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

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

Abstract

English. We present ISACCO (Italian school–age children corpus)[1], a new corpus of oral and written retellings of Italian-speaking children attending the primary school. All texts were digitalized and automatically enriched with linguistic information allowing preliminary explorations based on NLP features. Written retellings were also manually annotated with a typology of linguistic errors. The resource is conceived to support research and computational modeling of "later language acquisition", with an emphasis for comparative assessment of oral and written language skills across early school grades.

Italiano. *Presentiamo ISACCO (Italian school–age children corpus), un nuovo corpus di riassunti orali e scritti prodotti da bambini italiani della scuola primaria. Tutti i testi sono stati digitalizzati e arricchiti automaticamente con informazione linguistica per consentire esplorazioni preliminari basate su caratteristiche estratte con strumenti di TAL. I riassunti scritti sono stati anche annotati a mano con una tipologia di errori linguistici. La risorsa è pensata per lo studio e la definizione di modelli computazionali degli stadi più avanzati del processo di acquisizione linguistica, con un'enfasi per la valutazione comparativa delle abilità linguistiche orali e scritte nei primi anni scolastici.*

1 Introduction

The use of naturalistic data to investigate child language features and development has a well-established tradition in L1 acquisition research. 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 research for the earliest stages of acquisition. However, over the last two decades also "later language acquisition" has gained increasing interest (Tolchinsky, 2004), 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). Indeed, under explicit teaching language keeps growing through school–age years in a way that affects all domains and modalities (Koutsoftas, 2013). Regarding the methodological approach to inspect children's data, more attention has been recently paid to text analysis techniques drawn from computational linguistics and Natural Language Processing (NLP). The use of a statistical parser is reported e.g. by Sagae et al. (2005) and Lu (2009) to automate sophisticated measures of syntactic development, reaching performances comparable to those obtained by manual annotation. Computational methods are also employed in diagnostic 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 and Roark, 2011), as well as metrics of semantic similarity (Rouhizadeh et al., 2015). Despite the focus of this paper is on the resource, we will also present preliminary analyses aiming at showing how a NLP perspective applied to a corpus like ISACCO can

[1]The resource will be made publicly available at: http://www.italianlp.it/software–data.

serve as the starting point to conduct computational explorations at multiple levels, which may become particularly useful in view of their applicability to large–scale corpora. It should be possible to test the effect of the diamesic variation on the linguistic complexity of children's texts and to assess changes across schooling levels (cf. section 3.1). The same can be done with respect to the "content", to evaluate whether these variables affect text comprehension and recall. To this aim, the output of an ontology learning system can provide a mean to compare the quantity of 'matched' ideas between the child's retelling and the content of the heard story (cf. section 3.2), so that to identify patterns of typical development to be used for comparison e.g. in clinical settings, with children showing atypical language development.

2 The corpus

2.1 Participants

Fifty-six TD (typically developing) children from the 2nd to the 4thgrade of primary school participated in the task. They were all recruited from a public primary school located in the suburbs of Pisa and examined in the last month of the school year. All children were Italian monolingual speakers, except from two, who were also included in the survey since they had no significant exposure to other languages. Details of the sample group are given in Table 1.

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

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

2.2 Methodology

To collect ISACCO, we inspired to the work of (Silva et al., 2010) for Spanish, who assessed children's oral and written performance in a retelling task by exposing them to the same story to avoid a possible text bias. Differently from them, we excluded the 1stgrade pupils, following the teachers' suggestions pointing 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 intended to be not too challenging for the youngest nor too easy for the oldest group[2]. Children were tested in two sessions, with a gap of two weeks, so that to prevent memory bias. The first session was devoted to collect oral productions; this was done by reading the story aloud once to the whole class and repeating it again to a restricted group of students, which was randomnly chosen by teachers, while their mates carried out another activity related to the story (e.g. drawing a picture). Each selected child was tested individually, in a quiet room, and after hearing the story again was asked to retell it to the experimenter. All retellings were recorded and then transcribed, as detailed in Section 2.3.

