CItA: an L1 Italian Learner Corpus to Study the Development of Writing Competence

Alessia Barbagli•, Pietro Lucisano•, Felice Dell'Orletta◇, Simonetta Montemagni◇, Giulia Venturi◇
•Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"
alessia.barbagli@gmail.com,pietro.lucisano@uniroma1.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

In this paper, we present the *CItA corpus (Corpus Italiano di Apprendenti L1)*, a collection of essays written by Italian L1 learners collected during the first and second year of lower secondary school. The corpus was built in the framework of an interdisciplinary study jointly carried out by computational linguistics and experimental pedagogists and aimed at tracking the development of written language competence over the years and students' background information.

Keywords: Italian Learner Corpus, Diachronic Evolution of Written Language Competence, Error Annotation

1. Introduction

Over the last ten years, language technologies have been successfully exploited to study the development of language learning processes. A variety of different approaches based on Natural Language Processing (NLP) tools has been developed for different purposes, such as to track the syntactic development in child language (Sagae et al., 2005; Lu, 2007; Lubetich and Sagae, 2014), to measure the developmental language progress using child speech patterns (Sahakian and Snyder, 2012). NLP–based approaches have been devised also to detect mild cognitive impairments using measures of syntactic complexity (Roark et al., 2007) or of semantic and pragmatic atypicality (Rouhizadeh et al., 2013), and to select reading material that are appropriate for students' reading proficiency considered a fundamental component of language competency (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009). As witnessed by the increasing success of the *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* arrived in 2016 at its eleventh edition[1], language technologies have been also exploited in educational settings to design and develop educational applications such as for instance Intelligent Computer–Assisted Language Learning systems (ICALL) (Granger, 2003) or Automatic Essay Scoring systems (Attali and Burstein, 2006).

For all these studies and applications, the availability of electronically accessible corpora of student essays is of pivotal importance. So far, several learner corpora have been built mainly differing at the level of typologies of collected essays (i.e. written or speech transcriptions), goals of analysis (e.g. theoretical studies or development of educational applications), typologies of considered learners (e.g. learners of first or second language, adults or children). A specific interest has been devoted to the construction of written learners' corpora where *errors* (i.e. erroneous forms written by learners) are annotated and classified; this is especially (but not only) the case of corpora of students learning a foreign language (L2). Corpora enriched with this kind of information can offer insight into learners' development of competencies and difficulties (Deane and Quinlan, 2010), they can be used to investigate the characteristics of *interlanguage* (Brooke and Hirst, 2012) or as reference resources for automatic error detection and correction tasks. The latter is the case of the *NUS Corpus of Learner English (NUCLE)* (Dahlmeier et al., 2013) exploited during 2013 and 2014 editions of the "Shared Task on Grammatical Error Correction" (Ng et al., 2013; Ng et al., 2014). Interestingly, corpora of L2 learners annotated for errors have been built for a number of languages other than English, e.g. for Arabic as L2 (Zaghouani et al., 2015), German (Lüdeling et al., 2005), Hungarian (Dickinson and Ledbetter, 2012), Basque (Aldabe et al., 2005), Czech and Italian (Andorno and Rastelli, 2009; Boyd et al., 2014).

In this paper, we would like to narrow the focus on those studies devoted to build corpora of essays written by learners a first language (L1). Among the others, similar corpora have been built for example by Parr (2010) who collected a corpus of essays written by 20,947 New Zealand students in years 4 to 12 of schooling that were manually evaluated considering seven different rubrics (ranging from content organization to features of grammar, spelling, etc.); his study aimed at tracking the relative rate of progress in writing over the years and types of schools (e.g. private vs. public schools, urban vs. rural). Or by McNamara et al. (2010) who collected 120 essays written by U.S. undergraduate students that were manually evaluated to investigate linguistic factors (e.g. syntactic complexity and lexical diversity) related to the level of student writing quality.

If great attention has been payed so far to the construction of corpora of written essays to study English language development of L1 learners, little work has been carried out for other languages. The *KoKo* corpus (Abel et al., 2014) and the corpus collected by Berkling et al. (2014) represent two main exceptions for the German language. The former is a collection of authentic texts (for a total of 716,000 tokens) written by 1,319 German–speaking students attend-

[1] http://www.cs.rochester.edu/∼tetreaul/naacl-bea10.html

ing the last year of secondary school, linguistically annotated using a battery of linguistic annotation tools and manually annotated for background information and errors. It was built to get insight into pupils writing competencies and difficulties. The latter is a corpus of essays collected via elicitation and written by 1,730 students (for a total of 159,111 tokens) from grade 1 to 8 attending elementary and secondary schools. It was manually annotated for a wide range of spelling errors (i.e. orthographic, morpho–syntactic, etc.) to study the different categories of errors years.

