherwise it is explicit (13) (what we have called *postverbal subject* in Table 1).

(12) È così magnificamente astratto, il limite che divide l'individuo dalla società
It is so wonderfully abstract, the limit which divides the individual from society

(13) A questo ordigno accenna anche Heisenberg
To this device refers also Heisenberg

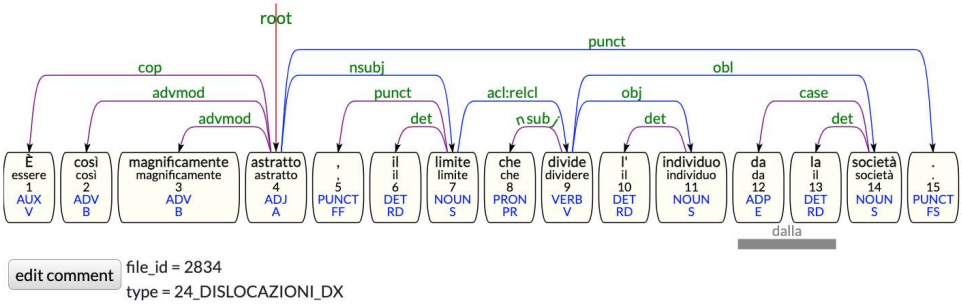

Figure 9
Right dislocated subject marked with *nsubj* relation and labeled "right dislocation"

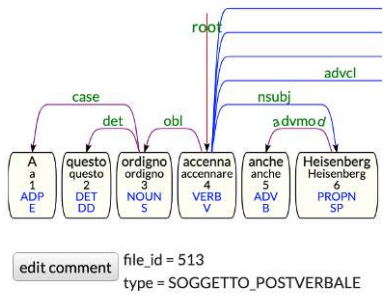

Figure 10
Explicit subject with postverbal position marked with *nsubj* relation and labeled "postverbal subject"

1 EN: "Lucia, **(atonic pronoun of you)** is not able to say no"
2 EN: "It is an ambitious project for sure, **that** of searching the mechanisms of brain"

5. MarkIT annotation and Statistics

Marked structures 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 (Palmero Aprosio 2021), which is built following Universal Dependencies guidelines (De Marneffe and Manning 2008), and trained on the Italian Stanford Dependency Treebank, ISDT (Bosco, Montemagni, and Simi 2013). The dependency trees parsed by TINT are then manually corrected by a linguist using the CoNLL-U Editor (Heinecke 2019). CoNLL-U editor is a very comprehensive tool for the manual annotation of dependency trees. We correct the dependencies pre-processed by TINT using both the flat graph and the tree graph forms and we use the comment section to label the type of marked structure represented in every sentence, as displayed in Fig. 11.

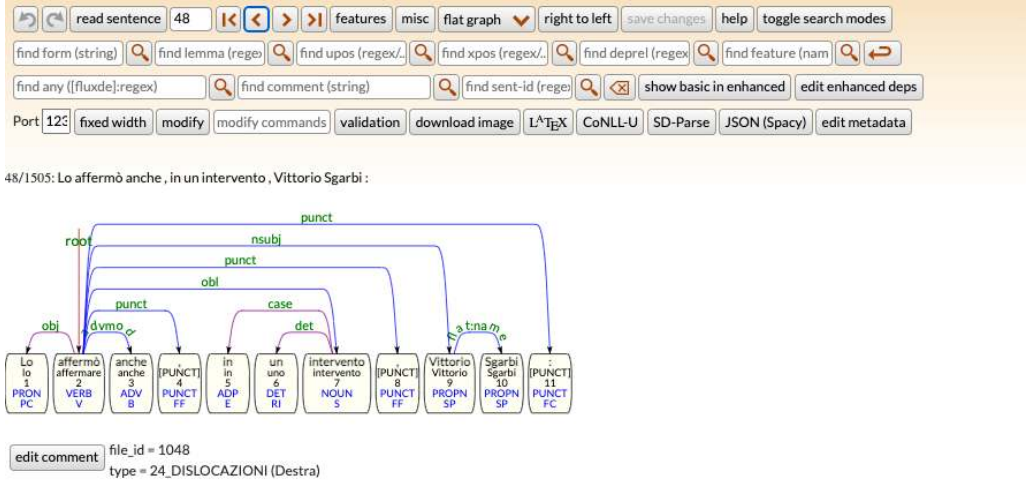

Figure 11
Screenshot of CoNLL-U Editor

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 1. 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. 12 and Fig. 13. We do not use the *dislocated* label to mark right dislocation of the subject without pronominal reprise (see Fig. 11). The distinction between right dislocation of the subject and postverbal explicit subject is maintained only in the label used in the comment section.

