IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 2, Number 1
june 2016

Emerging Topics at the Second Italian Conference on Computational Linguistics

aA
ccademia
university
press

editorial board

editorial office

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99200-99-2

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_02

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 2, Number 1
june 2016

Emerging Topics at the Second Italian Conference on Computational Linguistics

CONTENTS

ISACCO: a corpus for investigating spoken and written language development in Italian school–age children

Dominique Brunato[*]
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)

Felice Dell'Orletta[**]
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)

In this paper we present ISACCO (Italian School–Age Children COrpus), a corpus of oral and written retellings of Italian-speaking children attending primary school. All texts were digitalized and automatically enriched with multi–level linguistic annotation. Preliminary explorations of both the form and the content of children's productions were carried out based on a set of features automatically extracted by NLP tools. Written retellings were manually annotated with a typology of errors belonging to three different linguistic levels. The resource, which has been made publicly available[1]*, is conceived to support research and computational modeling of "later language acquisition", with an emphasis on comparative assessment of the evolution of oral and written language competencies in early school grades.*

1. Introduction

The use of naturalistic data to investigate child language features and development over time has a well-established tradition in first language acquisition research (L1). The most notable example is the CHILDES database (MacWhinney 2000), which contains transcripts of spoken interactions involving children of different ages for over 25 languages, Italian included. Yet, CHILDES data refer especially to preschool children, with only a minor section dedicated to their older mates, thus making this resource less adequate for studying how language skills evolve during early schooling. The rapid and remarkable changes children's language undergoes before age five justify the amount of studies for the earliest stages of acquisition. However, over the last years research on "later language acquisition"[2] is also gaining more attention prompted by the awareness that "becoming a native speaker is a rapid and highly efficient process but becoming a proficient speaker takes a long time" (Berman 2004). Under the influence of schooling and literacy instruction indeed, language competence keeps growing in a way that affects all domains (i.e. phonology, morphology, syntax, semantics and discourse) and extends to new modalities, i.e. reading and writing (Koutsoftas 2013).

In this paper we focus on oral and written language competence in early school–years and present a new resource of child language data for Italian, the ISACCO (Italian

[*] ItaliaNLP Lab (www.italianlp.it), ILC–CNR - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: dominique.brunato@ilc.cnr.it

[**] ItaliaNLP Lab (www.italianlp.it), ILC–CNR - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: felice.dellorletta@ilc.cnr.it

[1] The resource is available at the following link:
http://www.italianlp.it/software-data/isacco-italian-school-age-children-corpus/

[2] With this expression we refer to child language development after age five (Tolchinsky 2004).

School–Age Children COrpus) corpus. ISACCO has been collected with the aim of investigating in a comparative fashion the effect of the diamesic variation on children's narrative abilities. The peculiarities of speech and writing, and the way they reflect on discourse structure, have been widely studied in linguistics research from different theoretical standpoints (Cf. section 2). However, the majority of studies has focused on texts produced by adult speakers or children with proficient writing skills, while less attention has been paid to younger children, i.e. children attending early elementary-school years. Although it is acknowledged that writing is particularly demanding in the early stages of development, since it requires additional processes not involved in speech (e.g. graphomotor skills, phoneme–grapheme conversion mechanisms), how and to what extent these factors impact on written productions is not straightforward to understand.

With respect to the methodological approach to inspect child language data, over the last few years the interest of researchers is moving towards tools and techniques drawn from computational linguistics and Natural Language Processing (NLP), which have been adopted especially for the English language. The use of a statistical parser is reported e.g. by (Sagae, Lavie, and MacWhinney 2005) and (Lu 2009) to automate sophisticated measures of syntactic development like the Index of Productive Syntax (Scarborough 1990), reaching performances comparable to those obtained by manual annotation. A more challenging step has been recently made by (Lubetich and Sagae 2014), who derived a data–driven metric of syntactic development using Part–of–Speech (PoS) and parse tree features. Beyond typical language development, computational linguistics approaches are also employed in clinical settings, e.g. to identify markers of Autism Spectrum Disorders in children's speech by integrating features from automatic morpho–syntactic and syntactic annotation (Prud'hommeaux et al. 2011), as well as metrics of semantic similarity (Rouhizadeh, Sproat, and van Santen 2015).