In this paper, we introduce *CItA (Corpus Italiano di Apprendenti L1)*, the first freely available and digitalized corpus of essays written by Italian L1 learners collected in the first and second year of the lower secondary school[2]. Notably, it contains not–scored essays, it was manually annotated for errors and corrections, and it is accompanied by a questionnaire containing students' background information. The diachronic nature, the considered school period and the manual annotation for errors and corrections represent the main novelties of the *CItA* corpus. Italian corpora of L1 written productions built so far are characterized by a quite different internal composition. It is worth mentioning here the collection of 5,000 essays written by students from the first to the fifth years of elementary school (1,000 for each school year) collected in all the Italian regions by Marconi et al. (1993) and the corpus built by Borghi (2013), a collection of 2,500 essays (for a total of 276,849 tokens) written by students of the fist year of different high schools in Rome. The latter is a synchronic corpus, the former (although diachronic) does not include essays written by the same group of students over the five years of elementary school.

To our knowledge, *CItA* is the first corpus built to track the development of L1 writing competence of a same group of students over two school years and several students' background information. This makes possible to compare the characteristics of a set of chronologically ordered essays written by the same student over the years. Thus, as discussed in what follows, *CItA* is currently used within an interdisciplinary study jointly carried out by computational linguistics and experimental pedagogists and devoted to investigate how a wide set of linguistic features automatically extracted from the corpus can be related to different aspects of written language development.

2. The *CItA* Corpus

The *CItA corpus (Corpus Italiano di Apprendenti L1)* is a collection of essays written by Italian L1 learners collected during the first and second year of lower secondary school. It was collected during the two school years 20012–2013 and 2013–2014 as part of a broader on–going study carried out in the framework of the IEA–IPS (*Association for the Evaluation of Educational Achievement*) activities (Lucisano, 1988; Lucisano and Benvenuto, 1991). The study is devoted to introduce an innovative NLP–based methodology to track the evolution of written language competence over the first two years of the Italian lower secondary school, a temporal span that is quite crucial in the school career of L1 students (Barbagli et al., 2015).

First year				
	School	Students	Essays	Tokens
Center	A	25	123	39,855
	B	27	143	35,693
	C	24	138	36,441
Suburbs	D	21	58	14,232
	E	19	77	14,988
	F	24	66	17,753
	G	13	64	12,201
Sub–total		153	669	171,163
Second year				
	School	Students	Essays	Tokens
Center	A	25	108	44,338
	B	28	130	47,316
	C	23	117	28,819
Suburbs	D	22	62	19,278
	E	19	64	13,767
	F	24	146	31,897
	G	14	56	12,878
Sub–total		155	683	198,293
Total		**308**	**1,352**	**369,456**

Table 1: *CItA* corpus: internal composition.

CItA contains essays written by the same students chronologically ordered and covering a two–year temporal span. This makes the corpus particularly suitable to track the evolution of L1 written language competence over the time, as suggested by the results of the first experiments carried out by Richter et al. (2015). The underlying hypothesis is that a number of quite relevant transformations in writing competence occurs during the transition from the first to the second year of lower secondary school and that these transformations are mainly due to a different instructional approach to teach writing. The idea is that these transformations can be captured by inspecting how a wide set of linguistic features automatically extracted from text and different typologies of learners' errors are differently distributed in the two considered years.

It should also be noted that none of the already existing Italian corpora of L1 written productions have been annotated for errors. As discussed in what follows, we defined a new annotation schema to mark–up different typologies of errors made by students, together with the corresponding corrections. To our knowledge, this is the first time that an error annotation scheme is designed to annotate errors made by L1 Italian learners. Annotated errors can be used as a further index of the development of written language competence and they make *CItA* suitable for being used in the construction of Automatic Error Correction systems.

The corpus is also accompanied by a questionnaire including 34 questions about biographical, socio–cultural and sociolinguistic background of students. This makes it possible to investigate whether and to which extent some of the student background information are related to the observed written competence changes.

