Linguistic Profile of a Text and Human Ratings of Writing Quality: a Case Study on Italian L1 Learner Essays

Aldo Cerulli*
Università di Pisa

Dominique Brunato**
Istituto di Linguistica Computazionale
"Antonio Zampolli" (CNR–ILC)
ItaliaNLP Lab

Felice Dell'Orletta†
Istituto di Linguistica Computazionale
"Antonio Zampolli" (CNR–ILC)
ItaliaNLP Lab

This paper presents a study based on the linguistic profiling methodology to explore the relationship between the linguistic structure of a text and how it is perceived in terms of writing quality by humans. The approach is tested on a selection of Italian L1 learners essays, which were taken from a larger longitudinal corpus of essays written by Italian L1 students enrolled in the first and second year of lower secondary school. Human ratings of writing quality by Italian native speakers were collected through a crowdsourcing task, in which annotators were asked to read pairs of essays and rated which one they believed to be better written. By analyzing these ratings, the study identifies a variety of linguistic phenomena spanning across distinct levels of linguistic description that distinguish the essays considered as 'winners' and evaluates the impact of students' errors on the human perception of writing quality.

1. Introduction

With the effect of the global COVID-19 pandemic, the phenomenon of distance learning has become more prevalent showing the importance of endowing teachers and students with advanced language technologies able to support the practice of teaching and learning in online environments. With respect to language learning and teaching, many of the opportunities and challenges that are associated with these new learning paradigms have been tackled by Intelligent Computer-Assisted Language Learning (ICALL), an interdisciplinary research field that aims at integrating insights from computational linguistics and artificial intelligence into computer-aided language learning. Within the last twenty years, this field has experienced a considerable growth especially in the area of assessment thanks to the development of Automated Essay Scoring (AES) systems (Attali

* Dipartimento di Filologia Letteratura Linguistica Piazza Torricelli, 2, 56126, Pisa.
E-mail: a.cerulli1@studenti.unipi.it
** ItaliaNLP Lab (www.italianlp.it), CNR–ILC - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: dominique.brunato@ilc.cnr.it
† ItaliaNLP Lab (www.italianlp.it), CNR–ILC - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: felice.dellorletta@ilc.cnr.it

and Burstein 2006; Rudner, Garcia, and Welch 2006; Landauer, Laham, and Foltz 2003; McNamara, Crossley, and Roscoe 2013), i.e. computer-based assessment tools able to automatically score or grade the student's responses by considering appropriate features derived from a training set of annotated responses, or tools for automatic error detection and correction (Ng et al. 2013), which are able to automatically identify linguistic errors of different types in text essays in order to suggest adequate correction but also to provide individualized feedback to learners on exercises and to automatically create and use detailed learner models.

A fundamental requirement for developing such a kind of educational applications is the availability of electronically accessible corpora of authentic learners' productions. Corpora created so far differ in many respects. For instance, considering the types of examined learners, they can gather productions written by second language (L2) students or by native speakers: the former have been built for many languages (e.g. English, Arabic, German, Hungarian, Basque, Czech, Italian), while the latter are mainly available for English.

A further dimension of variation concerns the data collection method. The majority of existing corpora are cross-sectional while very few ones are longitudinal. In the context of Italian as first language (L1) – which is the focus of our contribution –, for the first typology, it is worth mentioning the synchronic corpus of 2,500 compositions written by students of the first year of several high schools in Rome (Borghi 2013), as well as the diachronic one composed by 5,000 productions written by pupils during the five years of elementary school all over Italy (Marconi et al. 1994). The only available longitudinal is represented by CItA (*Corpus Italiano di Apprendenti L1*), which was jointly developed by the Institute for Computational Linguistics of the Italian National Research Council (CNR) of Pisa and the Department of Social and Developmental Psychology at Sapienza University of Rome (Barbagli et al. 2016): it is the first digitalized collection of essays written by the same group of Italian L1 learners in the first two years of the lower secondary school[1].

CItA contains essays written by the same students chronologically ordered and covering a two–year temporal span. Its diachronic and longitudinal nature makes the corpus particularly suitable to study the evolution of L1 learners' writing competence over the two years, assuming that many remarkable changes in writing skills occur in this period. For instance, in their recent work, Miaschi, Brunato, and Dell'Orletta (2021) showed that it is possible to automatically learn the writing development curve of students: they extracted a wide set of linguistic features from the essays and used them to train a binary classification algorithm able to predict the chronological order of two productions written by the same pupil at different times. The present study ranks among previous research based on CItA, but chooses a different approach from the one just mentioned: instead of automatically tracking the development of students' writing competence, we focused here on assessing the perception of writing quality by Italian L1 speakers with the aim of understanding whether it is possible to find a set of linguistic features that are crucially involved in distinguishing 'good' and 'bad' essays according to the evaluation of our target readership.

Contributions. To the best of our knowledge, this is the first study that (i) introduces an Italian dataset of learner essays evaluated in terms of perceived quality by means of a

1 The corpus is freely available for research purposes at the following link: `http://www.italianlp.it/resources/cita-corpus-italiano-di-apprendenti-l1/`

crowdsourcing task, (ii) investigates the contribution of a wide set of linguistic features covering lexical, morpho-syntactic and syntactic phenomena – that altogether define the linguistic profile of a text – in modeling the individual perception of writing quality and (iii) assesses the impact of students' errors covering different domains on human judgments.

In what follows we first discuss some related works in the literature that have approached the problem of modeling writing quality according to human evaluation using NLP techniques (Section 2). We then present our starting corpus, i.e. CItA (*Corpus Italiano di Apprendenti L1*), and discuss the theoretical and methodological framework that informed its construction (Section 3). Section 4 focuses on the approach we adopted to set up the crowdsourcing task for collecting human judgements of perceived writing quality on a selection of CItA essays. Then, we present the results of our analysis along two main lines: the first one aimed at characterizing the linguistic profile of essays which were on average perceived as well-written (Section 5); the second one focused on understanding whether and to which extent linguistic errors play a role in native speakers' perception of writing quality (Section 6). In the conclusions, we discuss some relevant applications that this study would enable and propose further improvements in several directions.

2. Related work

As reported by Crossley and McNamara (2011), progresses in disciplines such as computational linguistics, discourse processing and information retrieval paved the way for computational investigations into the textual features that impact on human judgments of essay quality. According to Crossley et al. (2014), the most common approach to assessing writing proficiency is to identify relationships between linguistic 'microfeatures' extracted from a text – covering aspects such as length, complexity, cohesion, relevance, topic, and rhetorical style – and the scores attributed to it by expert human raters. A first insightful contribution towards a better understanding of this relationship was provided by the above-mentioned study by Crossley and McNamara (2011). It was aimed at investigating the role of human perception of coherence in predicting the overall judgments of essay quality by modelling raters' coherence judgments through several computational indices, which were calculated using the Coh-Metrix tool[2] (McNamara et al. 2014). The particular focus on coherence was motivated by previous studies (McNamara, Crossley, and Mccarthy 2010; Crossley and McNamara 2012) showing that human ratings of text coherence were the most informative predictors of the holistic judgments of writing quality, while no evident relation between cohesion cues and essay quality emerged. The analyses were conducted on a corpus of 135 argumentative essays written by as many college freshmen attending either 'Composition One' or 'Composition Two' course at Mississippi State University (MSU). Every student was randomly assigned one among two selected SAT (*Scholastic Assessment Test*) prompts to be responded in 25 minutes. Each essay was read and scored by at least two among eight trained composition professors according to both an analytic rubric – whose creation involved the collaboration of experts in composition studies, cognitive scientist and specialized raters – and a holistic one. The choice of first-year students is based on the assumption that learning how to competently convey messages in written texts is a crucial skill for academic and professional success. This makes the understanding of

2 http://cohmetrix.com/

writing and, in particular, the difference between good and poor writing an important objective both for theoretical and applied purposes.

Further analyses on a similar corpus, described by McNamara, Crossley, and Roscoe (2013), led to the development of the Writing Pal[3], an intelligent tutoring system (ITS) designed to assist high school and college students in the acquisition and improvement of writing skills. It provides lessons dealing with the most effective strategies to perform the various phases of writing – i.e., generating and organizing ideas, drafting and revising an essay – in addition to an area where students can put the learned concepts into practice by writing prompt-based essays. The system automatically scores them and returns a (hopefully) meaningful, formative feedback reporting suggestions to improve the structural and rhetorical quality of the essay. For instance, students are taught to write conclusions that succinctly summarize the main arguments without presenting additional or new information. Since students' responses are open-ended and potentially ambiguous, the performances of such systems in producing a valid feedback depend on the sophistication level of the NLP algorithms that process and interpret the input.

As regards L2 written proficiency, it is worth mentioning the investigation by Crossley et al. (2014) on the potential for many computational indices calculated by two automated text analysis tools, the aforementioned Coh-Metrix and the Writing Assessment Tool (WAT), to predict human scores of essay quality. The analyses were carried out on a corpus of 480 texts collected from two administrations of the TOEFL-iBT (Test of English as a Foreign Language Internet-Based Test) on two groups of 240 candidates, pertaining to a variety of home countries and linguistic backgrounds. Each production was firstly assessed by two expert raters trained by the Educational Testing Service (ETS) according to a standardized TOEFL independent writing rubric. Then, it was associated with an overall score, corresponding to the average of the two grades if their difference was smaller than two points; otherwise, a third expert evaluated it and the final score was the average of the two closest ones. By following this approach, the authors of the study could discriminate between higher and lower quality essays. The distinction led them to identification of the linguistic microfeatures that correlate with L2 essay quality and the training of a regression model to automatically score TOEFL essays according to the same dimensions. They finally evaluated the model strengths and weaknesses. Overall, the contribution represents a significant effort towards the modeling of L2 writing quality by means of textual microfeatures.