Although this paper focuses mainly on the resource, we will also discuss some preliminary analyses aiming at highlighting how a NLP perspective applied to a corpus like ISACCO can serve as the starting point to conduct computational linguistics explorations at multiple levels, whose potential will be particularly appealing in view of their application to large–scale corpora. Specifically, it could be possible to assess the effect of the diamesic variation on the form and linguistic complexity of children's productions and to detect changes across schooling levels (cf. section 4.1). Beyond the "form", students' texts can be analyzed with respect to the "content" so as to evaluate whether and to what extent children's comprehension and recall change according to linguistic modality and increased literacy development. To this aim, the output of an ontology learning system can provide a mean to investigate whether a child's retelling is coherent with the story she/he heard, in terms of 'matching' ideas and hierarchical organization of the content (cf. section 4.2). This would make it possible to identify patterns of typical development to be used for comparison e.g. in diagnostic settings, with children showing atypical language development.

The rest of the paper is organized as follows: Section 2 provides some theoretical background on the relation between speech and writing. In Section 3 we first introduce the corpus providing details on the children participating in the study and the empirical design we adopted to collect data. We then describe the different tagsets used for manually annotating children's productions, which were intended to distinguish typical phenomena of spoken language for what concerns the oral texts, and to inspect the most common linguistic errors made by children, for what concerns the written texts. Section 4 discusses preliminary explorations of the corpus focused on form and content, which were carried out by comparing the statistical distribution of linguistic and lexico–semantic features automatically extracted from texts using on NLP tools.

2. The Oral and Written Modality of Language

The interplay between oral and written language has been widely investigated from the perspective of many language–related disciplines, as well as in research on literacy development. More than the similarities, scholars emphasized the differences between the two modalities and analysed them with respect to several dimensions according to how speech and writing are compared, i.e. as linguistic systems, communicative acts, cognitive processes and so on (Lintunen and Makila 2014). A common feature of research on this topic is also the tendency to frame the relation between the two modalities in terms of complexity. For instance, from a cognitive perspective, writing is considered as more demanding than speaking because of the different nature of its acquisition and the pool of resources it involves. Indeed, if (typically-developing) children learn spoken language effortless and without explicit teaching, writing is achieved later through formal instruction and depends upon technical components not involved in spoken language, such as spelling and handwriting skills, which require training and experience to be proficiently mastered. A further dimension of variation is related to the role of context in oral and written language. Unlike spoken discourse, which is said to be "context-bound", written texts must function "acontextually", i.e. the receiver has to interpret them without relying on the context (Nystrand 1987). This makes the process of writing more planned, explicit and structured, whereas speech is fragmented and characterised by disfluencies, repetitions and implicit information.

The different situational context surrounding oral and written communication is ultimately responsible of the linguistic and textual differences between spoken and written discourse, which emerge in cross-linguistic corpus–based studies. Also from this perspective, scholars usually attribute a higher degree of complexity to written texts, especially with respect to morpho–syntactic and syntactic parameters taking into account e.g. the proportion between main and subordinate clauses or the predominance of acknowledged complex clauses, such as relative clauses, in writing (Cf., among others, (Chafe 1982) and (Beaman 1984)). On the other hand, in his seminal work on the English language, (Halliday 1989) claims that in spoken language "the sentence structure is highly complex, reaching degrees of complexity that are rarely attained in writing". This is because the online nature of spoken language prevents the speaker from arranging ideas as accurately as she/he would do in writing; such a constraint correlates, at grammatical level, with the use of sentences featuring complex chains of clausal embeddings.

However, the tendency of seeing speech and writing as a dichotomy is actually misleading. For instance, when the role of genre is taken into consideration ((Biber 1988) and subsequent research), it turns out that the spectrum of variation depends upon the examined typology of text, so that greatest differences are likely to appear when comparing e.g. academic writing to informal conversation, which can be seen as the opposite poles of a continuum. From this viewpoint, we shall expect to find few differences between spoken and written productions in early school grades, since the sensibility to style and register variation, together with the ability of differentiating oral and written forms according to the context, audience and purpose of communication, is achieved at more advanced stages of writing development (Kroll 1981).

3. The Corpus

3.1 Participants

Fifty-six children from the 2nd to the 4thgrade of elementary school took part in this study. They were all recruited from a public primary school located in Pisa and examined in the last month of the school year. All children were Italian monolingual speakers, except for two, who were also included in the survey since they had no significant exposure to other languages. None of the children had language disorders or cognitive problems that could interfere with their performance. Details of the sample group are given in Table 1.

Table 1
Children sample group (SD=Standard deviation; m=months).

