Quale testo è scritto meglio? A Study on Italian Native Speakers' Perception of Writing Quality

Aldo Cerulli[•], **Dominique Brunato**[◇], **Felice Dell'Orletta**[◇]
[•] University of Pisa
a.cerulli1@studenti.unipi.it
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

This paper presents a pilot study focused on Italian native speakers' perception of writing quality. A group of native speakers expressed their preferences on 100 pairs of essays extracted from an Italian corpus of compositions written by L1 students of lower secondary school. Analysing their answers, it was possible to identify a set of linguistic features characterizing essays perceived as well written and to assess the impact of students errors on the perception of text quality. The paper describes the crowdsourcing technique to collect data as well as the linguistic analysis and results.

1 Introduction

The institution of distance learning paradigms, which has become crucial during the Covid-19 pandemic, showed the need to provide schools and universities with Natural Language Processing (NLP)-based tools to assist students, teachers and professors. Nowadays, language technologies are more and more exploited to develop educational applications, such as *Intelligent Computer-Assisted Language Learning (ICALL)* systems (Granger, 2003) and tools for automated essay scoring (Attali and Burstein, 2006) or automatic error detection and correction (Ng et al., 2013). A fundamental requirement for developing this kind of applications is the availability of electronically accessible corpora of learners' productions. Corpora created so far differ in many respects. For instance, considering the types of examined learners, they can gather productions written by 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. In both cases, a peculiarity of existing corpora is that they are cross-sectional rather than longitudinal. A notable exception in the context of Italian as L1 – which is the focus of our contribution – 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].

The diachronic and longitudinal nature of *CItA* makes it particularly suitable to study the evolution of L1 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 et al. (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 research based on *CItA*, but chooses a different approach from the one just mentioned: instead of tracking the development of students' writing competence, we focused on the perception of writing quality by Italian L1 speakers with the aim of understanding whether it is possible to find the linguistic features that are crucially involved in the distinction between 'better' and 'worse' essays according to our target reader.

Contributions To the best of our knowledge, this is the first paper that (i) introduces a dataset of

[1]The corpus is freely available for research goals at http://www.italianlp.it/resources/cita-corpus-italiano-di-apprendenti-l1/

evaluated essays in terms of perceived writing quality by means of a crowdsourcing task, (ii) deals with the correlation between linguistic features and perceived quality of writing and (iii) assesses the impact of students errors on quality perception.

2 Corpus Collection

As previously mentioned, the starting point of our study was the *CItA* corpus. It comprises 1,352 essays, written by 156 pupils of seven lower secondary schools in Rome (three in the historical center and four in the suburbs) during the school years 2012-2013 and 2013-2014. The productions respond to 124 writing prompts that pertain to five textual typologies: reflexive, narrative, descriptive, expository and argumentative. An additional 'common prompt' was presented at the end of each school year, in which students were asked to write a letter to advise a younger friend how to compose better essays. The common prompts were aimed at understanding how learners internalize the different writing instructions given by teachers.

Each essay contained in *CItA* is also provided by a set of metadata tracking students' biographical, sociocultural and sociolinguistic information. Beyond the longitudinal nature, the most significant novelty introduced by *CItA* regards error annotation, which was manually performed by a middle school teacher according to a new three-level schema including: the macro-class of error (i.e. grammatical, orthographic and lexical); the class of error (i.e. verbs, prepositions, monosyllabes); and the corresponding type of modification required to correct it. More details about the *CItA* collection are reported in Barbagli et al. (2016).

2.1 Essay Selection

For the purpose of our investigation, we selected 200 essays from *CItA* to be submitted to human evaluation. The essays ranged from a minimum of 141 tokens to a maximum of 1153 tokens and their average length was 359.4 tokens. Then, to gather judgments on writing quality, we created ten questionnaires, each one consisting of ten pairs of essays of the same grade, and distribute them to native speakers of all ages and cultural background.

Table 1 reports the criteria we adopted to select the pairs of essays. As it can be seen, Survey 1 allows the comparison between essays responding to the common prompts written by students attending the first or the second grades. In surveys 2-8, we chosen essays pertaining to the same textual typology – assuming that their similarity with regard to the content could let the annotator focus on stylistic issue to orient their judgment – and paired them according to the school year in which they were written. Instead, essays in questionnaires 9 and 10 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 understand which categories of errors have a greater impact on human judgment.

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

Table 1: Criteria used for pairing the essays and number of essays for each survey.