As previously mentioned, **hanging topics** differ from left dislocated sentences because the element to the left is not syntactically linked to the 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, given that *parataxis* is defined as a relation between a word and other elements, without any explicit coordination, subordination, or argument relation with the head word (usually

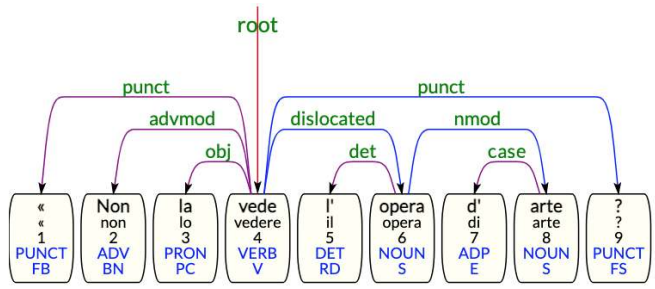

Figure 12
Dependency tree with right dislocated sentence (En: *Does he see it, the artwork?*)

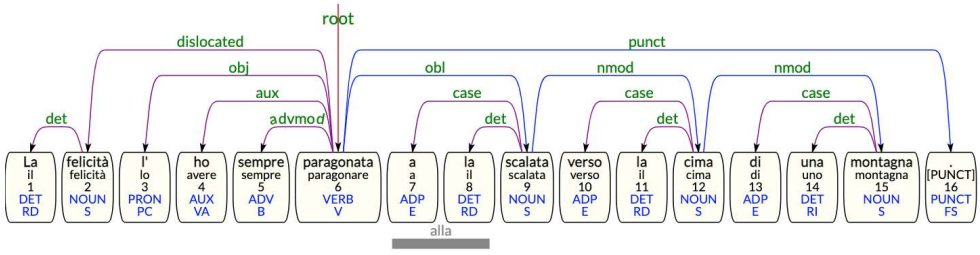

Figure 13
Dependency tree with left dislocated sentence (En: *The happiness, I always compared it with the climb towards the top of the mountain*)

the verb). An example is reported in Figure 14. Parataxis is mostly used in Italian for reported speech or parenthetical clauses, but in ISDT and in PoSTWITA corpora it is possible to find some examples of this relation used also for structures similar to hanging topics, even without an explicit focalization.

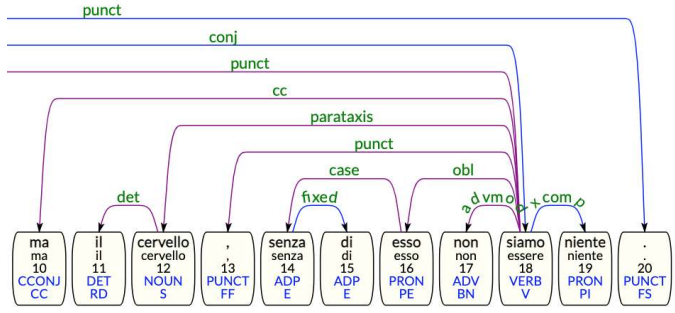

Figure 14
Dependency tree with hanging topic sentence (En: *But the brain, without it we are nothing*)

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 initially chose 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, in the following way. 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. 5). Instead, if there is no dislocation of the subject or the object, we used the *acl* relation, initially considering that it always modifies a nominal element (even if introduced by a preposition). However, we do not select the function of "che", treating it as a mere introducer for the subordinate clause with the *mark* relation (Fig. 15). We annotated in fact the nominal element introduced by a preposition as a copula+predicative form and the "subordinate clause" with *acl*. During a second revision, we actually decide that, given the controversial nature of the clause introduced by "che", we consider as the root the verb of the clause following the prepositional one and we mark with the *obl* relation the clause introduced by the preposition, as represented in Fig. 16.