Grade	Male	Female	Age Mean (SD)
Second	11	8	8.1 years (3.6 m)
Third	10	11	9.0 years (5.6 m)
Fourth	9	7	10.0 years (4.2 m)

3.2 Methodology

The collection of the corpus was inspired by the work of (Silva, Sańchez, and Borzone 2010), who assessed the syntactic complexity of oral and written retellings by monolingual Spanish children attending the first and second grades of primary school. Differently from their work, we excluded from our survey 1stgrade pupils in accordance to the teachers' indications, who pointed out that free written retelling is usually introduced in the curriculum by the end of the second year. We then selected a narrative text from a 3rdgrade book, which was meant to be not too challenging for the youngest nor too easy for the oldest group[3]. Children were tested in two sessions, with a gap of two weeks between each of them, so as to prevent memory bias in their recall. In both sessions, they were exposed to the same story in order to ensure that potential differences in children's performance could not be due to the effect of a different text.

The first session was devoted to collect oral productions. To this aim, the story was read aloud once to the whole class and repeated again, individually, to a restricted group of students, which was randomly chosen by teachers, while their mates carried out another activity related to the story (e.g. drawing a picture). Each selected child was tested alone, in a quiet room, and after hearing the story again was asked to retell it to the examiner. All retellings were recorded and then transcribed according to the standard detailed in Section 3.3.

In the second session, the same story was read again to the whole class and this time all students were asked to produce a written retelling. No limit of time was given and they were left free to write in capital letters or italics.

Although for the purpose of our comparative analysis only the writings of the 56 children tested in the first session were needed, we digitalized all written retellings;

[3] The story, titled "La statua nel parco" by Roberto Piumini, is reported in Appendix.

Table 2
Corpus of oral and written retellings.

Grade	Number of texts	Number of tokens
	Oral retellings	
Second	19	2.029
Third	21	2.994
Fourth	16	2.406
Tot	56	7.429
	Written retellings	
Second	43	4.508
Third	44	4.984
Fourth	38	4.417
Tot	125	13.909

such texts offers indeed valuable material for research on writing development with a view to its computational modeling. The size of the whole corpus comprising the oral and written section is reported in Table 2.

3.3 Oral data transcription

As argued by (Ochs, 1979: 44) [4], any "transcription procedure is responsive to cultural biases and itself biases readings and interpretation"; such a claim highlights that the transcription of oral language cannot be detached from theory, analysis and interpretation. This is particularly true for child language, for which several criteria have been proposed to address issues related e.g. to how speech can be properly segmented into utterances or how written transcripts should be enriched with those paralinguistic and pragmatic features that are necessary for interpretation but not properly conveyed by standard orthography (Cf. (Moneglia and Cresti 1997) for a review).

For the purposes of our research, we manually transcribed children's oral retellings adding some "natural punctuations" (Powers 2005) (i.e. periods and commas) according to speech pauses and intonations; this allowed us to identify major sentence boundaries. These "raw" transcripts were then enriched with additional "xml-style" labels to annotate typical phenomena of spoken language (e.g. false starts, disfluencies). At present, we have not followed the standard CHAT Transcription Format used in CHILDES (MacWhinney 2000) although we are planning to convert the corpus in this format in the near future to refine the analysis of spoken language features. In the current version, oral retellings were annotated according to the following tagset:

- tag *fs*: to mark a false start (covering both a single or a sequence of words).
- tag *rip*: to mark a repeated word. It has the attribute *number* for the number of repetitions made by the child;
- tag *int*: to mark a long interruption (e.g. when the child did not recall some part of the story)

[4] Quoted in (Roberts 2010)).

An example of annotation from the corpus is illustrated below:

(1) *Poi* <fs> <rip number="2"> *fecero* </rip> </fs> *le rondini fecero* <fs> *le due rondinini* </fs> *i due rondinini e in autunno le rondini andarono via.*

Table 3 shows the percentage distribution of false stars and repetitions in oral retellings. Interruptions were not reported here since we found only three occurrences in the whole corpus which, significantly, were all made by 2–grade children. It is worth noticed that the Standard Deviation (SD) is high within all grades; this might suggest that the proportion of disfluencies is not a function of age, although it is necessary to validate these data in a larger sample of children.

Table 3
Distributions of false starts and repetitions in oral retellings.

	False Starts			*Repetitions*		
	% (Abs)	Average	SD	% (Abs)	Average	SD
Second	65.79 (75)	3.95	2.71	31.56 (36)	1.90	2.02
Third	78.48 (113)	5.39	2.96	21.52 (31)	1.48	1.32
Fourth	65.98 (64)	3.56	2.96	34.02 (33)	1.83	1.72

3.4 Linguistic annotation of errors

As part of qualitative analysis, all written texts were digitalized and annotated with different typologies of linguistic errors and the corresponding correction. The notion of error and its systematisation plays a key role in language acquisition research and literacy studies, although more emphasis has been put on L2 learners corpora with the purpose of modeling the development of learners' competencies (Deane and Quinlan 2010), especially in writing, modeling the properties of interlanguage (Brooke and Hirst 2012) or enabling the development of automated systems for error detection and correction.