While sharing the purpose of modeling the human perception of writing quality from learner texts, our work differs in many respect: from the language and authors' characteristics of the analysed essays, to the approaches adopted for gathering human judgments of writing quality and studying how they relate to the features characterising the linguistic structure of text.

3. The CItA Corpus

As previously mentioned, our study is based on CItA, a longitudinal corpus of essays written by the same L1 learners in the first two years of lower secondary school and chronologically ordered. It was collected during the two school years 2012-2013 and 2013-2014 as part of a broader 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 two-year period was chosen based on the hypothesis that native

3 http://www.adaptiveliteracy.com/writing-pal

speakers' writing competence changes significantly in the transition from the first to the second year of middle school, as a consequence of a more formal approach to writing adopted by teachers. According to Barbagli et al. (2016), these transformations can emerge by inspecting the differences over the considered time frame in the distribution of a wide range of linguistic features automatically extracted from the texts.

CItA creators also supposed that the evolution of writing skills could be related to the cultural context in which students are born and/or live. To look for evidence of that, the essays were gathered from seven schools – each represented by a class – located in Rome, three of which in the historical center and four in suburbs. The two areas are assumed to be representative of a medium-high socio-cultural context and a medium-low one, respectively. Moreover, all students involved in the collection were asked to fill in a questionnaire to provide information about their biographical, socio-cultural and sociolinguistic background. It consists of 34 questions, divided into two groups: the first thirteen concern learners' biographical data (e.g. language(s) spoken at home, date and place of birth, parents' education and employment, etc.), while the remaining twenty-one explore their writing habits. Among the others: if they like writing outside school, which kind of texts is their favourite, how much time they spend writing, reading or listening to music, and so on. The distribution of the answers to the first set of questions seems to reveal the existence of an actual bond between the position of the school and the socio-cultural context: the schools of the center are mostly attended by pupils who usually speak 'Italian' or 'Italian and a foreign language' at home and whose parents occupy high-paying jobs; on the other hand, peers in the suburbs more frequently speak dialects and foreign languages and their parents hold lower ranked working positions. Interestingly, these results align with previous research in sociolinguistics, such as the study conducted by Chini (2004) and Chini and Andorno (edited by) (2018) which aimed to characterize plurilingualism within the Italian school context.

3.1 Corpus composition

The corpus comprises 1,352 essays (369,456 tokens altogether) written by a total of 156 students, 153 in the first year and 155 in the second. Overall, the compositions respond to 124 writing prompts that pertain to five textual typologies: reflexive, narrative, descriptive, expository and argumentative. Each one requires specific communication and writing abilities.

Furthermore, all pupils were asked to develop a "common prompt" at the end of each school year. In particular, the one assigned at the end of the first year was the Italian version of Task 9 of the IEA-IPS study (Lucisano 1984; Corda Costa and Visalberghi 1995), i.e. a letter to advise a younger student how they should write in order to get good grades in the school; the one given at the end of the previous year was a modified version of the same Task 9, adapted to learners' class and age. Table 1 reports their formulation, as well as an example prompt for each typology. The common prompts were aimed at understanding how learners internalize the writing instructions received in the considered period. In this regard, Barbagli et al. (2015) showed that first-year students' suggestions tend to concern the emotional sphere (e.g. *non aver paura*, 'have no fear', *rifletti prima di scrivere*, 'think before writing'), while the second-year pieces of advice focus more on meta–linguistic aspects, such as the use of verbs or the adherence to the prompt.

Observing the distribution of the five typologies (Table 2), some differences emerge over the two years and the seven schools. The first is merely numerical: the number of prompts given by teachers in the historical center tends to be higher than in the suburban

Table 1
Prompt examples based on the different textual typologies.

Textual typology	Prompt example
Reflexive	What does reading a good book or listening good music represent to you? Make some examples if you want.
Narrative	Invent a myth on the following topic: the laughter.
Descriptive	Hi, I am… describe yourself in a detailed way.
Expository	Child exploitation and slavery: a problem that directly affects us.
Argumentative	In your opinion, how much do mass media and advertising influence people's choices and behaviors?
Common Prompt (I year)	A friend of yours is beginning the fifth year of primary school with your teachers and confessed that is particularly afraid of writing works they will be asked to do. Write them a letter telling about your experience, the positive aspects and also your difficulties in the writing assignments you were asked to do in the fifth grade. Tell them about the works that you liked most and those you liked least and also about the suggestions that teachers gave you to teach you how to write well and how they used to correct writing assignments. Give them useful tips to get by.
Common Prompt (II year)	A boy younger than you has decided to enroll at your school. He wrote to you to ask you how to write an essay that can get good grades by your teachers. Send him a friendly letter describing at least five points that you believe are important for your teachers when they evaluate an essay.

schools. According Barbagli, Lucisano, and Sposetti (2017), two teachers in the suburbs decided to get their pupils to practice in class and at home, proposing them only one examination per quarter, after realising that their starting language competence was very low. Secondly, if 'reflexive' is the most frequent textual type in both years, from the first to the second year the amount of narrative prompts is halved while the expository and argumentative ones are doubled. This different distribution is a consequence of teachers' approach to teach writing: composing a narrative text is considered an easier task – since it requires more rudimentary cognitive and writing skills – than writing an argumentative or expository essay, for which more complex linguistic and discourse-structuring abilities become relevant (Kellogg 2008; Barbagli et al. 2016).

3.2 Error annotation

In addition to the longitudinal nature, the most significant trait that distinguishes CItA from the other Italian L1 learners' corpora is the annotation of many types of errors with the corresponding corrections. It has to be noted that error annotation is a quite challenging matter for at least two reasons: first of all, it assumes the occurrence of a deviation from a linguistic norm, that in itself is a conventionally accepted arbitrary concept. Secondly, while this kind of annotation is commonly practiced on L2 corpora in order to e.g. investigate the properties of interlanguage (Brooke and Hirst 2012)

Table 2
Distribution of the textual typologies in CItA.

Textual typology	Center	Suburbs	Total
First year			
Reflexive	25	13	38
Narrative	18	4	22
Descriptive	2	1	3
Expository	0	1	1
Argumentative	2	2	4
Sub-total	47	21	68
Second year			
Reflexive	24	5	29
Narrative	3	6	9
Descriptive	0	0	0
Expository	4	5	9
Argumentative	5	4	9
Sub-total	36	20	56

or automatically detect and correct errors (Dahlmeier, Ng, and Wu 2013), an L1 error taxonomy did not exist for the Italian language.

To fill this lack and be able to annotate the errors contained in CItA essays, a new annotation schema was defined. In line with the literature on evaluation of written skills of L1 Italian learners (Corda Costa and Visalberghi 1995; De Mauro 1983; Colombo 2011), Berruto's definition of "neo-standard Italian" (Berruto 1987) was adopted as linguistic norm. Similarly to those already existing in other languages (e.g. the one defined by Granger (2003) for French L2 learners'), it is a three-level schema including: the macro-class of error (i.e. grammatical, orthographic and lexical); the class of error, that is to say the linguistic element involved; and the corresponding type of modification required to correct it (Table 3). According to the format introduced by Ng et al. (2013), CItA errors are annotated as follows:

[...] Non mi sembra giusto che uno <M t="112" c="sia">è</M> "uguale" agli altri avendo un Samsung Galaxy e che se uno <M t="112" c="compra">comprato</M> un iPhone diventa subito popolare [...]

("[...] it does not seem fair to me that one who has a Samsung Galaxy is "equal" to others and that, if one buys an iPhone, they immediately becomes popular [...]")

The tags <*M*> and </*M*> ('Mistake') mark the textual area occupied by the error, the attribute *t* ('type') specifies its macro-class and class, and *c* ('correction') indicates the correct form. In the reported example, there are two mistakes related to the misuse of verbal moods: the indicative form *è* instead of the subjunctive *sia* and the past participle of

the verb *comprare* ('to buy') instead of the third person singular of the present indicative. Applying the scheme and following this format, CItA errors were manually annotated by a teacher of middle school, helped by two undergraduate students in *Digital humanities* who had been adequately trained.

The statistical distribution of errors (Table 3) seems to support the hypothesis underlying the collection of CItA – several common trends in the evolution of writing competence occur during the transition from the first to the second year – since most categories of errors (marked with an asterisk) vary in a statistically significant way over the two years. It can be noted that in both years (rows 'Total') orthographic and grammatical errors have the highest frequencies (47.63/44.72% and 46.41/48.7%, respectively) while lexical ones are far less (about 6%). Going into detail, the unclassified orthographic mistakes (i.e. the class 'Other') are the most frequent ones (22.32%), followed by the incorrect use of verb tenses (11.26%), the unclassified grammatical errors (6.37%) and the wrong use of prepositions (6.6%).

Concerning error frequency distribution per year, it emerges that almost all categories are similarly distributed in the two years. However, second-year essays include a considerably higher percentage of mistakes referring to verb morphology, especially in terms of incorrect tense inflection. As previously stated, from one year to the next narrative prompts are replaced by argumentative and expository ones, that involves more complex linguistic and discourse–structuring abilities. Moreover, older students are more aware that "good" writing requires organizing ideas in larger passages, in which the temporal relationships between actions and events should be reconstructed through appropriate shifts in verb tenses and moods. Therefore, the higher number of errors related to verbs could depend both on the more challenging prompts assigned in the second year, and on students' intention to put into practice teachers' writing instructions in order to produce more elaborate essays. Nevertheless, this ability develops across school years, as indicated by previous studies in the literature (Wilcox, Yagelski, and Yu 2013).

To conclude, it is worth mentioning that the statistical distribution of grammatical errors varies significantly with respect to the city areas. As shown in Table 4, their average frequency diminishes over the two years in all the schools located in the historical center and in two suburban institutes, increasing in the remaining two. Surprisingly, the highest amount (on average) is observed in a school of the center, even though its difference over the years is doubled as compared to the other six. Instead, orthographic errors do not vary significantly in relation to any background information. This aligns with previous studies suggesting that mastering orthography requires a longer time (Colombo 2011; Ferreri 1971; Lavinio 1975; De Mauro 1977). However, it could also indicate a general insensitivity to spelling mistakes at this level of education and within this particular age group, potentially reflecting a common characteristic of the neo-standard Italian.