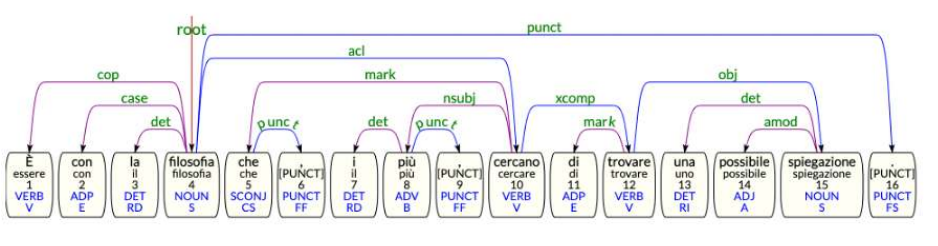

Figure 15
First version of the annotation of cleft sentence with *acl* relation (adnominal clause) (En: *It is trough philosophy that most people look for a possible explanation*)

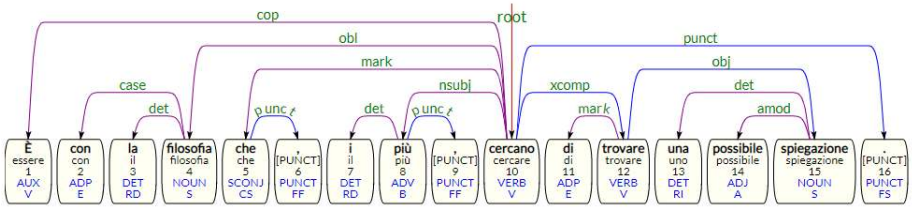

Figure 16
New annotation of the same cleft sentence in Fig.15 annotated with *obl* relation (oblique nominal)

Furthermore, we include in the "False marked" category the structures which resulted challenging to tag for the annotators and which are not marked. "False Marked" includes in fact those structures which usually present an explicit subject in the main clause and are erroneously identified as cleft, for example the following sentence:

(14) L'Agenzia Spaziale Europea (ESA) è l'organizzatrice di questa nuova missione che sembra creare grandi aspettative verso Marte
The European Space Agency (ESA) is the organizer of this new mission that seems to create great expectations towards Mars.

The elements such as "l'organizzatrice di questa missione" were erroneously identified as the focalized element of the cleft structure but they actually are the predicates nominative of the subject (in this example, "l'Agenzia Spaziale Europea"). This error in the identification of marked sentences by IPRASE annotators is the most common and it represents the majority of structures in the set of "False marked".

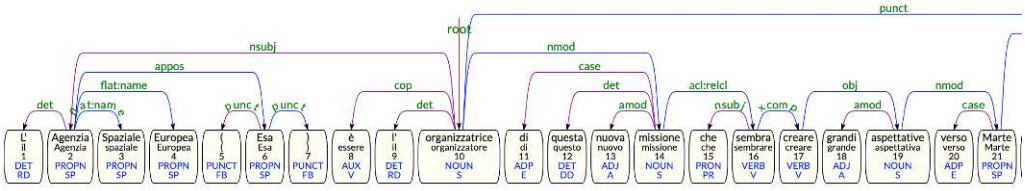

Figure 17
Dependency tree with false marked with explicit subject and *acl:rel* relation

"False Marked" includes also passive clauses, which were originally tagged as right dislocated because of the postverbal position of the subject:

(15) Il 6 e il 9 agosto del 1945 furono sganciate le due bombe atomiche americane su Nagasaki e Hiroshima
On August 6th and 9th 1954, the two American atomic bombs were dropped on Hiroshima and Nagasaki.

Sentences in this last category are particularly challenging both for parsers and for human annotators, given that they were wrongly classified even by IPRASE experts (i.e. high-school teachers) and have been assigned the correct label only after our revision.

We report in Table 1 a summary of the sentences belonging to the eight categories contained in MarkIT. "Cleft sentences" is the most frequent category in the treebank, followed by "Left" and "Right dislocated sentences". This may be due to the fact that these were the categories originally included in the IPRASE corpus, together with hanging topic. It is interesting to note that "False marked" is well-represented in the resource, providing the possibility to have a good number of extremely challenging negative examples to train a system for marked structures recognition.