The annotation of errors in our corpus was manually performed following the tagset recently defined by (Barbagli et al. 2015) which, to our knowledge, is the only existing annotation scheme for classifying errors made by L1 Italian learners. This tagset distinguishes errors into three macro–areas which reflect the domain of linguistic knowledge affected, i.e.: ortography, grammar and lexicon. Each macro–class is further sub-divided into more subclasses codifying the linguistic category and the target modification for the unit erroneously written by the student.

In (2) we provide an example of a sentence from the corpus marked with a sub–type of orthographic error affecting the use of apostrophe (whose relative code in the tagset is *t=27*) and of a syntactic error concerning subject–verb agreement (i.e. *t=13*):

(2) *Quando arrivó l'autunno andarono via peró* <M t="27" c="lasciarono"> *l'asciarono* </M> *dei semini sulla mano del geografo e a primavera* <M t="13" c="nacque"> *nacquero* </M> *un cespuglio di fiorellini.*

Table 4
Linguistic errors tagset and quantitative distributions in written retellings. For each grade are reported: frequency distribution (%) and number of occurrences (Abs), average occurrence per year (Avg), Standard Deviation (SD).

		II grade		**III grade**		**IV grade**	
Class of Error	**Target modification**	% (Abs)	Avg (SD)	% (Abs)	Avg (SD)	% (Abs)	Avg (SD)
Orthography							
Consonant doubling	Omission	10.62 (46)	1.07 (2.72)	1.90(4)	0.09 (0.36)	5.52 (8)	0.21 (0.58)
	Excess	2.31 (10)	0.23 (0.53)	1.42 (3)	0.07 (0.25)	2.07 (3)	0.08 (0.27)
Use of *H*	Omission	0.70 (3)	0.07 (0.34)	0.95 (2)	0.05 (0.21)	0.00 (0)	0.00 (0.00)
	Excess	0.23 (1)	0.02 (0.15)	0.00 (0)	0.00 (0.00)	0.00 (0)	0.00 (0.00)
Monosyllabic words	Misspelling of stressed monosyllables	2.31 (10)	0.23 (0.53)	6.64 (14)	0.32 (0.56)	1.38 (2)	0.05 (0.32)
	Misspelling of *po'*	3.70 (16)	0.37 (0.58)	4.74 (10)	0.23 (0.09)	4.14 (6)	0.16 (0.44)
Apostrophe	Misuse	3.93 (17)	0.4 (0.73)	1.90 (4)	0.23 (0.09)	0.69 (1)	0.03 (0.16)
Other		31.40 (136)	2.25 (3.16)	30.33 (64)	1.45 (1.78)	40.00 (58)	1.53 (1.94)
Grammar							
Verbs	Use of tense	23.33 (101)	2.38 (2.35)	15.16 (32)	0.73 (1.15)	8.28 (12)	0.32 (0.70)
	Use of mode	0.00 (0)	0.00 (0.00)	0.00 (0)	0.00 (0.00)	0.69 (1)	0.03 (0.16)
	Subject-Verb agreement	2.78 (12)	0.28 (0.83)	6.64 (14)	0.32 (0.70)	5.52 (8)	0.21 (0.47)
Prepositions	Misuse	1.85 (8)	0.19 (0.40)	3.32 (7)	0.16 (0.53)	1.38 (2)	0.05 (0.23)
	Omission or Excess	1.62 (7)	0.16 (0.37)	0.47 (1)	0.02 (0.53)	1.38 (2)	0.05 (0.23)
Pronouns	Misuse	0.23 (1)	0.02 (0.15)	0.47 (1)	0.02 (0.53)	0.69 (1)	0.03 (0.16)
	Omission	0.23 (1)	0.02 (0.15)	0.47 (1)	0.02 (0.15)	1.38 (2)	0.05 (0.23)
	Excess	0.23 (1)	0.02 (0.15)	0.47 (1)	0.02 (0.15)	1.38 (2)	0.05 (0.23)
	Misuse of relative pronoun	0.23 (1)	0.02 (0.15)	0.47 (1)	0.02 (0.15)	1.38 (2)	0.05 (0.23)
Articles	Misuse	0.00 (0)	0.00 (0.00)	0.00 (0)	0.00 (0.00)	0.69 (1)	0.03 (0.16)
Conj. and/or Conn.	Misuse	0.23 (1)	0.02 (0.15)	0.47 (1)	0.02 (0.15)	0.69 (1)	0.03 (0.16)
Other		9.24 (40)	0.92 (1.08)	12.32 (26)	0.59 (0.87)	11.72 (17)	0.45 (0.69)
Lexicon							
Vocabulary	Misuse of terms	4.85 (21)	0.49 (0.96)	11.85 (25)	0.57 (1.13)	11.03 (16)	0.42 (0.76)
	Total number of errors	587		211		145	