[2]The corpus is freely available for research purposes at http://www.italianlp.it/software-data/

2.1. Corpus Collection

The *CItA* essays were collected in 7 different lower secondary schools located in different areas of Rome: 3 schools are in the historical center and 4 schools in suburbs (see Table 1). The underlying idea is that the city area where the school is located is highly correlated with the socio–cultural context: the historical center is considered representative of a medium–high context while suburbs of a medium–low context. The corpus contains a total of 1,352 essays (369,456 word tokens) written by 153 students the first year and 155 the second year.

Typology of prompt	n° of prompts		
	Center	Suburbs	Total
First year			
Reflexive	25	13	38
Narrative	18	4	22
Descriptive	2	1	3
Expository	–	1	1
Argumentative	2	2	4
Sub–total	47	21	68
Second year			
Reflexive	24	5	29
Narrative	3	6	9
Descriptive	–	–	–
Expository	4	5	9
Argumentative	5	4	9
Sub–total	36	20	56

Table 2: Distribution of typologies of prompts.

The students were asked to respond to different writing prompts that can be grouped into five textual typologies: reflexive, narrative, descriptive, expository and argumentative corresponding to different communicative language abilities and different writing skills. As shown in Table 2, there are some differences over the two considered years and the seven schools. First of all, it can be noted that the number of prompts differs among the seven schools: teachers of the schools located in the city center tend to give a higher numbers of prompts than their colleagues in the suburban schools. Secondly, if reflexive prompts are the most frequent textual type in the two years, from the first to the second year the distribution of narrative prompts are halved while the expository and argumentative ones are doubled. This different distribution follows from the approach to teach writing adopted by teachers: writing a narrative essay is considered simpler, i.e. it requires simpler cognitive and writing skills, than writing an argumentative or expository essays where more complex linguistic and discourse–structuring competences are required. As we will discuss later, this different distribution of prompts is also related to the different distribution of some categories of errors made by students.

A prompt common to all schools was also assigned at the end of the first and second year. At the end of second year, students were asked to respond to the Italian version of Task 9 of the IEA–IPS (Lucisano, 1984; Corda Costa and Visalberghi, 1995) study, i.e. a letter of advice to a younger fellow student on how one should write in order to get good grades in the school; and at the end of the first year a modified version of Task 9. The two common prompts were meant to provide evidence of how students perceive the different writing instructions received in the two considered school years. First investigations in this directions were carried out by Barbagli et al. (2015) combining automatic linguistic annotation tools and knowledge extraction techniques. It resulted that in the first year students tend to mostly provide emotive advises expressed by terms such as e.g. *non aver paura 'not to have fear', paura dei compiti 'fear of the essay', rifletti prima di scrivere 'reflect before writing'*; while in the second year, their advises refer more to meta–linguistic traits, such as e.g. the use of calligraphy, the use of verbs, the adherence to the prompt, thus reflecting the different typology of writing instructions that they received.

The students were also asked to answer to a questionnaire that we designed and that includes 34 questions about their biographical, socio–cultural and sociolinguistic background. We considered two main types of questions: a first group of thirteen concern biographical information such as the language(s) the students usually speak at home, when and where they were born, their parents' education and employment, etc.; the other questions are meant to investigate how students perceive the writing activity in general and particularly the school writing, if they like writing outside school, which kind of texts they prefer writing, etc.

Interestingly enough, the distribution of the answers to the first set of questions is in line with our starting hypothesis that the city area where the school is located is highly correlated with the socio–cultural context. As shown in Tables 3 and 4, it resulted that the schools located in the historical center are mostly attended by students who at home usually speak Italian or Italian and a foreign language, and whose parents are employed in highly ranked jobs; while students attending schools in suburbs belong to a different socio–cultural context where dialects and foreign languages are more frequently spoken, and where low ranked jobs (i.e. artisan and workman jobs) are the main typology of employment. As far as the attitude towards writing is concerned, the majority of students claims that teaching writing is "very important" (78,9%) even though students attending the schools in the city center (88,7%) believe that it is more important than those attending a school in suburbs (69%). Interestingly, all students agree that writing is mostly useful to "find a job" than to "put in order ideas" , and they prefer writing essays that require few discourse–structuring competences and allow conveying emotions and feelings.