4. Dataset construction

To fulfill the main two purposes of our investigation – i.e., identifying which are the linguistic features that make an essay perceived as good and evaluating the impact of linguistic errors on such a perception – we needed to collect evidences of what a well written production is according to a native speaker. We thus decided to model the perception of writing quality as a manual classification task: proposing two texts to our target user, we wanted them to choose the best written one. By gathering a substantial amount of preferences on a couple of essays, the underlying idea was that we could assume that the most chosen one was actually the best. In order to collect

Table 3
Error annotation schema. Error categories varying significantly over the two years (i.e. $p < 0.05$) are marked with an asterisk.

Class of error	**Type of Modification**	**I year**	**II year**
		Freq %	Freq %
	Grammar		
Verbs	Use of tense *	7.78	15.67
	Use of mood *	4.25	4.92
	Subject-Verb agreement *	2.85	4
Prepositions	Erroneous use	6.48	6.75
	Omission/Redundancy	1.03	0.72
Pronouns	Erroneous use	5.09	3.54
	Omission *	0.41	0.59
	Redundancy	2.70	1.57
	Erroneous use of relative pronoun *	2.13	1.70
Articles	Erroneous use	5.81	3.54
Conjunctions	Erroneous use	0.57	0.52
Other		7.31	5.18
Total		46.41	48.7
	Orthography		
Double consonants	Omission *	6.74	5.05
	Redundancy	3.27	3.67
Use of h	Omission *	3.21	1.64
	Redundancy	1.66	1.11
Monosyllables	Erroneous use of monosyllabic words *	4.87	4.07
	adverb *po* and *pò* instead of *po'*	1.66	1.64
Apostrophe	Erroneous use *	4.82	4.52
Other		21.77	23.02
Total		47.63	44.72
	Lexicon		
Vocabulary	Erroneous use	5.60	6.56

Table 4
Average number of grammatical errors with respect to school years and city areas.

City area	School	I year	II year	Difference
Center	1	2.6	0.9	1.7
	2	5.2	3.1	2.1
	3	15.1	9.3	5.8
Suburbs	4	3.5	8.2	-4.8
	5	6.4	4.6	1.9
	6	5.4	4.6	0.8
	7	1.5	2.8	-1.3

judgments on many couples, we rounded essay pairs up to obtain different questionnaires and we administered a crowdsourcing task. In a broader meaning, crowdsourcing is a methodology that refers to any typology of online collaborative activity but many different – and often contrasting – definitions were given. Analysing about forty of them, Estellés and González (2012) extracted the common elements and proposed the following integrated definition:

> "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

In our study, the "participative online activity" is the completion of a survey and the "group of individuals" involved is formed by Italian native speakers of different ages and cultural background. We would like to underline that the reliability of data obtained via crowdsourcing has been well acknowledged in recent years also in the linguistics and computational linguistics communities. For instance, the thorough survey by (Munro et al. 2010) has shown that the quality of findings obtained from the crowd is often comparable, if not higher, to controlled laboratory experiments. Besides, crowdsourcing allows to reach a broader population, in terms of age, education, profession and etc. and it is thus more suitable to catch the 'layman' perception of writing quality, which is an aspect that qualifies our study with respect to similar ones, which instead focused on judgments given by experts (namely, teachers).

4.1 Essay selection

To collect native speakers' evaluations, we designed ten surveys, each including ten pairs of essays of the same grade. We selected 200 essays from CItA that ranged from a

Table 5
Composition of the ten questionnaires.

Survey	Selection criteria	Number of pairs	
		I year	II year
1	Common prompts	5	5
2	Narrative	10	0
3	Narrative	0	10
4	Reflexive	10	0
5	Reflexive	0	10
6	Descriptive	8	2
7	Expository	3	7
8	Argumentative	3	7
9	Error bins	10	0
10	Error bins	0	10

minimum of 141 tokens to a maximum of 1153 tokens and whose average length was 359.4 tokens.

Table 5 reports the criteria we defined for the selection of the couples of texts to be included in the questionnaires: the first comprises ten pairs – five for each school year – responding to the common prompts given at the end of the years: such a composition allows the comparison between texts simultaneously written by students attending different schools and discussing the same topic. Questionnaires 2-8 gather essays that develop prompts pertaining to the same textual typology, paired according to the school year in which they were written. This choice was based on the assumption that their similarity with regard to the content could let the annotator focus on stylistic issues to orient their judgment. For example, it was meant to avoid a text on a serious and committed topic being preferred to a better written fairy tale. They were designed according to the already seen (Table 2) distribution of textual typologies in CItA: both narrative and reflective texts are dedicated two questionnaires, one per year. Instead, each of the other three typologies (i.e. descriptive, expository, argumentative) occupies only one questionnaire, in which the proportions of pairs with respect to the school year reflect their general distribution in the corpus. Finally, in surveys 9 and 10 the essays were paired according to their number of errors: for each year, we divided the range between the minimum amount of errors (0) and the maximum one (49 for the first year, 43 for the second one) into ten error bins and designed the two surveys choosing a couple of productions for each bin. Surveys comparing essays with a similar amount of errors were meant to investigate which categories of errors have a greater impact on human judgment.

While designing the essay selection criteria, we also took the spatial dislocation of schools into consideration. Indeed, 30 out of the total 100 pairs – 16 for the first and 14 for the second year – gather a text written in a suburban school and one in the historical center.

4.2 Creation and distribution of the questionnaires

After designing the surveys, we moved on to their implementation. We went through the main free web applications dedicated to the creation of questionnaires (e.g. Google Forms, Microsoft Forms), but none was equipped with the customization facilities we needed. Therefore, we choose to rely on the QuestBase platform[4]. As shown in Figure 1, we juxtaposed the essays through its built-in HTML editor and gave surveys a graphical layout with a CSS stylesheet.

The definitive structure of the questionnaire comprises twelve pages: the first one reports the *filling-in instructions*; the second contains the *personal data entry form*: we asked the annotators to provide some personal information (i.e. age, sex, education), in the total guarantee of anonymity and only for statistical purposes. Finally, each of the remaining ten pages is occupied by two side by side essays and a field with a radio button that has to be used to express the answer (Figure 1). They have to choose the option '1' if they prefer the first essay, '2' otherwise. Once the form is submitted, the following message is displayed: *Hai completato il sondaggio. Grazie per il tuo prezioso contributo!* ('You completed the survey. Thank you for your precious help!'). It is worth focusing a little more deeply on the submission instructions. Trying to provide clear and exhaustive directions, we proposed the following guidelines[5]:

> *Ciao!*
> *Il presente sondaggio è rivolto a partecipanti di madrelingua italiana. La sua compilazione richiede circa 20 minuti. Pima di proseguire, dando il consenso alla partecipazione, ti spieghiamo in cosa consiste.*
> *Nelle pagine che seguono leggerai dieci coppie di temi scritti da studenti del primo e del secondo anno di scuola media. I testi possono contenere un certo numero di errori.* ***Per ciascuna coppia ti chiediamo di indicare quale dei due temi ritieni sia scritto meglio****.*
> *Non esistono risposte giuste o sbagliate: conta semplicemente quello che pensi! Tieni presente che i temi di una stessa coppia possono trattare argomenti diversi, ma questo non deve influire sul tuo giudizio.*
> *La tua partecipazione al sondaggio è completamente libera. Se in qualsiasi momento dovessi cambiare idea e volessi interrompere il test, potrai farlo liberamente.*
> *Un'ultima cosa: prima di iniziare il sondaggio, ti chiediamo di darci alcune tue informazioni anagrafiche, che serviranno solo a fini statistici. I dati rimarranno completamente anonimi e in nessun modo le risposte verranno associate alla tua persona.*
> *Se hai dubbi, curiosità o proposte di miglioramento, scrivimi all'indirizzo: a.cerulli1@studenti.unipi.it.*
> *Buona lettura!*

The users were simply asked to choose the best text of each pair. It is an intentionally generic indication, since we wanted them to rely on their native speaker's intuition,

4 https://questbase.com/en/home-questbase/

5 For the sake of completeness, we report the English translation of the guidelines: "Hello! This survey is addressed to Italian native speakers. Its submission requires about 20 minutes. By completing it, you give your consent to participation. Before going on, we explain to you what it consists of. In the following pages you will read ten pairs of essays written by Italian L1 learners during the first two years of lower secondary school. The essays may contain linguistic errors. For each pair, you are asked to choose the best written of the two essays. No answers are right or wrong: you only have to express your opinion! Bear in mind that the essays of a pair can concern different topics, but this must not affect your judgment. Your participation to the survey is completely free. You may withdraw from it at any time. Before starting the survey, we ask you to provide some personal information that will be used for statistical purposes. Data will remain completely anonymous and will not be connected to you in any way. If you have doubts, curiosities or improvement proposals, please write me to the address: a.cerulli1@studenti.unipi.it. Have a good read!"