Results reported in Table 4 allow us to provide a preliminary descriptive picture of the role of errors in assessing writing competence in primary schools. We can observe that orthography is the most problematic domain for pupils across all grades, although with some improvements in the acquisition of the spelling rules concerning consonant doubling and apostrophe. A more significant development affects grammatical knowledge, which is testified by the overall decrease of morpho–syntactic and syntactic errors from the second to the fourth school grade, especially those affecting the use of tenses. On the other hand, four–grade writers still make some lexical mistakes of different types, such as the use of nonstandard and dialectal forms or confused words with formal similarity (e.g. "geologo" [*geologist*] instead of "geografo" [*geographer*]).

4. Preliminary explorations of the corpus

This section presents preliminary explorations of the corpus comparing oral and written retellings with respect to both linguistic structure and content. To this aim, the corpus was automatically annotated using an automatic NLP pipeline, which was the prerequisite for the extraction of multi–level linguistic and lexico–semantic features. Specifically, all texts were automatically tagged with the part–of–speech tagger described in (Dell'Orletta 2009) and dependency–parsed by the DeSR parser (Attardi 2009) using Multilayer Perceptron as learning algorithm.

Before discussing the results of these analyses, it is worth pointing out that the typology of texts under examination is particularly challenging for *general-purpose* text analysis tools; this is not only due to the features of spoken language but also to missing punctuation (especially in the 2nd grade writings), which already impacts on the coarsest levels of text analysis, i.e. sentence splitting. Although we plan to evaluate more in detail the impact of these non–standard patterns on linguistic annotation, we believe that some features extracted from linguistically annotated texts are robust enough to offer a first insight into the linguistic structure of children's texts according to age and modality, as well as with respect to the content.

4.1 First results on linguistic structure

Based on the output of the automatic linguistic annotation described in Section 4, the corpus was inspected using MONITOR–IT[5]. This is a tool able to carry out the linguistic profiling of texts following the methodology devised by (Dell'Orletta, Montemagni, and Venturi 2013), that relies on the wide set of linguistic features extracted from the output of the different levels of automatic linguistic analysis (i.e. tokenization, lemmatization and POS tagging, syntactic dependency parsing). Table 5 shows a subset of the monitored features indicating those for which the average difference value between children's oral and written samples was significant[6].

Starting from superficial features (i.e. features available from sentence splitting and tokenization), it can be noted that oral retellings are on average longer than written ones ([1]); this is in line with previous findings in the literature and it possibly reflects the heavy cognitive demands posed by writing in the early developmental phases, which impact on memory reducing the quantity of information recalled. Oral retellings also tend to exhibit slightly shorter words. This finding can be elaborated by looking at the PoS distribution, where we find a greater distribution of words belonging to functional categories (particularly, Pronouns [9] and Conjunctions [5,10]) in oral than in written texts. The higher use of pronouns in oral texts, whereas nouns are preferred in writing, goes in the same direction of what already observed in larger corpora of adult spoken language (see e.g. (Voghera 2004, 2005) for Italian); indeed, pronouns are among the linguistic devices related to deixis, anaphorical chains and dislocation phenomena typical of spoken language. The different distribution of lexical and grammatical categories also affects lexical density [12], which is a little higher in writing, as typically reported for adults (Halliday 1989).

When we focus on the grammatical structure, it turns out that children tend to produce more complex sentences when they retell the story orally; this is suggested e.g. by the predominance of conjunctions, especially of subordinating ones. Such a distribution, together with that of adverbs [4], can also give some indications on the way modality affects children's language at discourse structure, which appears less cohesive when they write rather than when they retell the story verbally. It is also interesting to note that some well-known factors of syntactic complexity, i.e. the average length of dependency links [14] [7] and the average parse tree depth [15] [8], are not greatly influenced by the way children retell the story. This finding, which also largely stems from the omission or inconsistent use of orthographic marks in writing, seems

[5] http://monitor-it.italianlp.it/

[6] Wilcoxon's signed rank test was applied for statistical analysis because of the small number of subjects.

[7] The dependency length is here calculated in terms of the words occurring in the sentence between the head and the dependent of a syntactic link.