2.2. Error Annotation

The *CItA* corpus was manually annotated for different typologies of errors by a lower secondary school teacher. She also hand–corrected the errors made by students. Error annotation is a quite challenging task since it assumes that a deviation from a linguistic norm is occurring, a norm which is in its turn an arbitrary concept defined only according to social conventions. Besides, an L1 error taxonomy applicable in corpus annotation is lacking for the Italian language. This is the reason why we defined a new annotation schema starting from Berruto (1997)'s definition of

Spoken language	Center	Suburbs	Total
Italian	66 (48)	46 (33)	56 (81)
Italian and dialect	7 (5)	30 (21)	18 (26)
Dialect	–	3 (2)	1 (2)
Italian and foreign language	26 (19)	15 (11)	21 (30)
Foreign language	1 (1)	6 (4)	3 (5)
Total	100 (73)	100 (71)	100 (144)

Table 3: Percentage distribution (and number of occurrences) of student answers to the question: "Which language do you usually speak at home?"

		Center	Suburbs	Total
Mother's employment	High	54.9 (39)	6.1 (4)	31.4 (43)
	Medium	26.8 (19)	25.8 (17)	26.3 (36)
	Low	18.3 (13)	68.2 (45)	42.3 (58)
	Total	100 (71)	100 (66)	100 (137)
Father's employment	High	47.3 (35)	2.9 (2)	25.7 (37)
	Medium	36.5 (27)	21.4 (15)	29.2 (42)
	Low	16.2 (12)	75.7 (53)	45.1 (65)
	Total	100 (74)	100 (70)	100 (144)

Table 4: Percentage distribution (and number of occurrences) of student answers to the question: "Which is your mother's and father's employment?"

"neo–standard Italian" as linguistic norm, according to the literature on evaluation of written skills of L1 Italian learners (Corda Costa and Visalberghi, 1995; De Mauro, 1983; Emilia-Romagna, 2010; Colombo, 2011) and checking the frequency distribution of errors in *CItA*. To our knowledge, this is the first time that an error annotation scheme is designed to annotate errors made by L1 Italian learners.

Table 5 reports the typology of errors considered in the error annotation schema we defined as well as some statistical distributions. We designed a three–level schema including: the **macro–class of error**, i.e. grammatical, orthographic and lexical; the **class of error**, i.e. the linguistic element involved (e.g. verbs, prepositions, monosyllables); and the corresponding **type of modification** required to correct the error (e.g. the misuse of verb with respect to the use of verbal tense). We chose to consider these three macro–classes of errors since they correspond to the main areas of linguistic skills required by the report "Rilevazione degli errori più diffusi nella padronanza della lingua italiana nella prima prova di italiano"[3] issued by the INVALSI national institute[4] and the Accademia della Crusca in 2012. This three–layered schema is also in line with the one defined by Granger (2003) for the annotation of errors made by second language learners.

According to the annotation format defined by Ng et al. (2013) for the "Shared Task on Grammatical Error Correction", *CItA* is annotated as follows:

> [...] dopo aver fatto le squadre <M t="11" c="abbiamo">avevamo</M> subito iniziato a giocare [...] (*once we splitted into teams we have suddenly started playing*)

where the textual span of error is marked by *<M>* and *</M>* (*Mistake*), the attribute *t* (*type*) is the macro–class and class of error (in this is case the error is a grammatical errror and it refers to a misuse of verbal tense), and *c* (*correction*) reports the corrected form. Examples of annotation are reported in Table 6.

Inspecting the statistical distribution reported in Table 5, it can be noted that in both years (Column *Total %*) orthographic and grammatical errors are the most frequent ones (46.55% and 47.33% respectively) while the lexical errors are far less (about 6%). In particular, the most frequent errors are the orthographic not–classified (*Other*) ones (22.32%) followed by the erroneous use of verb tenses (11.26%), the grammatical not–classified errors (6.37%) and the erroneous use of prepositions (6.6%). Interestingly enough, the majority of errors (the ones bolded in Table 5) has a statistically significant variation over the two years thus showing that several common trends in the development of writing competence occur during the transition from the first to the second year.