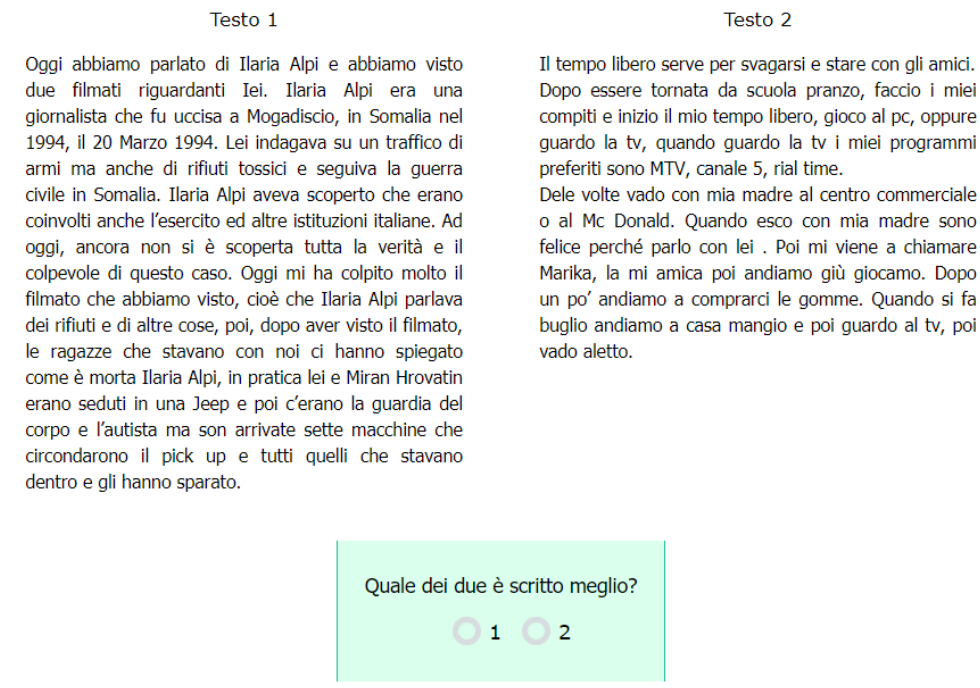

Figure 1
Comparison of a pair of essays extracted from one of the ten surveys.

instead of focusing on specific aspects (e.g. topics discussed or linguistic errors). In other words, their answers had to arise from an instinctive reaction to a quick reading of essay pairs, based on the entirety of linguistic knowledge learnt over time.

To assess the adequacy of the defined structure for our purposes, we created a test survey and distributed its link through WhatsApp and social networks (i.e. Facebook and Instagram). It included eight essay couples randomly extracted from CItA and two 'control pairs', consisting of clearly unbalanced texts, whose aim was to evaluate the annotation accuracy. The administration returned interesting results. As sign of the efficiency of the propagation method, the survey was correctly submitted 43 times by an heterogeneous sample of people ranging from 17 to 51 years of age, mostly holding a high school diploma (48.8%) or an academic degree (41.9%). The answer to 'control pairs' also satisfied our expectations, since the two better essays were preferred 40/43 and 37/43 times, respectively.

At this point, we started collecting evaluations. Using Linktree[6] we added the ten questionnaires URLs to a single web page and shared its link through the previously mentioned social media platforms: clicking on it, users were redirected to the page and could access every survey.

6 https://linktr.ee/

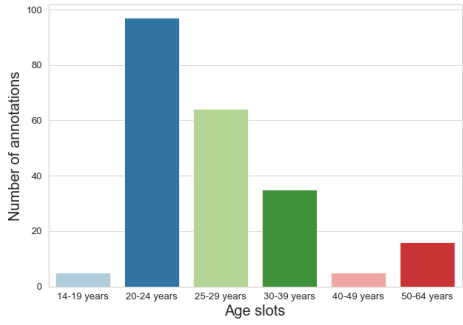

Figure 2
Distribution of annotations with respect to readers' age bins.

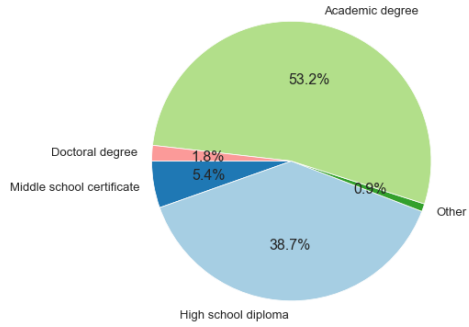

Figure 3
Distribution of annotations with respect to readers' education.

4.3 Collection and analysis of human judgments

We collected 223 annotations distributed quite homogeneously among the ten surveys, except for the first that was submitted 28 times. It is interesting to focus on the heterogeneous composition of the readers cross-section. Concerning 'gender', the majority of answers (183 units = 82.1%) were given by women, against the 38 (17%) by men; just two people preferred not to specify it. As regards 'age', the sample was partitioned into six bins (Figure 2): '20-24 years' was the most frequent class (97 units), followed by '25-29 years' (64 units). This means that the great majority of the readers (72.5%) ranged from 20 to 29 years of age. Furthermore, 35 evaluations (15.8%) were made by natives between 30 and 39 years of age, while people belonging to the remaining bins contributed to the collection for an overall 11.7%. Finally, Figure 3 shows the distribution of submissions with respect to readers' education: almost all the annotations (91.9%) were given by people holding an academic degree (118 units, equal to 53.2%) or a high school diploma (86 units, equal to 38.7%). 12 annotators (5.4%) had a middle school certificate and just 4 (1,8%) held a doctoral degree. The last two indicated a non-specific 'Other'.

Since the questionnaires received an amount of responses ranging from 20 to 28, we decided to select the same number of most coherent annotations for each. For this purpose, we defined a selection function to discard the most inaccurate submissions and consequently improve the quality of dataset. We firstly built the average vector of every survey (a) as the set of ten values '1' or '2' chosen according to the most assigned label to each pair of essays; then, we calculated the distance between each survey average vector and all its annotations relying on the euclidean metric generalized to the n-dimensional space (Equation 1) that computes the distance between two vectors as the square root of the sum of their sizes squared difference.

However, simply calculating the difference between the average vector of a survey and each of its submissions is a partial evaluation unless the other annotations of the same questionnaire are also taken into account. In other words, if an essay pair is given an answer diverging from the average, the impact of the deviation should be higher the fewer annotators made it. So, to give relevance to the deviating degree of answers differing from the average, we assigned each couple a weight (w_i) equal to the number of preferences received by the 'winning' essay. Thus, we computed the weighted distance between annotations and average vectors (Equation 2).

$$d(a,v) = \sqrt{\sum_{i=1}^{n}(a_i - v_i)^2} \quad (1) \qquad wd(a,v) = \sqrt{\sum_{i=1}^{n} w_i(a_i - v_i)^2} \quad (2)$$

Then, for every survey we created two rankings of annotations – from the closest to the most different from the average – by sorting weighted and unweighted distances in ascending order. To choose the most consistent group per survey, we estimated the Inter-Annotator Agreement (IAA) of the first 15 and 20 annotations according on Krippendorff's alpha (α), a coefficient that expresses IAA in terms of observed (D_o) and casual (D_e) disagreement (Krippendorff 2011):

$$\alpha = 1 - \frac{D_o}{D_e} \quad (3)$$

We noticed that IAA values of the first 15 annotations ordered by their increasing values of weighted distance were the highest. Therefore, we took them into account (150 total annotations) for the following analysis and discarded the remaining 73[7]. It is noteworthy that the selection led us to an average IAA of 0.26, that is a much higher value than the initial 0.12.

The analysis of discarded submissions reveal interesting trends with respect to annotators' personal data. Concerning 'gender', in addition to those by the two people who did not specify it, 54 rejected responses were made by women and 14 by men. However, the percentage of the former is higher than that of the latter (36,8% and 31,1%, respectively). Regarding 'age', '50-64 years' is the class of which the highest percentage of annotations was dropped (9 out of 16 = 56,25%), while the lowest proportions – just over 20% in both cases – refer to the bins '30-39' and '40-49' (8 out of 35 and 1 out of 5, respectively). The two most populated classes ('20-24' and '25-29') lost about 32% of responses (31 out of 97 and 21 out of 64, respectively). As for 'education', most submissions by people with a 'Middle school certificate' were rejected (7 out of 12 = 58,3%). The classes 'High school diploma' and 'Academic degree' had about the same number of discards (32 out of 86 and 33 out of 118, respectively), but in percentage terms the gap is wider: 37% for the former and 28% for the latter. These values could suggest that the higher the cultural level of natives, the more accurate their annotation are.

Relying on the selected annotations, we established the 'winning' and 'loser' essays of every pair: the former was the one that received an higher number of preferences and the latter was the less chosen one. Consequently, we could split our annotated corpus into two subsets of 100 texts each, one comprising all 'winning' essays and the other the 'loser' ones.

As discussed in Section 3, the collection of CItA was also based on the assumption that the development of L1 learners' writing competence could be affected by some variables of their socio-cultural background, among which the school position. Thus, the essays were gathered from schools in both the historical center and the suburbs. In Section 4.1, we already commented that 30 pairs of essays – 16 for the first year and 14 for the second – set a comparison between texts composed in the two city areas. Interestingly enough, in 18 cases (60%), the 'winning' production was made by a center student. This would support the hypothesis that they have higher writing skills than their suburban peers. Considering each year independently, we found out a significant difference: while

7 The corpus of evaluated essays is available at the following link: `http://www.italianlp.it/EvaluatedEssays.zip`

essays of the downtown schools were preferred in 11/16 first-year pairs (68,75%), in those of the second year the amounts of 'winning' texts coincide (7 for both areas). This could be a further proof of the "two different speeds of development" mentioned by Barbagli et al. (2016): suburban students' starting level of linguistic competence is lower but it improves more rapidly from the first to the second year of lower secondary school.

5. Data Analysis

5.1 Studying the linguistic phenomena underlying the perception of writing quality

The first purpose of our investigation aimed at identifying whether essays perceived as well-written have a peculiar style which can be represented in terms of a specific set of linguistic features. To this end we adopted the *linguistic profiling* framework, a NLP-based methodology in which a large array of linguistically-motivated features automatically extracted from annotated texts are used to obtain a vector-based representation of it. Such representations can be then compared across texts representative of different textual genres and varieties to identify the peculiarities of each (Montemagni 2013; van Halteren 2004). To perform this analysis, we relied on Profiling-UD[8], a recently introduced tool that implements the underpinnings of the linguistic profiling methodology and allows the extraction of a wide set of features covering lexical, morpho-syntactic and syntactic phenomena from a text (or collection of texts) linguistically annotated according to the Universal Dependencies (UD)[9] formalism. An overview of the features computed by Profiling-UD and used in this study is shown in Table 6. For a complete description of them, the reader is referred to Brunato et al. (2020).