[8] The average parse tree depth is here calculated in terms of the longest path from the root of the dependency tree to the leaf.

to support the claim that "writing is very nearly talk written down" (Kroll 1981) in the early stages of writing development.

In a further stage, the distribution of linguistic features automatically extracted from the corpus was compared to that of the original story, which was also linguistically annotated. Table 6 contains an extract of this analysis. From this comparison, it emerges that children's productions, independently from the modality, are shorter than the original story (feature [1]). This is possibly a consequence of the intrinsic nature of the task of retelling, which requires children to select the salient passages of a story, as well as of the memory biases previously discussed. But more interesting patterns are revealed by the analysis of the distribution of morpho–syntactic categories. At this level e.g. there is a significant difference in the use of adjectives ([3]), which turned out as a peculiarity of children's retellings, whereas it is less attested in the original story (see Section 4.2 for more details).

Table 5
A subset of the monitored linguistic features and their distribution in oral and written retellings. Significant differences at $p < 0.05$ are bolded, those at $p < 0.005$ are also marked with $*$. Note that features from [3] to [11] are percentage distributions while the others are absolute values.

Linguistic Feature	*Oral*	*Written*	*Diff.*
[1] Text length (in token)	125.11	109.46	**+15.64**
[2] Word length	4.54	4.55	**-0.01**$*$
[3] Adjectives	3.03	3.56	-0.53
[4] Adverbs	8.62	4.86	**+3.77**$*$
[5] Coordinating Conj.	6.14	4.83	**+1.31**$*$
[6] Determiners	10.88	14.52	**+3.64**$*$
[7] Nouns	21.80	28.50	**-6.70**$*$
[8] Prepositions	13.45	15.12	**-1.66** $*$
[9] Pronouns	6.70	4.79	**+1.91**$*$
[10] Subordinating Conj.	1.56	0.96	**+0.60**
[11] Verbs	15.51	14.26	**+1.25**$*$
[12] Lexical density	0.539	0.552	**-0.012**
[13] Type/Token ratio (for the first 200 lemmas)	0.514	0.543	**-0.028**
[14] Length of depend. links	2.40	2.42	-0.02
[15] Parse tree depth	6.39	6.71	-0.32

4.2 Analysis of the content

For the analysis of the corpus with respect to the content, we relied on $T2K^2$ (Text–to–Knowledge), a suite of tools based on NLP modules for automatically extracting domain–specific knowledge from a corpus (Dell'Orletta et al. 2014). Following the assumption that the most relevant concepts of a text have a linguistic counterpart, which is typically conveyed by single and multi–word nominal terms, the process of terminology extraction can be seen as the first step to access the knowledge contained in text. We applied the term extraction functionalities of $T2K^2$ both to the original story and to the corpus of children's retellings; the latter was first distinguished into the oral

Table 6
A subset of linguistic features and their distribution in the original story and in the corpus of oral and written retellings. All differences reported in columns three (*Orig. vs. Oral*) and four (*Orig. vs. Written*) are statistically significant at $p < 0.005$.

Linguistic Feature	*Original*	*Diff. (Orig. vs. Oral)*	*Diff. (Orig. vs. Written)*
[1] Text length (in token)	236	+110.90	+126.53
[2] Word length	4.86	-0.32	-0.31
[3] Adjectives	2.54	-0.49	-1.02
[4] Adverbs	4.24	-4.38	+0.62
[5] Coordinating Conj.	4.67	-1.47	-0.16
[6] Determiners	13.14	+2.26	-1.38
[7] Nouns	25.85	+4.05	-2.65
[8] Prepositions	13.14	-0.31	-1.98
[9] Pronouns	5.08	-1.62	+0.29
[10] Subordinating Conj.	1.27	-0.29	+0.31
[11] Verbs	15.25	-0.26	+1.0
[12] Lexical density	0.560	+0.02	+0.008
[13] Type/Token ratio (for the first 200 lemmas)	0.67	+0.16	+0.127
[14] Length of depend. links	2.12	-0.28	-0.30
[15] Parse tree depth	5.32	-1.07	-1.39

and written sub–corpora (each one taken as a whole) and then by considering each school–grade separately for both modalities.

As shown by the excerpt of the output in Table 7, there is a strict correspondence between the ten most salient concepts characterizing the original story and those reported by children, independently from the linguistic modality. Such findings were also replicated when we analyzed separately the oral and written retellings of the 2nd, 3rd and 4thgrade students, thus suggesting that from age seven, children have already mastered the ability to grasp, retain and organize the main concepts of a narrative text like the one here proposed.