	Oral retellings	
Grade	**Number of texts**	**Number of tokens**
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

Table 2: Corpus of oral and written retellings.

In the second session, the same story was read again to the whole class and this time all students produced 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 comparative analysis only the writings of the 56 children tested in the first session were needed, we digitalized all written retellings; such a corpus offers indeed valuable material for research on writing development with a view to its computational modeling.

2.3 Oral data transcription

Children's oral retellings were manually trascribed adding some "natural punctuations" (Powers, 2005) (i.e. periods and commas) according to speech pauses and intonations, to identify major sentence boundaries. These "row" transcripts were then enriched with additional "xml-style" labels to annotate typical phenomena of spoken language (e.g. false starts, disfluencies), as defined in the following tagset:

- tag *fs*: to mark a false start (covering both a single or a sequence of words).

[2]The story is titled "La statua nel parco", by Roberto Piumini.

- 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 the story)

2.4 Linguistic annotation of errors

After being digitalized, written texts were manually annotated with typologies of linguistic errors, following the tagset defined by Barbagli et al. (2015). Errors are distinguished into three macro–areas, according to the domain of linguistic knowledge affected, i.e.: ortography, grammar and lexicon. Each macro–class is further sub-divided into more classes codifying the linguistic category and the target modification for the misused units. Table 3 reports the error tagset and the quantitative distributions for each category according to the school grade.

3 Preliminary explorations of the corpus

This section presents preliminary explorations comparing oral and written retellings with respect to both linguistic structure and content. All analyses were conducted by comparing the statistical distribution of linguistic and lexico–semantic features automatically extracted from the corpora by means of NLP tools. 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, 2006) using Support Vector Machines as learning algorithm. It goes without saying 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 2ndgrade 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.

3.1 First results on linguistic structure

Table 4 shows a subset of linguistic features for which the average difference value between oral and written samples was significant[3]. Starting from superficial features, it emerges that oral retellings are on average longer than the written ones ([1]); in line with previous findings in the literature, such a difference may be due to the heavy cognitive demands initially posed by writing affecting memory and causing a loss of information. 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 [7] and Conjunctions [4,8]) in oral than in written texts. Such a difference affects lexical density [10], which is higher in writing, as typically reported for adults (Halliday, 1989). Coming to the grammatical structure, when children retell the story orally they tend to produce more complex sentences, as suggested by the predominance of conjunctions, especially subordinating ones. Such a distribution, together with that of adverbs [3], 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. Last, it is interesting to note that a well-known factor of syntactic complexity, i.e. the length of dependency links [11], is not significantly influenced by the way children retell the story.

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] Adverbs	8.62	4.86	**-3.77**$*$
[4] Coordinating Conj.	6.14	4.83	**-1.31**$*$
[5] Determiners	10.88	14.52	**+3.64**$*$
[6] Nouns	21.80	28.50	**+6.70**$*$
[7] Pronouns	6.70	4.79	**-1.91**$*$
[8] Subordinating Conj.	1.56	0.96	**-0.96**
[9] Verbs	15.51	14.26	**-1.25**$*$
[10] Lexical density	0.539	0.552	**0.012**
[11] Length of depend. links	2.40	2.42	0.02

Table 4: Linguistic features. Significant differences at $p < 0.05$ are bolded, those at $p < 0.005$ are also marked with $*$.