It has to be noted that these features turned out to be highly predictive in many scenarios, all related to modeling formal aspects of a text rather than its content, such as in authorship profiling analyses where they showed to be helpful in identifying specific traits of an author or groups of authors (e.g. gender, native language) from the texts they write (Cocciu et al. 2018; Cimino et al. 2018), or in the case of the automatic assessment of 'perceived' linguistic complexity according to conscious readers' judgments (Brunato et al. 2018; Iavarone, Brunato, and Dell'Orletta 2021). In light of this, we expect them to be useful also for investigating how they might influence human judgments of writing quality.

Using Profiling-UD, we first analyzed each text comprised in the two subsets – i.e. the 'winning' and the 'loser' essays – thus converting each text into its feature-based vector representation, where each dimension of the vector corresponds to the average value of a given linguistic feature in the examined essay. We then estimated three statistical indices for each considered feature in the two groups: the arithmetic mean to summarize the set of values associated to the same feature, the standard deviation as an indicator of data dispersion around the average and the coefficient of variation to normalize and make comparable phenomena measured on different scales. Table 7 shows the mean and standard deviation of an excerpt of the tracked characteristics. As it can be noticed, average values computed for the same feature in the two subsets are often similar, but in some cases they diverge considerably. To have a better understanding of these data, we carried out two separate statistical evaluations sharing the goal of identifying which linguistic features impact more on the rating assigned by annotators.

8 http://linguistic-profiling.italianlp.it/
9 https://universaldependencies.org/

Table 6
Overview of the linguistic features used in this study.

Level of annotation	Linguistic feature	Label
Raw text properties	Total number of sentences	n_sentences
	Total number of tokens	n_tokens
	Avg. number of tokens per sentence	tokens_per_sent
	Avg. number of characters per word	char_per_tok
Lexical variety	Type/Token Ratio in the first 100 or 200 lemmas	ttr_lemma_chunks_100, ttr_lemma_chunks_200
	Type/Token Ratio in the first 100 or 200 words	ttr_form_chunks_100, ttr_form_chunks_200
	Morphosyntax	
POS tagging	Dist. of the 17 UD POS-tags	upos_dist_*
	Lexical density (i.e., content words/total words)	lexical_density
Inflectional morphology	Dist. of verbs and auxiliaries according to their tense, mood, form, gender, number and person	verbs_*_dist_* aux_*_dist_*
	Syntax	
Verbal predicate structure	Avg. dist. of verbal heads	verbal_head_per_sent
	Avg. dist. of roots headed by a verbal lemma	verbal_root_perc
	Verbal arity	avg_verb_edges
	Dist. of verbs for arity class (from 0 to 6)	verb_edges_dist_*
Global and local parsed tree structures	Mean of the maximum tree depths of each sentence	avg_max_depth
	Avg. number of tokens per clause	avg_token_per_clause
	Avg. length of dependency links	avg_links_len
	Mean of the longest dependency links of each sentence	avg_max_links_len
	Length (n. tokens) of the longest dependency link	max_links_len
	Avg. length of prepositional chains	avg_prepositional_chain_len
	Total number of prepositional chains	n_prepositional_chains
	Dist. of prepositional chains by depth (from 1 to 4)	prep_dist_*
Order of elements	Dist. of subjects/objects preceding the verb	subj_pre, obj_pre
	Dist. of subjects/objects following the verb	subj_post, obj_post
Syntactic relations	Avg. dist. of UD 37 universal dependency relations	dep_dist_*
Use of subordination	Dist. of principal/subordinate clauses	principal_proposition_dist subordinate_proposition_dist
	Dist. of subordinate clauses following the main clause	subordinate_post
	Dist. of subordinate clauses preceding the main clause	subordinate_pre
	Avg. length of subordinate chains	avg_subordinate_chain_len
	Dist. of subordinate chains by depth (from 0 to 5)	subordinate_dist_*

In what follows we describe the method underlying the two evaluations and discuss our most interesting findings.

5.2 Linguistic features that vary significantly

The first evaluation was meant at assessing whether the variation between the average values of feature extracted from the 'winning' and the 'losing' essays was statistically

Table 7
Mean and (standard deviation) of an excerpt of the tracked phenomena with respect to 'winning' and 'loser' essays. Features varying in a statistically significant way between the two groups are marked with an asterisk. Features highlighted in bold are also the ones that turned out to be more uniformly widespread in the 'winning' group, according to the ranking established by the coefficient of variation (see Subsection 5.3).

Feature	'Winning' essays	'Loser' essays
	Avg (StDev)	Avg (StDev)
Raw Text Features		
n_sentences	17.46 (9.40)	16.12 (8.14)
n_tokens *	374.95 (127.33)	342.74 (116.27)
tokens_per_sent	24.48 (10.77)	23.60 (8.35)
char_per_tok	4.40 (0.20)	4.38 (0.22)
Lexical Variety		
ttr_lemma_chunks_100	0.61 (0.05)	0.60 (0.06)
ttr_lemma_chunks_200	0.48 (0.13)	0.47 (0.13)
ttr_form_chunks_100 *	0.72 (0.057)	0.71 (0.06)
ttr_form_chunks_200	0.58 (0.154)	0.57 (0.15)
Morphosyntax		
upos_dist_ADJ	5.08 (1.91)	5.13 (2.07)
upos_dist_ADV	7.049 (2.29)	6.80 (2.46)
upos_dist_CCONJ	4.17 (1.28)	4.51 (1.61)
upos_dist_DET	14.03 (2.36)	14.42 (2.43)
upos_dist_NOUN *	16.31 (2.49)	16.98 (2.63)
upos_dist_PRON	8.36 (2.38)	7.98 (2.52)
upos_dist_PUNCT	10.17 (3.35)	9.27 (2.86)
upos_dist_SCONJ	2.33 (1.25)	2.15 (1.15)
upos_dist_VERB	12.95 (2.17)	12.97 (2.58)
lexical_density	0.49 (0.03)	0.49 (0.03)
verbs_tense_dist_Fut *	2.75 (4.37)	2.47 (6.90)
verbs_tense_dist_Past	41.29 (23.73)	40.79 (24.55)
verbs_tense_dist_Pres	42.61 (28.27)	43.07 (28.19)
verbs_mood_dist_Ind	93.91 (8.87)	94.33 (6.78)
verbs_mood_dist_Sub	2.51 (3.79)	3.16 (4.80)
verbs_form_dist_Fin	52.31 (15.20)	54.47 (16.41)
verbs_form_dist_Ger *	3.13 (3.52)	2.32 (3.25)
verbs_form_dist_Inf	24.32 (11.05)	22.79 (12.55)
verbs_form_dist_Part	20.29 (13.75)	20.42 (15.69)
verbs_num_pers_dist_+3	0.02 (0.22)	0.022 (0.22)
verbs_num_pers_dist_Plur+	0.05 (0.52)	0.05 (0.40)
aux_tense_dist_Fut	2.55 (7.38)	3.35 (11.74)
aux_tense_dist_Imp	26.32 (27.94)	21.98 (25.41)
aux_tense_dist_Past	5.0 (8.67)	6.43 (11.10)
aux_mood_dist_Ind	90.52 (11.82)	91.10 (10.61)
aux_mood_dist_Sub *	4.41 (7.22)	2.48 (4.51)
aux_form_dist_Fin	92.28 (8.17)	92.72 (7.83)
Syntax		

Continued on next page

Table 7. Continued from previous page

Feature	'Winning' essays	'Loser' essays
	Avg (StDev)	Avg (StDev)
verbal_head_per_sent	3.56 (1.53)	3.48 (1.48)
verbal_root_perc	88.63 (10.45)	87.94 (10.57)
avg_verb_edges	2.73 (0.23)	2.72 (0.24)
verb_edges_dist_0	1.23 (1.62)	1.06 (1.74)
verb_edges_dist_1	13.45 (5.44)	12.48 (6.30)
avg_max_depth	4.632 (1.11)	4.56 (1.03)
avg_max_links_len	11.28 (5.99)	10.68 (3.92)
avg_links_len	2.766 (0.47)	2.73 (0.41)
max_links_len	31.23 (17.07)	32.01 (17.82)
n_prepositional_chains *	10.70 (6.29)	9.5 (5.92)
obj_pre	31.35 (13.02)	30.017 (15.87)
obj_post	68.65 (13.02)	69.983 (15.87)
subj_pre	83.59 (11.64)	83.707 (11.47)
subj_post	16.41 (11.64)	16.293 (11.47)
dep_dist_compound	0.09 (0.17)	0.18 (0.33)
dep_dist_cop	1.85 (0.98)	1.93 (1.24)
dep_dist_det:predet	0.27 (0.26)	0.24 (0.30)
dep_dist_flat:foreign	0.03 (0.14)	0.02 (0.17)
dep_dist_flat:name	0.31 (0.52)	0.32 (0.79)
dep_dist_parataxis	0.13 (0.21)	0.15 (0.31)
dep_dist_punct	10.167 (3.36)	9.24 (2.84)
dep_dist_root	4.62 (1.44)	4.69 (1.42)
principal_proposition_dist	36.41 (11.95)	37.64 (12.21)
subordinate_proposition_dist	63.59 (11.95)	62.35 (12.21)
subordinate_dist_1	74.89 (13.39)	75.87 (13.71)

significant. Relying on the Wilcoxon rank sum test (Wilcoxon, Katti, and Wilcox 1970), we found out that seven linguistic characteristics (marked with an asterisk in Table 7) varies significantly ($p < 0.05$) between the groups, though no variation turned out to be strongly significant ($p < 0.001$).