6. Parsing Evaluation

As already mentioned in Section 1, the lack of marked structures in treebanks used to train syntactic parsers may affect the robustness of existing systems, 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 after training it on different datasets. In particular, we first split our novel treebank into training, dev, and test, respectively 50%, 25%, and 25%, 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 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 together (see Section 2),

Table 1
Number of examples annotated for each category in MarkIT

Type	Sents	Train	Dev	Test
Cleft sentences	484	230	127	127
Left dislocated	206	98	54	54
Right dislocated	50	23	14	13
Postverbal subject	47	22	12	13
Presentative "there"	31	13	9	9
Pseudo-clefts	14	6	4	4
Hanging topic	7	3	2	2
False marked	453	216	118	119
Total	1,292	611	340	341
Total (tokens)	40,488	19,893	10,399	10,196

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 (i.e. the ones derived from Twitter) because their particularly informal language is very different from MarkIT sentences, where we would not find hashtags, abbreviations, user mentions, etc. In both cases, we use the concatenation of the development sets of the four datasets as development set during the training phase. We also compare the models by testing them on the original test sets belonging to ISDT, VIT, and ParTUT (concatenated).

Following the standard evaluation used in dependency parsing, we compute unlabeled attachment score (UAS) and labeled attachment score (LAS) in the two tests.

Table 2
Parsing performance of the model without (1) and with (2) MarkIT in the training set. We compare the results on two test sets: MarkIT and ISDT+VIT+ParTut.

	Training set	MarkIT		ISDT+VIT+ParTut	
		UAS	LAS	UAS	LAS
(1)	ISDT+VIT+ParTut	82.78	77.76	86.01	82.58
(2)	ISDT+VIT+ParTut+MarkIT	83.47	78.55	86.10	82.57

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 in the MarkIT 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, both parsing models perform consistently

worse on MarkIT structures than on the other test set. Results in Table 2 suggest also that the performance of the parser does not drop when adding MarkIT to the training set, and the increased performance on MarkIT alone is not simply due to a slightly bigger training set.

In order to have better insights into the above results, we perform a second evaluation comparing the performance of the two models introduced in Table 2 on the different sentence types in MarkIT as separate test sets. Results are reported in Table 3.

Table 3
Parsing performance without (1) and with (2) MarkIT added to the training set on the different sentence types. Values in "Num" column refer to the number of instances in the test set.

Type	Num	(1)		(2)	
		UAS	LAS	UAS	LAS
Cleft sentences	127	81.16	76.60	**83.80**	**79.06**
Left dislocated	54	82.42	76.05	82.42	**76.56**
Right dislocated	13	**83.94**	**77.27**	83.64	75.76
Postverbal subject	13	84.39	79.77	**86.99**	**80.64**
Presentative "there"	9	**88.00**	**85.09**	87.27	83.64
Pseudo-clefts	4	83.57	80.71	83.57	80.71
Hanging topic	2	67.44	60.47	67.44	60.47
False marked	119	**84.03**	**79.01**	83.16	78.69

As expected, performance improves adding MarkIT to the training set for the categories that are represented by more examples in the treebank, as in the case of cleft sentences, left dislocated and postverbal subject. Right dislocated sentences represent an exception, and we argue that a lower accuracy when MarkIT is added to the training set (2) is due to the existence of two possible annotations of the right dislocated element. In fact, in this category we include both dislocations of the subject on the right, according to the criteria illustrated in Section 4.7, which we annotated with the *nsubj* relation, and pronominal reprise, labeled with the *dislocated* relation (Sec. 4.3). In the case of pesudo-clefts and hanging topics, the instances in the training set are not enough to have a real impact on the performance of the model, and we argue that this is the same reason why "presentative there" has no improvement. We also hypothesize that a slight drop in accuracy for this latter category can derive from the existence of several structures similar to the "presentative there" (having then "there+verb to be" as we have seen in 4.6) in the training set. Considering false marked sentences, even if there is a good number of instances, we argue that a better accuracy with training (1) is due to the fact that these structures are better parsed if a model is trained without marked structures. When MarkIT is added to the training set, the system is less likely to distinguish between marked and unmarked structures. The relatively poor performance of the parser trained without MarkIT can be also explained by the fact that there are some differences in the annotation scheme we adopt for this treebank (see Section 5). The present study enlightened the need for a more homogeneous treatment of the annotation of non-canonical structures of Italian. We think that treebanks built on specific grammatical constructions represent a step forward in this direction. Further investigation is needed

in the future to determine to which extent consistency in annotation plays a role in evaluating parsers.

7. Release

Although the IPRASE corpus is not available because of copyright issues, the sentences in MarkIT have been extracted without any additional information related to the authors or the textual context, and they can therefore be freely distributed. MarkIT is released under CC BY 4.0 license,[3] and can be downloaded from Github.[4] Starting from version 2.10, MarkIT is also included in the official release of Universal Dependencies.