However, a more in depth investigation comparing the adjective lemma entries in the original story and in the examined corpus revealed some interesting characteristics typical of children's retellings, which comply with what already observed from linguistic profiling. Specifically, in Table 6 we showed that children used more adjectives with respect to the original story (feature [3]). As shown in (Table 8), although children remembered all the adjectives of the original story both in the oral and the written task, they also introduced new adjectives in their productions. The most evident case is the adjective "triste", which is ranked in the highest positions of the corpus (it was used 28 times in oral and 70 in written texts). Interestingly, a qualitative inspection of the corpus revealed that almost all children chose this adjective as the predicate of a copular verb (example 3) to paraphrase the same meaning conveyed in the original story by a more complex support verb construction (example (4)).

(3) *la faccia della statua divento' triste.*

(4) [...] *il suo sorriso si trasformo' in tristezza.*

Table 7
Excerpt of automatically extracted domain–terminology in the examined corpora.

Original story	Oral corpus	Written Corpus
mappamondo	mano	statua
pietra	statua	mano
terra	mappamondo	mappamondo
mano	rondine	geografo
rondine	geografo	rondine
Geografo	terra	parco
statua	primavera	primavera
busto	nido	terra
parco	ragazzo	nido
primavera	giorno	ragazzo

Table 8
Comparison of the adjectives extracted from the original story and the ten most frequent ones in the examined corpora.

Original story	Oral corpus	Written Corpus
suo	pesante	triste
loro	triste	pesante
contento	felice	felice
scuro	bello	bello
pieno	suo	suo
pesante	loro	loro
bello	scuro	vuoto
vuoto	nuovo	sorridente
	vuoto	geografico
	pieno	piccolo

5. Conclusion

In this paper we presented ISACCO, a new resource for the Italian language containing oral and written retellings of children attending the second, third and fourth grade of the primary school. The particular composition of the corpus would allow scholars to explore issues related to "later language acquisition" (i.e. language development after age five), in particular how children's narrative abilities vary in oral and written language and how writing cognitive demands affect beginning writers' productions.

We also showed the potentiality of an approach based on NLP techniques to inspect child language features, both with respect to linguistic form and content structure, as well as in relation to diachronic and diamesic variations. The obtained results complied with the view that in the early stages of writing development the effect of linguistic modality in a task such as retelling is rather limited and it should be interesting to test if the same holds in other typologies of texts by primary students, e.g. free narratives.

Ongoing work is devoted to enlarge the corpus, also in a longitudinal view, in order to carry out analyses in two different directions: from a more computational linguistics perspective, we would like to evaluate the impact of child language features on standard linguistic annotation tools and to elaborate methods to mitigate this impact; from a theoretical perspective, the study will focus on a deeper linguistic comparison between oral and written language and on a qualitative analysis of linguistic errors also with respect to other existing learner corpora.

Appendix

La statua nel parco, Roberto Piumini, *Mi leggi un'altra storia?*, Einaudi Ragazzi.

> Nel parco di una cittá c'era un monumento ad uno studioso di geografia:
> un busto di pietra che portava sulla mano un mappamondo, anche quello di pietra.
> La faccia del busto sorrideva, guardando il mappamondo.
> Ma una notte, un gruppo di ragazzi pensó di rubare il mappamondo.
> Si arrampicarono sulla statua, spinsero, tirarono e scrollarono,
> finché il mappamondo di pietra rotoló giú. Allora i ragazzi
> se ne andarono e lasciarono il mappamondo sull'erba,
> perché era troppo pesante per portarlo via.
> Il Geografo di pietra rimase a mani vuote e piano piano
> il suo sorriso si trasformó in tristezza.
> Venne la primavera e due rondini portarono fili d'erba, pagliuzze
> e terra e si costruirono il nido tra le mani della statua.
> Le rondini deposero le loro uova e presto nacquero due rondinini.
> che presto incominciarono a volare durante il giorno per tornare
> solo alla sera: il Geografo li aspettava, contento.
> Quando arrivó l'autunno le rondini dovettero partire, ma, prima di lasciare
> la loro casa portarono altra terra e vi lasciarono cadere alcuni semi.
> Durante l'inverno, il Geografo di pietra guardava la sua mano piena
> di terra scura e quando tornó la primavera, dalla terra nacque
> un cespuglio di fiori bellissimi.

Acknowledgments

We would like to thank the headmaster of the primary school "Vasco Morroni" of Ghezzano (Pisa), the teachers and all the children taking part in this research for their contribution.

References

Attardi, Giuseppe. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Prooceedings of EVALITA 2009, Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, 12th December.

Barbagli, Alessia, Piero Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2015. Cita: un corpus di produzioni scritte di apprendenti l'italiano l1 annotato con errori. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it)*, pages 31–35, Trento, Italy.