3.2 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

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

		II grade		III grade		IV grade	
Category	**Target modification**	Freq.%	Abs. Value	Freq.%	Abs. Value	Freq.%	Abs. Value
Orthography							
Consonant doubling	Omission	10.59	(45)	1.40	(3)	5.52	(8)
	Excess	2.35	(10)	1.40	(3)	2.07	(3)
Use of *H*	Omission	0.71	(3)	0.93	(2)	0.00	(0)
	Excess	0.24	(1)	0.00	(0)	0.00	(0)
Monosyllabic words	Mispelling of stressed monosyllabic words	2.35	(10)	6.51	(14)	1.38	(2)
	Mispelling of *po'* (e.g. pó or po)	3.76	(16)	4.65	(10)	4.14	(6)
Apostrophe	Misuse	3.76	(16)	0.93	(2)	0.69	(1)
Other		32.94	(140)	33.02	(71)	40.69	(59)
Grammar							
Verbs	Use of tenses	24.00	(102)	15.35	(33)	12.00	(12)
	Use of modes	0.00	(0)	0.00	(0)	0.69	(1)
	Subject-verb agreement	1.88	(8)	6.51	(14)	5.52	(8)
Prepositions	Misuse	1.88	(8)	3.26	(7)	1.38	(2)
	Omission or Excess	1.41	(6)	1.47	(1)	1.38	(2)
Pronouns	Misuse	0.24	(1)	0.47	(1)	1.38	(2)
	Omission	0.24	(1)	0.47	(1)	1.38	(2)
	Excess	0.240	(1)	0.47	(1)	1.38	(2)
	Misuse of relative pronoun	0.24	(1)	0.47	(1)	0.69	(1)
Conjunctions	Misuse	0.24	(1)	0.47	(1)	2.38	(2)
Other		8.00	(34)	11.63	(25)	10.34	(16)
Lexicon							
Vocabulary	Misuse of terms	4.94	(21)	11.63	(25)	11.03	(16)

Table 3: Linguistic errors tagset and quantitative distributions in written retellings.

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 thus 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 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 5, there is a strict correspondence between the ten most salient concepts characterizing the original story and those reported by children, independently from 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. This analysis, complemented with first data of linguistic profiling, seems to imply that the effect of modality is more relevant at the level of linguistic structure.

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 5: Excerpt of automatically extracted domain–terminology.

4 Conclusion

We presented ISACCO, a new resource for the Italian language containing oral and written retellings of children attending the primary school. We showed the potentiality of NLP-based analyses to inspect child language features, both with respect to linguistic and content structure, as well as in relation to diachronic and diamesic variations. Ongoing work is devoted to enlarge the corpus, also in a longitudinal perspective, to elaborate a qualitative analysis of linguistic errors by also looking comparatively at other learner corpora, and to evaluate the impact of child language features on standard linguistic annotation tools.

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 the survey for their contribution in this research.

References

B. MacWhinney. 2000. The CHILDES Project: Tools for Analyzing Talk. *3rd edition. Lawrence Erlbaum Associates, 2000.*

L. Tolchinksy. 2004. The nature and scope of later language development. In *R.A. Berman (Ed.), Language Development across Childhood and Adolescence. Amsterdam: John Benjamins Publishing Company.*

R. Berman. 2004. Between emergence and mastery: the long developmental route of language acquisition. In *R.A. Berman (Ed.), Language Development across Childhood and Adolescence. Amsterdam: John Benjamins Publishing Company.*

A. D. Koutsoftas. 2013. School–age language development: Application of the five domains of language across four modalities. In *N. Capone-Singleton and B.B. Shulman (Eds.), Language development: Foundations, processes, and clinical applications, Second Edition*, pp. 215–229, Burli, April 2013.

K. Sagae, A. Lavie and B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 197–204.

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

E. T. Prud'hommeaux and B. Roark. 2011. Classification of atypical language in autism. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics.*

M. Rouhizadeh, R. Sproat, J. van Santen. 2015. Similarity Measures for Quantifying Restrictive and Repetitive Behavior in Conversations of Autistic Children. *Computational Linguistics and Clinical Psychology Workshop (CLPsych), NAACL, 2015, Denver, CO.*

M. Silva, V. Sánchez Abchi, A. 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.

W. R. Powers. 2005. Transcription techniques for the spoken word. *Lanham, MD: Altamira Press.*

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi. 2015 (Submitted). CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, New York City, New York:166–170.

M. A.K. Halliday. 1989. Spoken and Written Language. *Oxford: Oxford University Press.*

F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni. 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)*, pp. 2062–2070, 26-31 May, Reykjavik, Iceland.