As far as the frequency distribution (Column *Freq.%*) and the average occurrence (Column *Avg.*) per year is concerned, the most frequent errors are the orthographic and grammatical not–classified ones, the erroneous use of verbs, prepositions, articles, pronouns and the redundant use of double consonants. More in particular, the total number of errors decreases over years even if this is not the case for all the classes of errors. The most interesting exception is represented by the erroneous use of verbs and, in particular, by the misuse of verbal tense that increases. This may be due to the different typology of prompts given by teachers. As reported above, in the first year students were mostly asked to respond to narrative prompts that require quite simple linguistic abilities including the use of 'simple' verb moods and tenses to express temporal sequences; while, in the second year students have to write more argumentative essays where more complex linguistic and discourse–structuring competences are required. This can suggest that students in the transition from the first to the second year are requested to use more complex verb forms thus making more errors.

Interestingly, the statistical distribution of some typologies of errors is correlated with the student background information we collected. This is the case, for example, of the distribution of lexical errors that correlates with the attitude towards reading: the students who claim to read "frequently" make less errors of this type over the two considered years. And, this is also the case e.g. of the grammatical errors that vary significantly with respect to the city area where the schools are located, as Table 7 shows. The average occurrence of this type of errors decreases over the two years in all the schools located the center of Rome and in two of those in suburbs; while, in two of the suburban schools they increase. Surprisingly, the highest number of grammatical errors (on average) is made in a school of the center even though in this school the difference over the years is doubled with respect the other schools. If we compare

[3] http://www.invalsi.it/download/rapporti/es2_0312/RAPPORTO_ITALIANO_prove_2010.pdf

[4] http://www.invalsi.it/invalsi/index.php

		I year			II year			Total %
Class of Error	**Type of Modification**	Freq.%	Avg	SD	Freq.%	Avg	SD	
Grammar								
Verbs	**Use of tense**	7.78 (150)	0.99	2.29	15.67 (239)	1.47	4.05	11.26 (389)
	Use of mood	4.25 (82)	0.54	1.39	4.92 (75)	0.49	0.99	4.55 (157)
	Subject-Verb agreement	2.85 (55)	0.37	1.38	4 (61)	0.41	1.27	3.36 (116)
Prepositions	**Erroneous use**	6.48 (125)	0.83	2.58	6.75 (103)	0.66	1.21	6.6 (228)
	Omission or redundancy	1.03 (20)	0.13	0.40	0.72 (11)	0.07	0.25	0.90 (31)
Pronouns	Erroneous use	5.09 (98)	0.65	1.13	3.54 (54)	0.36	0.97	4.4 (152)
	Omission	0.41 (8)	0.05	0.36	0.59 (9)	0.06	0.39	0.49 (17)
	Redundancy	2.70 (52)	0.35	0.61	1.57 (24)	0.16	0.46	2.2 (76)
	Erroneous use of relative pronoun	2.13 (41)	0.27	0.70	1.70 (26)	0.17	0.44	1.94 (67)
Articles	**Erroneous use**	5.81 (112)	0.75	3.72	3.54 (54)	0.35	1.09	4.81 (166)
Conjunctions and/or connectives	Erroneous use	0.57 (11)	0.07	0.33	0.52 (8)	0.05	0.23	0.55 (19)
Other		7.31 (141)	0.94	3.66	5.18 (79)	0.49	1.79	6.37 (220)
Orthography								
Double consonants	**Omission**	6.74 (130)	0.83	2.49	5.05 (77)	0.48	1.56	5.99 (207)
	Redundancy	3.27 (63)	0.42	0.89	3.67 (56)	0.37	1.13	3.45 (119)
Use of *h*	**Omission**	3.21 (62)	0.39	1.03	1.64 (25)	0.17	0.62	2.52 (87)
	Redundancy	1.66 (32)	0.21	0.53	1.11 (17)	0.10	0.34	1.42 (49)
Monosyllables	**Erroneous use of stressed monosyllabic words**	4.87 (94)	0.63	1.07	4.07 (62)	0.40	0.83	4.52 (156)
	Use of *po* or *pò* instead of *po'*	1.66 (32)	0.21	0.72	1.64 (25)	0.17	0.52	1.65 (57)
Apostrophe	**Erroneous use**	4.82 (93)	0.61	1.01	4.52 (69)	0.46	0.89	4.69 (162)
Other		21.77 (420)	2.76	4.58	23.02 (351)	2.27	4.60	22.32 (771)
Lexicon								
Vocabulary	**Erroneous use**	5.60 (108)	0.70	1.64	6.56 (100)	0.66	1.09	6.02 (208)
Total number of errors		1929			1525			