8. Conclusions

In this work, we present the final release of MarkIT, a treebank composed of almost 1,300 sentences with syntactic annotation of both marked and non-marked (but challenging) structures. The former include seven types of marked sentences, which we annotate manually after defining some guidelines to label their syntactic relations consistently. We also perform a parsing evaluation on two different test sets, comparing a dependency parser performance with and without MarkIT in the training data.

Our results show that the configuration with MarkIT yields some slight improvements, which are probably due to the small size of the treebank compared to other existing ones. Indeed, the marked structures that are more frequent in MarkIT are also those that are parsed more correctly. The treebank is made available to the research community, being included in the release of Universal Dependencies 2.10. Our goal is to make dependency parsers more robust to the different syntactic structures present in Italian, in particular in the neo-standard variant, but also to contribute to a standardisation of syntactic annotation of markedness phenomena.

References

Benincà, Paola, Salvi Gianpaolo, and Frison Lorenza. 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*. Il Mulino, pages 115–225.

Berruto, Gaetano. 1986. Un tratto sintattico dell'italiano parlato: il c'è presentativo. *Parallela*, 2:61–73.

Berruto, Gaetano and Massimo Cerruti. 2011. *La linguistica: un corso introduttivo*. UTET Università.

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

Bosco, Cristina, 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.

Cardinaletti, Anna. 2018. On different types of postverbal subjects in Italian. *The Italian Journal of Linguistics*, 30:79–106.

Cignarella, Alessandra Teresa, 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.

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

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

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

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

De Cesare, Anna-Maria. 2005. La frase pseudoscissa in italiano contemporaneo. aspetti semantici, pragmatici e testuali. *Studi di grammatica italiana*, 24:293–322, 01.
De Cesare, Anna-Maria. 2014. Subject dislocations in contemporary Italian and in a contrastive perspective with French. In *Tra romanistica e germanistica: lingua, testo, cognozione e cultura/Between Romance and Germanic: Language, text, cognition and culture*. Peter Lang, pages 35–54.
De Cesare, Anna-Maria and Angela Ferrari. 2007. *Lessico, grammatica, testualità*, volume 18. Universität Basel.
De Marneffe, Marie-Catherine 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.
Delmonte, Rodolfo. 2016. Syntactic and lexical complexity in Italian noncanonical structures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 67–78, Osaka, Japan, December. The COLING 2016 Organizing Committee.
Delmonte, Rodolfo, Antonella Bristot, and Sara Tonelli. 2007. VIT-Venice Italian Treebank: Syntactic and Quantitative Features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54, Bergen, Norway, December. Northern European Association for Language Technol.
Ferrari, Angela and Luciano Zampese. 2016. *Grammatica: parole, frasi, testi dell'italiano*. Carocci editore.
Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden, August.
Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of linguistics*, 42(1):25–70.
Heinecke, Johannes. 2019. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies*, Paris, France, August.
Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June.
Merlo, Paola. 2016. Quantitative computational syntax: some initial results. *IJCoL. Italian Journal of Computational Linguistics*, 2(2-1).
Munro, Robert M. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
Paccosi, Teresa, Alessio Palmero Aprosio, and Sara Tonelli. 2021. It Is MarkIT That Is New: An Italian Treebank of Marked Constructions. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, Milano, Italy, June-July 2022.
Palmero Aprosio, Alessio. 2021. Tint, the Swiss-Army Tool for Natural Language Processing in Italian. In *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021)*, Online, November.
Pieri, Giulia, 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 *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.
Renzi, Lorenzo. 2001. Le tendenze dell'italiano contemporaneo. note sul cambiamento linguistico nel breve periodo. *Studi di lessicografia italiana*, pages 279–319.
Sanguinetti, Manuela and Cristina Bosco. 2014. Converting the parallel treebank ParTUT in Universal Stanford Dependencies. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, pages 316–321, Pisa, Italy, December. Pisa University Press.
Sanguinetti, Manuela, 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).
Sprugnoli, Rachele, Sara Tonelli, Alessio Palmero Aprosio, and Giovanni Moretti. 2018. Analysing the evolution of students' writing skills and the impact of neo-standard italian with

the help of computational linguistics. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 354–359, Torino, Italy, December. aAcademia University Press.

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