In particular, it emerged that 'winning' compositions are, on average, longer (+32.2 tokens) than 'losing' ones in terms of number of tokens (*n_tokens*). This finding may suggest that longer productions are evaluated as more developed, organised and content-rich. Although this is usually true, Crossley, Roscoe, and McNamara (2014) reasonably warn that it may not be the case for all writers. Interestingly, this also reflects the perception that CItA learners have about school writing instructions. Indeed, in an investigation focused on the essays that respond to 'common prompts', Barbagli et al. (2015) showed that two of the most frequent suggestions given by students to an hypothetical younger friend are *Leggi/scrivi molto* ('Read/write a lot') and *Lavora sodo, fai vedere che ti impegni* ('Work hard, show your dedication'). Thus, pupils possibly write more so as to show their dedication and get higher grades. Not by chance, the 9th most salient term extracted from second-year texts is *Voti al tema* ('grades assigned to essays'). Secondly, we observed that a richer vocabulary (in terms of Type/Token Ratio, *ttr_form_chunks_100*) plays a crucial role in annotators' judgment. This reflects another piece of advice included in the above-mentioned ranking, that is *Usa un vocabolario ricco ed*

espressivo ('Use a rich and expressive vocabulary'). It could be a consequence of teachers' encouragement to vary the vocabulary in writing assignments by using synonyms to write clearer and more readable compositions and avoid word repetition as much as possible. The impact of these two features on quality perception was already shown by previous studies dealing with corpora by English L2 learner: higher rated essays comprise more words (Carlson et al. 1985; Ferris 1994; Reid 1990) and exhibit greater lexical diversity (Engber 1995; Grant and Ginther 2000; Jarvis 2002; Reppen 2002). Also Crossley et al. (2014) found out that the strongest quality predictor is the number of word types in a text – i.e., its vocabulary – which strongly correlates (r = .836) with the number of words. This would indicate that essays containing more types (and thus more words) receive higher scores. Values related to the third feature (*upos_dist_NOUN*) reveal that 'winning' essays contain less nouns, although the difference with respect to 'loser' ones is very narrow (-0.67%). As observed in the literature, (see (Montemagni 2013; Biber, Conrad, and Reppen 1998), among others), the nominal style is typical of written texts, and especially of highly informative ones (e.g., newspaper articles, laws), while genres closer to speech contain more verbs. Creative texts like learner essays lie in between and we can expect that the typology of prompts will play a role in emphasizing the similarities with written prose or spoken texts. Our results suggest that readers prefer essays less complex and closer to spoken discourse, which is something similar to what already shown by Crossley et al. (2014), who demonstrated that more 'verbal' essays are rated higher than essays relying more on nouns and nominalizations. However, we intend to deepen this analysis by also considering the typology of prompts under evaluation. With regard to verbal inflection, 'better' productions include on average more future verbs (+0.28%) (*verbs_tense_dist_Fut*), gerund verbs (+0.81%) (*verbs_form_dist_Ger*) and subjunctive auxiliaries (+1.93%) (*aux_mood_dist_Sub*). Verbal tenses differing from present and moods differing from indicative require elevated linguistic skills, which positively influence annotators' judgments. In this regard, also Crossley and McNamara (2011), Crossley et al. (2014) noticed the high effect of complex verb forms on the positive evaluation of a text. Once again, according to above mentioned survey by Barbagli et al. (2015), this is something L1 learners are well aware of: specifically, *Usa correttamente pronomi, verbi e congiunzioni* ('Use correctly pronouns, verbs and conjunctions') and *Usa correttamente i verbi, modi e tempi* ('Use correctly verbs, moods and tenses') are among the most frequent suggestions given in the first and the second year; in addition, *Uso dei verbi* ('Usage of verbs') is the 16th most salient expression in second-year texts. The last feature significantly varying between the two groups is the number of prepositional chains (*n_prepositional_chains*), which is a feature of syntactic complexity: 'winning' compositions have, on average, +1.2 of them.

To sum up, it can be stated that phenomena pertaining to all levels of linguistic description are involved in the choice of a 'better' essay over a 'worse' one: the average length in tokens is a raw text property, while the Type/Token Ratio index belongs to the class of lexical features; the distribution of nouns, verbs and auxiliaries in different moods and tenses are morpho-syntactic characteristics and the presence of prepositional chains is a syntactic one. However, it is thought-provoking that only one feature belongs to the last category, that is the most populated one (see again Table 6). The most likely reason of this has to be sought in the same nature of syntax: being the deepest and most fine-grained level, two much larger subsets are needed to capture the phenomena whose mean values vary in a statistically significant way.

5.3 Degree of variability of linguistic features

As a second evaluation, we calculated the degree of variability of each linguistic feature in the two subset of essays in order to identify which features are more uniformly distributed in the 'winning' set, on the assumption that these features exemplify those linguistic phenomena that are likely to cause the annotator to perceive an essay as better written. For this purpose, we firstly ranked the features of each subset by ordering them according to their increasingly coefficients of variation. Table 8 reports the characteristics that occupy the first twenty positions – that is, the most uniform ones – in both subsets. Given their homogeneous distribution, we can assume that they are intrinsic properties of both the Italian language and the literary genre 'middle school essay'. The former include, e.g., the average word length – in terms of number of characters – and the lexical density, calculated as the ratio between the number of content words over the total number of words. Not coincidentally, they are positioned at the beginning of both lists. The same applies to the distribution of subjects that precede verbs (*subj_pre*), since the canonical, unmarked constituent order of the Italian sentence is SVO (Subject–Verb–Object). Among the latter, instead, it is worth mentioning the vocabulary variation with respect to word forms (*ttr_form_chunks_100*) and lemmas (*ttr_lemma_chunks_100*), the distribution of verbs and auxiliaries in indicative mood (*verbs_mood_dist_Ind*, *aux_mood_dist_Ind*) and finite form (*aux_form_dist_Fin*) and the distribution of first-degree subordinates (*subordinate_dist_1*), i.e., directly depending on the main clause.

To pursue our objective, we then computed another ranking based on the difference between each feature position in the previous classification of 'better' essays and the corresponding one in that of 'worse' ones and putting the results in ascending order. Table 7 reports (in bold) the last ten linguistic characteristics of the new list, i.e., those that, maximally vary in the 'losing' subset, are more uniformly widespread in the 'winning' one. Among them, it is worth mentioning the distribution of future verbs (*verbs_tense_dist_Fut*). As already mentioned in Subsection 5.2, their frequency is higher in 'better' essays. This may give a further evidence supporting the view that native speakers tend to interpret the use of complex verbal forms as an indicator of higher writing skills. Moving on, another feature that is more homogeneous among the 'winners' is the distribution of the parataxis dependency relation (*dep_dist_parataxis*); since its average value is slightly higher in the 'loser' subset, it can be deduced that annotators prefer hypotaxis. This is not surprising: it allows to build more complex and elegant periods that require refined knowledge and mastery of subordination relationships. This syntactic observation seems to find evidence also at the morphosyntactic level, given that 'better' compositions include -0.34% coordinating conjunctions (*upos_dist_CCONJ*), that connect sentences in paratactic periods. In this regard, also Grant and Ginther (2000) found out that higher rated essays include more subordination. It also appears that 'worse' productions have +0.08% copulas (dep_dist_cop), whose use is to link the subject to a subject complement in a nominal predicate structure. This could suggest that annotators do not appreciate this kind of predication in a sentence. Moreover, it is curious that 'better' essays have, on average, +0.1% foreign terms (*dep_dist_flat:foreign*) and -0.1% compound proper nouns (*dep_dist_flat:name*). Finally, it is worth highlighting a higher and more uniform percentage of verbs with few modifiers in the 'winning' corpus (*verb_edges_dist_0*, *verb_edges_dist_1*).

Table 8
First twenty linguistic features of the two subsets ordered by increasing coefficient of variation.

Ranking	'Winning' essays	'Loser' essays
1	char_per_tok	char_per_tok
2	lexical_density	lexical_density
3	ttr_form_chunks_100	verbs_mood_dist_Ind
4	avg_verb_edges	ttr_form_chunks_100
5	ttr_lemma_chunks_100	aux_form_dist_Fin
6	aux_form_dist_Fin	avg_verb_edges
7	verbs_mood_dist_Ind	ttr_lemma_chunks_100
8	verbal_root_perc	avg_prepositional_chain_len
9	subordinate_post	aux_mood_dist_Ind
10	aux_mood_dist_Ind	verbal_root_perc
11	avg_prepositional_chain_len	prep_dist_1
12	subj_pre	subj_pre
13	prep_dist_1	subordinate_post
14	upos_dist_NOUN	avg_links_len
15	avg_token_per_clause	upos_dist_NOUN
16	upos_dist_VERB	avg_subordinate_chain_len
17	upos_dist_DET	upos_dist_DET
18	avg_links_len	avg_token_per_clause
19	avg_subordinate_chain_len	subordinate_dist_1
20	subordinate_dist_1	subordinate_proposition_dist

6. Studying the impact of errors on human ratings

The last phase of our investigation was aimed at assessing the influence of students' errors on human ratings of writing quality.

The 200 evaluated texts contained a total of 1,595 errors, out of which 785 (48.7%) refer to 'Grammar', 721 (44.7%) to 'Orthography' and 98 (6,14%) to 'Lexicon'. As predictable, 'loser' essays contain more errors than 'winner' ones (56.9%) vs 43.1%, respectively). This could be interpreted as a first evidence of the connection between errors and annotators' choice. Further evidence is given by simply counting the essay couples whose 'winning' essay includes less errors than the 'loser', those in which the latter has more than the former and those in which both share the number of errors: in 56 out of the 100 pairs, the essay with fewer errors is the most preferred, while only in 21 cases the 'winner' comprises more errors. The remaining 23 couples pertain to the third category and they are particularly concentrated in the last two questionnaires (see again Table 5).