Table 5: Error annotation schema. For each year: frequency distribution and number of occurrences (*Freq.%*), average occurrence per year (*Avg*), Standard Deviation (*SD*). The column *Tot. %* reports the percentage and the number of occurrences of errors in the two years. Errors varying significantly over the two years (i.e. $p < 0.05$) are bolded.

the average occurrences of these errors made by students born in Italy and abroad, we can claim that students born abroad make more errors than their mates in both the first and second year (see Table 8). However, the difference over the years varies importantly: during the transition from the first to the second year the students born abroad make significantly less errors; on the contrary, the number of errors made by students born in Italy increases a little bit. This demonstrates two different speed of development: students born abroad start from a lower level of grammatical competence but they improve faster their skills.

On the contrary, orthographic errors do not vary significantly with respect to any background information. This provides evidence of linguistic studies claiming that language competence is not related with the orthographic correctness: orthographic skills are learned only over a longer time span (Colombo, 2011; Ferreri, 1971; Lavino, 1975; De Mauro, 1977).

	School	I year	II year	Difference
Center	A	2.6	0.9	1.7
	B	5.2	3.1	2.1
	C	15.1	9.3	5.8
Suburbs	D	3.5	8.2	-4.8
	E	6.4	4.6	1.9
	F	5.4	4.6	0.8
	G	1.5	2.8	-1.3

Table 7: Average occurrence of grammatical errors per year and with respect to the city areas.

"Are you born in Italy or abroad?"	**I year**	**II year**	**Diff.**
Yes	3.98	4.18	-0.2
No	23.19	11.06	12.13

Table 8: Average occurrence of grammatical errors per year and with respect to the question: "Are you born in Italy or abroad?".

3. *CItA* for ...

As discussed in previous paragraphs of this paper, *CItA* is meant for a number of different research purposes. Firstly, the corpus was meant to study the development of writing language competence of Italian L1 learners over the time. As introduced by Barbagli et al. (2015), it is currently used in an interdisciplinary study that combines computational linguistics and experimental pedagogy approaches. The study stems from the intuition that linguistic features

Class of Error	Type of Modification	Example
Verbs	Use of tense	[...] dopo aver fatto le squadre <M t="11" c="abbiamo">avevamo</M> subito iniziato a giocare
	Use of mood	[...] il pensiero che mi tormentava di più era che tra poco si <M t="12" c="sarebbe fatto">faceva</M> il campo scuola.
	Subject-Verb agreement	[...] la mia famiglia ed io <M t="13" c="stavamo">stavo</M> al mare a Torvajanica
Prepositions	Erroneous use	<M t="14" c="in">a</M> Romania sono andata <M t="14" c="in">a</M> agosto
Pronouns	Erroneous use	Proteggere i più deboli è molto coraggioso da parte di chi <M t="16" c="li">lo</M> protegge
	Redundancy	Alla nostra maestra <M t="18" c="canc">gli</M> piaceva tanto la storia
	Erroneous use of relative pronoun	La scienza non so perché mi fa pensare a un fenomeno costruito su un'altura <M t="19" c="per cui">che</M> ci vuole molto ingegno.
Articles	Erroneous use	<M t="111" c="gli">i</M> dei, sapendo che qualcuno aveva preso senza merito il sacro vaso della Giustizia, si rattristarono molto, [...]
Use of *h*	Omission	<M t="23" c="ho">o</M> visto uno spettacolo bellissimo con i raggi laser
Lexicon	Erroneous use	C'era molta ombra nel giardino e io mi ci <M t="31" c="addormentavo">addormivo</M> sempre.

Table 6: Examples of errors annotated in *CItA*.

Raw text features
Average sentence and word length
Lexical features
Percentage of words belonging to the *Basic Italian Vocabulary* (De Mauro, 2000)
Internal distribution into the usage classification classes of 'fundamental', 'high usage', 'high availability' words
Type/Token Ratio (TTR) of the first 100 and 200 tokens
Morpho–syntactic features
Distribution of Part–Of–Speech
Lexical density
Distribution of verbs with respect to their mood, tense and person
Syntactic features
Distribution of dependency types
Verbal predicates features (i.e. arity of verbal predicates, percentage of verbal predicates with elliptical subject)
Parse tree depth features (i.e. depth of the whole parse tree, average length of dependency links)
Subordination features (i.e. distribution of subordinate vs main clauses, relative ordering of subordinates with respect to the main clause, average depth of 'chains' of embedded subordinate clauses)
Nominal modification features (i.e. average depth of embedded complement 'chains' governed by a nominal head)
Relative ordering of subject and object with respect to the main verbal predicates

Table 9: Linguistic features automatically extracted from the *CItA* corpus.

of text quality change over time according to the development of student writing skills and that these features can be identified by relying on the automatically annotated student essays.