Beaman, Karen. 1984. Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In Deborah Tannen, editor, *Coherence in spoken and written discourse*, volume 12 of *Advances in Discourse Processes*. Ablex, Norwood, pages 45–80.

Berman, Ruth A. 2004. Between emergence and mastery. The long development route of language acquisition. In Ruth A. Berman, editor, *Language Development across Childhood and Adolescence*, volume 3 of *Trends in Language Acquisition Research*. John Benjamins Publishing Company, Amsterdam, pages 9–34.

Biber, Douglas. 1988. *Variation across speech and writing*. Lund Studies in English. Cambridge University Press, Cambridge.

Brooke, J. and G. Hirst. 2012. Measuring interlanguage: Native language identification with l1–influence metrics. In *Proceedings of the 8th Conference on Language Resources and Evaluation (LREC 2012)*, pages 779–784.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, Ablex, pages 35–53.

Deane, P. and T. Quinlan. 2010. What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2):151–177.

Dell'Orletta, Felice. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, Italy, 12th December.

Dell'Orletta, Felice, Andrea Cimino, Simonetta Montemagni, and Giulia Venturi. 2014. T2k: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2062–2070, Reykjavik, Iceland.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistic profiling of texts across textual genre and readability level. An exploratory study on Italian fictional prose. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, pages 189–197.

Halliday, Michael Alexander Kirkwood. 1989. *Spoken and Written Language*. Lund Studies in English. Oxford: Oxford University Press, Oxford.

Koutsoftas, Anthony D. 2013. School–age language development: Application of the five domains of language across four modalities. In Nina Capone-Singleton and Brian B. Shulman, editors, *Language development: Foundations, processes, and clinical applications*. Jones & Bartlett, Burlington, MA, pages 2215–229.

Kroll, B.M. 1981. Developmental relationship between speaking and writing. In B. M. Kroll and R. J. Vann, editors, *Exploring speaking-writing relationships: Connections and contrasts*. National Council of Teachers of English, Urbana, IL, pages 32–54.

Lintunen, Pekka and Mari Makila. 2014. Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in language*, 12(4):377–399.

Lu, Xiaofei. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Lubetich, Shannon and Kenji Sagae. 2014. Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160, Dublin, Ireland.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 3rd edition of *Lund Studies in English*. Lawrence Erlbaum Associates.

Moneglia, Massimo and Emanuela Cresti. 1997. L'intonazione e i criteri di trascrizione del parlato. In U. Bortolini and E. Pizzuto, editors, *Il progetto CHILDES Italia*, volume II. Il cerro, Pisa, pages 57–90.

Nystrand, Martin. 1987. The role of context in written communication. In Rosalind Horowitz and Jay S. Samuels, editors, *Comprehending Oral and Written Discourse*. Academic Press, San Diego, CA, pages 197–214.

Powers, Willow Roberts. 2005. *Transcription techniques for the spoken word*. Lund Studies in English. Altamira Press, Lanham, Maryland.

Prud'hommeaux, Emily T., Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–96, Portland, Oregon.

Roberts, Celia. 2010. *Qualitative Research Methods and Transcription. Issues in Transcribing Spoken Discourse*. Lund Studies in English. School of Social Science and Social Policy, Kings College London.

Rouhizadeh, Masoud, Richard Sproat, and Jan van Santen. 2015. Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children. In

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, pages 117–123, Denver, Colorado.

Sagae, Kenji, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pages 197–204, Ann Arbor, Michigan.

Scarborough, Hollis S. 1990. Index of productive syntax. *Applied Psycholinguistic*, 11(3):1–22.

Silva, Maria Luisa, Veronica Sańchez, and Abchi Ana Borzone. 2010. Subordinate clauses usage and assessment of syntactic maturity: A comparison of oral and written retellings in beginning writers. *Journal of Writing Research*, 2(1):47–64.

Tolchinsky, Liliana. 2004. The nature and scope of later language development. In Ruth A. Berman, editor, *Language Development across Childhood and Adolescence*, volume 3 of *Trends in Language Acquisition Research*. John Benjamins Publishing Company, Amsterdam, pages 233–248.

Voghera, Miriam. 2004. La distribuzione delle parti del discorso nel parlato e nello scritto. In R. Van Deyck, R. Sornicola, and J. Kabatek, editors, *La variabilité en langue*, volume 1. Gand, pages 261–284.

Voghera, Miriam. 2005. La misura delle categorie sintattiche. In Isabella Chiari and Tullio De Mauro, editors, *Parole e numeri. Analisi quantitative dei fatti di lingua*. Aracne, Roma, pages 125–138.