In order to identify which error macro-classes are more involved in the distinction between 'better' and 'worse' essays, we firstly calculated the average number of errors and the standard deviation for each macro-class in both subsets. Then, relying again on the *Wilcoxon rank sum test*, we found out that grammatical and orthographic mistakes vary significantly between the two groups (Table 9). As expected, 'loser' essays have, on average, a higher quantity of grammatical and orthographic errors (+1.29 and +0.85, respectively). It is worth adding that orthographic mistakes variation ($p = 0.007$) is more

Table 9
Average number of errors in the two subsets per macro-class. Categories whose mean varies significantly between the two subsets are marked with an asterisk.

Error category	'Winning' essays	'Loser' essays
	Avg (StDev)	Avg (StDev)
Grammar *	3.28 (5.52)	4.57 (6.13)
Orthography *	3.18 (4.52)	4.03 (4.83)
Lexicon	0.41 (0.71)	0.48 (0.82)

significant than the other ($p = 0.029$). This could be an indication that native speakers probably judge orthographic deviations worse than grammatical ones. Once again, our findings are in line with Barbagli et al. (2015): *Usa una corretta ortografia* ('Use correct orthography') and *Ortografia aspetti generali* ('Orthography general aspects) are the 2nd and the 8th of the most frequent suggestions given in the second year; moreover, *Errori di ortografia* ("Orthography errors") occupies the 6th and the 1st position among the most salient terms respectively of the first and the second year. The non-significant variations of lexical errors ($p = 0.581$) is probably related to their scarce amount in the analysed corpus. This is the same reason why we preferred to generically take the three error macro-classes into account, rather than considering the many error classes included in CItA annotation scheme (as seen in Table 3). Such a study would certainly have given more significant and interesting results, but it would have required a much higher amount of annotated essays in order to curb the problem of data sparsity. We plan to do it in the continuation of our research.

Interestingly enough, the pair on which all annotators agreed assigning their preference to the second essay is also the maximally unbalanced one with respect to the number of errors: the first text has 34 mistakes (i.e., 14 grammatical, 19 orthographic and 1 lexical), while the second only 9 (i.e., 6 grammatical and 3 orthographic). It is worth noticing that the two productions respond to the same prompt, which seems to reduce the influence of the topic on the assessment. In what follows, we report and comment this couple so as to concretely show how errors are crucial in determining an essay quality and the native speaker's perception of it.

Il film si svolge in Belfast, Irlanda, il film narra di un ragazzo, Jerry che va a Londra con una borsa dentro salsicia, soldi e maglette. Va a Londra con un suo amico. Arrivato a Londra entra in una casa dentro un gruppo di persona che chiamato IRA poi va alla zia a dare le salsicie. Una notte Jerry e il suo amico decidono a dormire nel parco dove incontrarono un vecchio senza casa che dorme in una sedia nel parco, dopo in poi incontrano una prostituta che cade le sue chiave della casa, Jerry entra nella casa della prostituta e ruba dei soldi, poi cambia scena che scopia una bomba in appartamento. Torna a Belfast con soldi e vestito elegante. Mentre Dorme Jerry entrano le polizie che arrestano Jerry. Va in carciere

Il film narra la storia di un ragazzo, Jerry Conlon, nato a Belfast che venne accusato di un reato non commesso in Inghilterra, durante la guerra. Venne messo in un carcere a vita con il padre, la zia Annie e i suoi tre figli di cui uno aveva tredici anni. Dopo un po di tempo venne rinchiuso il vero colpevole che confessò di aver fatto scoppiare la bomba nel PUB, i giudici e il direttore erano al corrente che Jerry non era colpevole ma per far vedere al popolo che erano capaci di catturare i colpevoli misero Jerry in carcere, quando scoprirono che un uomo aveva confessato il reato cercarono di fare qualcosa in modo tale che nessuno al di fuori del carcere sapesse dell'accaduto. Dopo quindici lunghi anni si fece un altra

per 30 anni con il Padre. Un anno dopo, il vero che ha messo la bomba confessa alla polizia. Diciendo che ci sono inocenti. 15 anni dopo esce dal carciere perché hanno sapputo la verita.

Il episodio che mi è piacciuto e quando Jerry viene liberato con i tre compagni che escono nella porta principale. E quella di meno è quando le polizie turturano il suo amico e Jerry per dire solo che hanno messo la bomba.

Il personaggio mi ha colpito è Jerry perché lui rischia di restare in carciere, perché non vuole dire la verita alla vocato. In senzo negativo i capo della polizia perché hanno trattato male Jerry. Le scene che mi hanno colpito è quando Jerry torna a Belfast vestito elegante con soldi, Quando stava morendo il Padre, Quando Jerry ha saputo che morto il padre che gli altri persone in carciere votano un foglio con fuoco nella finestra, E quando Jerry viene liberato.

sentenza sul caso in cui vi partecipò l'avvocato interpretato da Emma Thompson che da poco aveva scoperto che i giudici e il direttore erano consapevoli che Jerry era innocente e che per far bella figura non dicevano niente. L'avocatessa prese l'articolo di Jerry Conlon e scoprì l'accaduto, in tribunale lo fece vedere al giudice che archiviò tutti gli accusati di reato non commesso. Quando Jerry fu archiviato uscì dalla porta principale e disse: «esco dalla porta principale perché ora sono un cittadino libero», fuori lo aspettavano giornalisti con macchine fotografiche e registrazione, venne intervistato e fu riportato in televisione, inquadrando solo il suo volto. Il messaggio del film è che ci può essere un buon rapporto tra padre e figlio in qualunque situazione.

Mi è colpito molto l'affetto di Jerry nei confronti del padre dopo tutto il periodo in cui non sono andati d'accordo, alla fine sono stati molto uniti e Jerry quando il padre è morto gli dispiaque tantissimo.

In the first essay, almost all error classes included in the annotation scheme (see again Table 3) are represented. As regards verbs, some mistakes concern the misuse of tense (e.g., *decidono di dormire nel parco dove incontrarono*, 'they decide to sleep in the park where they met') or mood (e.g., *un gruppo di persona che chiamato IRA*, 'A group of person that called IRA'), as well as the missed agreement between subject and verb (e.g, *Le scene che mi hanno colpito è quando*, 'The scenes that touched me is when'). Secondly, we detect many mistakes related to the 'Erroneous use' of prepositions (e.g., *in Belfast* instead of *a Belfast*, *decidono a dormire* rather than *decidono di dormire*) or articles (e.g. *il episodio* instead of *l'episodio*, *gli altri persone* rather than *le altre persone*). Moreover, several misspellings refer to the 'Omission' of double consonants (e.g., *inocenti* instead of *innocenti*) or their 'Redundancy' (e.g., *sapputo* rather than *saputo*) or pertain to the category 'Other' (e.g., *carciere* instead of *carcere*). Besides, the punctuation is totally arbitrary. Also the second composition has errors, related, for example, to the use of apostrophes (e.g., *un altra sentenza*) or the use of the adverb *po* instead of *po'* (e.g., *un po di tempo*), but their amount is clearly lower. The above, combined with a more canonical use of punctuation and a more structured organization of content, made all annotators prefer the latter text.

7. Conclusions

In this article we have presented a first study for the Italian language aimed at assessing the relationship between the linguistic structure of a text and the native speaker's perception of its writing quality. We motivated our investigation within the framework of linguistic profiling, a NLP-based methodology that allows to characterize a text in terms of the distribution of a wide set of features representative of phenomena spanning across language domains, with the purpose of understanding which of them are more involved in the human assessment of writing quality.

Although our study falls within a longstanding research area focusing on the interplay between the textual features of a composition and the written proficiency of its author (Crossley et al. 2014), the typology of texts we examined represents quite

a novelty in this scenario. In fact, the majority of existing contributions focuses on English learners' corpora, especially of L2 speakers, and takes into account few linguistic phenomena, such as those involved in text coherence or lexical sophistication. A further distinction from previous work concerns the approach adopted for modeling human perception. Instead of resorting to some kind of structured scoring rubric, as made by Crossley and McNamara (2011) and Crossley et al. (2014), we tackled quality assessment as a manual binary classification task: given a pair of essays, readers were asked to choose the one they considered better written. The simplicity of this evaluation method allowed us to propose it as a crowdsourcing task and to gather ratings from native speakers of various ages and cultural background, rather than limiting it to only expert raters. Based on a careful analysis of the distribution of raters' preferences among the collected annotations, we were able to establish the 'better' and 'worse' essay for each pair and, consequently, to split the corpus into two subsets, comprising the former and the latter, respectively. Statistical analyses carried out on the linguistic profiles characterizing the two subcorpora yielded some significant results. For example, we found out that longer compositions are preferred to shorter ones and that lexical variety as well as the use of non-indicative mood and non-present tense verbs positively affect the perceived quality of an essay, while an overuse of nouns over verbs does it negatively. It also seems that annotators appreciate more a subordinating style, reasonably because a prose constructed via hypotaxis is more organized and elegant. Interestingly enough, not only are some of these findings in line with those provided by previous studies on writing quality perception, but also reveal a quite unexpected correspondence between annotators' judgments and the way L1 learners receive writing instructions by teachers (Barbagli et al. 2015). Such a finding could be motivated by the fact that readers – especially the youngest ones – were given similar instructions during their schooling. Comparing the average number of students' errors per category in the two subsets, we confirmed our starting idea that mistakes substantially affect human judgements, also discovering that grammatical and orthographic ones do it in a stronger way.