In order to test this hypothesis, *CItA* was morpho-syntactically tagged by the POS tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi et al., 2009). The linguistically annotated corpus is further inspected using MONITOR–IT[5], a tool able to carry out the linguistic profiling of texts following the methodology devised by Dell'Orletta et al. (2013) that relies on the wide set of linguistic features reported in Table 9 and extracted on the basis of the different levels of automatic linguistic analysis, i.e. tokenization, lemmatization, morpho–syntactic tagging and dependency parsing.

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

Feature	I year	II year	Significance
Conjunctions	6.81	6.27	0.00
Pronouns	9.31	8.38	0.00
Clitic pronouns	4.79	4.32	0.00
Personal pronouns	1.70	1.27	0.00
Preposition	10.68	11.37	0.00
Nouns	19.92	20.66	0.01

Table 10: % distribution of morpho–syntactic features varying significantly over the school years (significance: $p < 0.05$).

Table 10 reports an excerpt of the results of the statistical distributions of some morpho–syntactic features. It can be noted that the essays written in the second year contain a lower percentage of conjunctions, pronouns, clitic and per-

sonal pronouns, and a higher percentage of prepositions and nouns with respect to the essays of the first year. These statistically significant differences suggest that in the second year students learned to write possibly exploiting the *pro–drop* potentiality of the Italian language. According to the literature on register variation (Biber, 1993), they write more *informative* essays, i.e. characterized by more prepositions and nouns. Note that this can also be influenced by the type of prompt the students are asked to write in the second year, i.e. descriptive and expository (see Table 2).

CItA is also currently used to investigate whether the linguistic features of student essays are significantly related to the students' background information. Table 11 reports an example of this investigation showing how the lemma occurring in the essays written by students attending schools in the historical center and in suburbs are differently distributed over the years with respect to the De Mauro's usage classification classes. It can be noted that in the first year students attending the suburban schools use a higher percentage of 'fundamental words' (i.e. very frequent and simple words) with respect to their peers attending the schools in the historical center, even if this variation resulted to be not statistically significant. On the contrary, it is significant that in the second year they use a lower percentage of this class of words and a higher percentage of 'high availability' words (i.e. relatively lower frequency words referring to everyday life). This suggests that they learned to write more complex words.

The corpus can be also used to develop innovative NLP approaches for use in educational applications that can be used for example in MOOCs (Massive Open Online Courses) such as e.g. automatic error detection and correction systems, automatic essay scoring tools, intelligent tutoring systems or also tools for assisting teachers and test developers.

4. Bibliographical References

Abel, A., Glaznieks, A., Nicolas, L., and Stemle, E. (2014). Koko: an l1 learner corpus for german. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 26–31.

Aldabe, I., Amoros, L., Arrieta, B., de Ilarraza, A. D., Maritxalar, M., Oronoz, M., and Uria, L. (2005). Learner and error corpora based computational systems. In *Proceedings of the PALC 2005 Conference*.

Andorno, C. and Rastelli, S. (2009). *Corpora di Italiano L2: tecnologie, metodi, spunti teorici*. Guerra Edizioni.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e–rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3):465–480.

Attardi, G., Dell'Orletta, F., Simi, M., and Turian, J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2015). Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati. *Italian Journal of Computational Linguistics (IJCoL)*, 1(1):99–117.

Berkling, K., Fay, J., Ghayoomi, M., Hein, K., Lavalley, R., Linhuber, L., and Stüker, S. (2014). A database of freely written texts of german school students for the purpose of automatic spelling error classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1212–1217.

Berruto, G. (1997). *Sociolinguistica dell'italiano contemporaneo*. Carocci, Roma.

Biber, D. (1993). Using register–diversified corpora for general language studies. *Computational Linguistics Journal*, 19(2):219–241.

Borghi, C. C. (2013). *Analisi di produzioni scritte. Valutazioni e misure automatizzate di elaborati scolastici.* Tesi di dottorato in pedagogia sperimentale, Università di Roma, La Sapienza.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. (2014). The merlin corpus: Learner language and the cefr. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Brooke, J. and Hirst, G. (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.

Colombo, A. (2011). *"A me mi" Dubbi, errori, correzioni nell'italiano scritto*. Franco Angeli editore.

Corda Costa, M. and Visalberghi, A. (1995). *Misurare e valutare le competenze linguistiche. Guida scientifico-pratica per gli insegnanti*. La Nuova Italia, Firenze.

Dahlmeier, D., Ng, H., and Wu, S. (2013). Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

De Mauro, T. (1977). *Scuola e linguaggio*. Editori Riuniti, Roma.

De Mauro, T. (1983). Per una nuova alfabetizzazione. In S. Gensini et al., editors, *Teoria e pratica del glotto-kit. Una carta d'identità per l'educazione linguistica*. Franco angeli, milano edition.

De Mauro, T. (2000). *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

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

Dell'Orletta, F., Montemagni, S., and Venturi, G. (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.

Dell'Orletta, F. (2009). Ensemble system for part-of-speech taggin. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*.

Dickinson, M. and Ledbetter, S. (2012). Annotating errors in a hungarian learner corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

		I year			II year		
		'fundamental'	'high usage'	'high availability'	'fundamental'	'high usage'	'high availability'
Center	Average	84.44	11.04	4.52	84.96	10.78	4.27
	Stand Dev.	1.69	1.62	0.88	2.02	1.95	0.84
Suburbs	Average.	84.61	10.35	5.04	83.73	11.25	5.02
	Stand Dev.	2.15	1.88	1.16	2.54	2.07	1.20
	Significance	0.60	0.02	0.00	0.00	0.16	0.00

Table 11: % distribution of lemma in the essays of the historical center and suburban schools with respect to the De Mauro's usage classification classes (significance: $p < 0.05$).

Emilia-Romagna, G. (2010). La correzione dei testi scritti. In E. Lugarini, editor, *Valutare le competenze linguistiche*, pages 188–203. Franco angeli, milano edition.

Ferreri, S. (1971). Italiano standard, italiano regionale e dialetto in una scuola media di palermo. In M. Medici et al., editors, *L'insegnamento dell'italiano in Italia e all'estero*, pages 205–224. Roma, bulzoni edition.

Granger, S. (2003). Error–tagged learner corpora and call: A promising synergy. *CALICO Journal*, 20:465–480.

Lavino, C. (1975). *L'insegnamento dell'italiano. Un'inchiesta campione in una scuola media sarda*. Edes, Cagliari.

Lu, X. (2007). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Lubetich, S. and Sagae, K. (2014). Data–driven measurement of child language development with simple syntactic templates. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 2151–2160.

Lucisano, P. and Benvenuto, G. (1991). Insegnare a scrivere: dalla parte degli insegnanti. *Ricerca educativa*, 6:265–279.

Lucisano, P. (1984). L'indagine iea sulla produzione scritta. *Ricerca educativa*, 5:41–61.

Lucisano, P. (1988). La ricerca iea sulla produzione scritta. *Ricerca educativa*, 2(3):3–13.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi–level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., and Tavella, M. (1993). *Lessico elementare: dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli, Bologna.

McNamara, D., Crossley, S., and McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Ng, H., Wu, S., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

Ng, H., Wu, S., Briscoe, T., Hadiwinoto, C., Susanto, R., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Parr, J. (2010). A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research*, 2(2):129–150.

Petersen, S. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:89–106.

Richter, S., Cimino, A., Dell'Orletta, F., and Venturi, G. (2015). Tracking the evolution of language competence: an nlp–based approach. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it)*, pages 236–240.

Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 1–8.

Rouhizadeh, M., Prud'hommeaux, E., Roark, B., and van Santen, J. (2013). Distributional semantic models for the evaluation of disordered language. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 709–714.

Sagae, K., Lavie, A., and MacWhinney, B. (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.

Sahakian, S. and Snyder, B. (2012). Automatically learning measures of child language development. In *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 95–99.

Schwarm, S. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pages 523–530.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native arabic texts: Guidelines and corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 129–139.