Altogether, our findings appear consistent enough to be interpreted as indicators of the reliability of our collected data and, more in general, could suggest the effectiveness of crowdsourcing techniques to gather large and reliable amounts of annotated data. They would be valuable resources to train and test NLP algorithms, above all if considering the lack of Italian corpora of graded essays. Despite the promising findings, the limited size of our dataset certainly reduced the amount of results, as already touched upon in Section 5.2. This motivates us to enlarge it by (i) creating and distributing new surveys grouping other essay pairs and (ii) collecting more annotations for the already existing ones. Carrying out again the same analyses on a wider dataset, we expect to be able to identify stronger linguistic predictors that are more likely associated to well-written perceived compositions. Besides, following Miaschi, Brunato, and Dell'Orletta (2021), we could rely on the results to train a binary classification model that, given a pair of texts, automatically performs the task of predicting the best one. Such a tool could be the starting point for the development of an automated scorer able to grade a composition and return a (hopefully) formative feedback, exactly like the Writing Pal (McNamara, Crossley, and Roscoe 2013). Without presuming to replace teachers, AES systems can be a valuable teaching aid for both teachers and students: the former, freed from many time consuming and cost prohibitive elements of essay grading, can focus more on some aspects that these tools are poor at assessing (e.g., argumentation, style, and idea development) (Crossley et al. 2014); the latter can get an immediate and preliminary self-assessment on their written productions so as to better understand their mistakes and hopefully avoid repeating them. Generally speaking, these systems reduce the demands and complications often

associated with human writing assessment, such as time, cost, and reliability (Page 2003; Burstein 2003; Bereiter 2003). An AES system for Italian L1 written productions would be particularly useful if integrated into educational processes based on distance learning paradigms, which in turn need adequate technological infrastructures to be really efficient.

References

Attali, Yigal and Jill Burstein. 2006. Automated Essay Scoring With e-rater® V. 2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Barbagli, Alessia, Pietro Lucisano, Felice Dell'Orletta, and Giulia Venturi. 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.

Barbagli, Alessia, Pietro Lucisano, and Patrizia Sposetti. 2017. Insegnare a scrivere nel biennio della scuola secondaria di primo grado. *Giornale italiano della ricerca educativa*, numero speciale:9–25.

Barbagli, Alessia, Lucisano Pietro, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. CItA: An L1 Italian Learners Corpus to Study the Development of Writing Competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 88–95, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Bereiter, Carl. 2003. *Automated essay scoring: a cross disciplinary approach. Foreword*. Mark D. Shermis and Jill C. Burstein Eds., Lawrence Erlbaum Associates: Mahwah, NJ.

Berruto, Gaetano. 1987. *Sociolinguistica dell'italiano contemporaneo*. Carocci, Roma.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. Corpus linguistics - Investigating language structure and use. In *Cambridge approaches to linguistics*.

Borghi, Carlotta Caterina. 2013. *Analisi di produzioni scritte. Valutazioni e misure automatizzate di elaborati scolastici*. Ph.D. thesis, Università di Roma, La Sapienza.

Brooke, Julian and Graeme Hirst. 2012. Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 779–784, Istanbul, Turkey. European Language Resources Association (ELRA).

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2020. Profiling-UD: A Tool for Linguistic Profiling of Texts. In *Proceedings of the 12th Conference of Language Resources and Evaluation (LREC 2020)*, pages 7145–7151, Marseille, France, May. European Language Resources Association (ELRA).

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium, October-November. Association for Computational Linguistics.

Burstein, Jill. 2003. *The E-rater® scoring engine: Automated essay scoring with natural language processing*. Lawrence Erlbaum Associates Publishers.

Carlson, Sybil B., Brent Bridgeman, Roberta Camp, and Janet Waanders. 1985. *Relationship of admission test scores to writing performance of native and non-native speakers of English*. ETS, Princeton, New Jersey (USA).

Chini, Marina. 2004. *Plurilinguismo e immigrazione in Italia. Un'indagine sociolinguistica a Pavia e Torino*. Franco Angeli, Milano.

Chini, Marina and Maria Cecilia Andorno (edited by). 2018. *Repertori e usi linguistici nell'immigrazione. Una indagine sui minori alloglotti dieci anni dopo*. Franco Angeli, Milano.

Cimino, Andrea, Felice Dell'Orletta, Dominique Brunato, and Giulia Venturi. 2018. Sentences and documents in native language identification. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December.

Cocciu, Eleonora, Dominique Brunato, Giulia Venturi, and Felice Dell'Orletta. 2018. Gender and genre linguistic profiling: A case study on female and male journalistic and diary prose. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December.

Colombo, Adriano. 2011. *«A me mi»: dubbi, errori, correzioni nell'italiano scritto*. FrancoAngeli, Milano.

Corda Costa, Maria and Aldo Visalberghi, editors. 1995. *Misurare e valutare le competenze linguistiche: guida scientifico-pratica per gli insegnanti*. La Nuova Italia, Firenze.

Crossley, Scott A., Kristopher Kyle, Laura K. Allen, Liang Guo, and Danielle S. McNamara. 2014. Linguistic Microfeatures to Predict L2 Writing Proficiency: A Case Study in Automated Writing Evaluation. *Journal of Writing Assessment*, 7(1).

Crossley, Scott A. and Danielle S. McNamara. 2011. Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, pages 1236–1241, Boston, Massachusetts, USA, July.

Crossley, Scott A. and Danielle S. McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115 – 135, 05.

Crossley, Scott A., Rod Roscoe, and Danielle S. McNamara. 2014. What Is Successful Writing? An Investigation Into the Multiple Ways Writers Can Write Successful Essays. *Written Communication*, 31(2):184–214.

Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu. 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, Atlanta, Georgia, June. Association for Computational Linguistics.

De Mauro, Tullio. 1977. *Scuola e linguaggio*. Editori Riuniti, Roma.

De Mauro, Tullio. 1983. Per una nuova alfabetizzazione. In Stefano Gensini and Massimo Vedovelli, editors, *Teoria e pratica del glotto-kit: una carta d'identità per l'educazione linguistica*. FrancoAngeli, Milano, pages 19–29.

Engber, Cheryl A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2):139–155.

Estellés, Enrique and Fernando González. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 20(10):1–14.

Ferreri, Silvana. 1971. Italiano standard, italiano regionale e dialetto in una scuola media di palermo. In *L'insegnamento dell'italiano in Italia e all'estero: Atti del quarto Convegno internazionale di studi*, pages 205–224, Roma, italy, June 1970. Bulzoni Editore.

Ferris, Dana R. 1994. Lexical and Syntactic Features of ESL Writing by Students at Different Levels of L2 Proficiency. *TESOL Quarterly*, 28(2):414–420.

Granger, Sylviane. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3):465–480.

Grant, Leslie and April Ginther. 2000. Using computer-tagged linguistic features to describe l2 writing differences. *Journal of Second Language Writing*, 9(2):123–145.

Iavarone, Benedetta, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199, Online, June. Association for Computational Linguistics.

Jarvis, Scott. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1):57–84.

Kellogg, Ronald T. 2008. Training writing skills: A cognitive developmental perspective. *Journal of Writing Research (JoWR)*, 1(1):1–26.

Krippendorff, Klaus. 2011. Computing Krippendorff's Alpha-Reliability. Technical report, University of Pennsylvania.

Landauer, Thomas K., Darrell Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, Mahwah, New Jersey (USA).

Lavinio, Maria Cristina. 1975. *L'insegnamento dell'italiano. Un'inchiesta campione in una scuola media sarda*. Edes, Cagliari.

Lucisano, Pietro. 1984. L'indagine IEA sulla produzione scritta. *Giornale italiano della ricerca educativa*, 5:41–46.

Lucisano, Pietro. 1988. La ricerca IEA sulla produzione scritta. *Giornale italiano della ricerca educativa*, 1:3–13.

Lucisano, Pietro and Guido Benvenuto. 1991. Insegnare a scrivere: dalla parte degli insegnanti. *Scuola e città*, 6:265–279.

Marconi, Lucia, Michela Ott, Elia Presenti, Daniela Ratti, and Mauro Tavella. 1994. *Lessico elementare. Dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli, Bologna.

McNamara, Danielle S., Scott A. Crossley, and Philip Mccarthy. 2010. Linguistic Features of Writing Quality. *Written Communication*, 27(1):57–86, 01.

McNamara, Danielle S., Scott A. Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, 45(2):499–515.
McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, Cambridge.
Miaschi, Alessio, Dominique Brunato, and Felice Dell'Orletta. 2021. A NLP-based stylometric approach for tracking the evolution of l1 written language competence. *Journal of Writing Research (JoWR)*, 13(1):71–105.
Montemagni, Simonetta. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, pages 145–172.
Munro, Robert, Steven Bethard, Victor Kuperman, Vicky T. Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pages 122–130, Los Angeles, CA, USA, June. Association for Computational Linguistics.
Ng, Hwee Tou, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 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, Sofia, Bulgaria, August. Association for Computational Linguistics.
Page, Ellis Batten. 2003. Project essay grade: Peg. *Automated essay scoring: A cross-disciplinary perspective*.
Reid, Joy. 1990. Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In Barbara Kroll, editor, *Second language writing: Research insights for the classroom*. Cambridge University Press, Cambridge (England), pages 191–210.
Reppen, Randi. 2002. A Genre-Based Approach to Content Writing Instruction. In Jack C. Richards and Willy A. Renandya, editors, *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge University Press, Cambridge (England), pages 321–327.
Rudner, Lawrence M., Veronica Garcia, and Catherine Welch. 2006. An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, 4(4):1–21, 03.
van Halteren, Hans. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, pages 200–207, Barcelona, Spain, July.
Wilcox, Kristen Campbell, Robert Yagelski Yagelski, and Fang Yu. 2013. The nature of error in adolescent student writing. *Reading and Writing*, 27(6):1073–1094.
Wilcoxon, Frank, S. K. Katti, and Roberta A. Wilcox. 1970. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. In *Selected Tables in Mathematical Statistics*, volume 1. American Mathematical Society, Providence, Rhode Island (USA), pages 171–259.