2.2 Human Evaluation

After designing the surveys, we moved on to their implementation using the QuestBase platform[2]. We defined a three-section structure including the filling-in instructions, the personal data entry form and the essays evaluation pages.

Filling-in instructions. The first section reported the following submission guidelines:

> 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

[2]https://story.questbase.com/

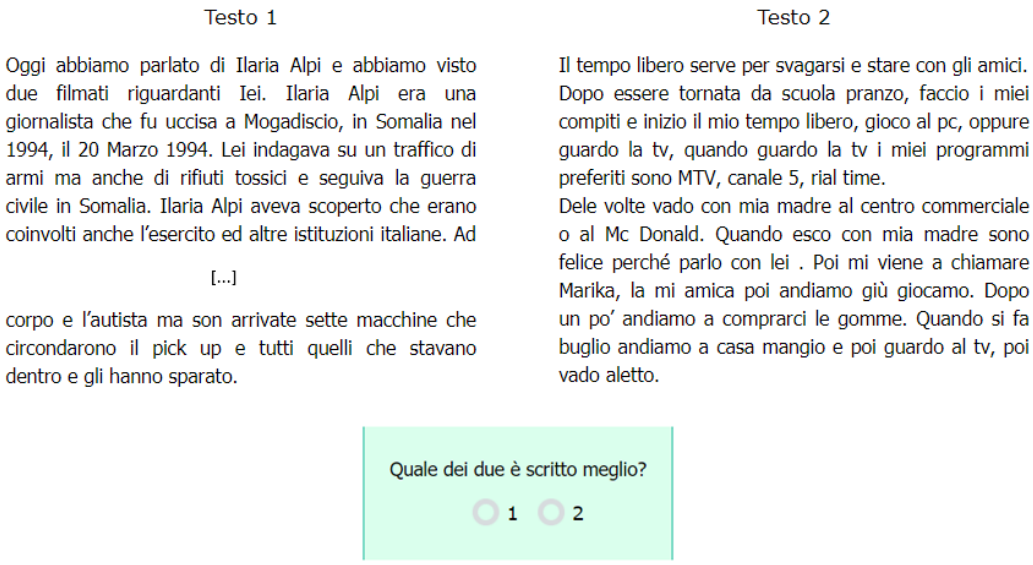

Figure 1: Comparison of a pair of essays extracted from one of the ten surveys.

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!

For the sake of completeness, we also report an English translation of the same 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!

Personal data entry form. The surveys were obviously anonymous. However, as we mentioned before, we asked the annotators to entry some personal information (age, sex, education) for statistical purposes.

Essays evaluation. The third section comprised ten pages, each occupied by two side by side essays and a field to give the answer (Figure 1). The user had to choose the label '1' if they had preferred the first essay, '2' otherwise.

After carrying out a pilot study to test the adequacy of the structure as well as the completeness and clearness of the instructions, we started collecting evaluations. Using Linktree[3] we added the ten questionnaires links to a single web page and shared its link through WhatsApp, Facebook and Instagram: clicking on it, users were redirected to the page and could access every survey.

3 Analysis of Human Judgments

We collected 223 annotations distributed quite homogeneously among the ten surveys, except for the first one, submitted 28 times. It is worth to focus on the heterogeneous composition of the readers sample. Concerning sex, the large majority of answers (183 units, equal to 82.1%) were given by women, against the 38 (17%) by men; just two people preferred not to specify their gender.

Regarding age, we divided the group into six bins (Figure 2). The most frequent class (97 units) was '20-24 years', followed by '25-29 years' (64 units). This means that most readers (72.5%) ranged from 20 to 29 years of age. 35 evaluations (15.8%) were made by natives between 30 and 39 years of age. People belonging to the remaining bins contributed to the task for an overall 11.7%.

[3]https://linktr.ee/

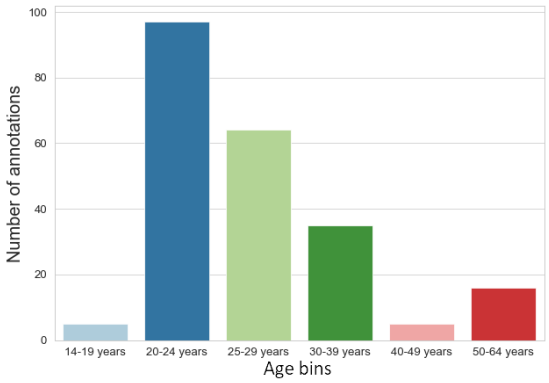

Figure 2: Distribution of annotations with respect to readers' age bins.

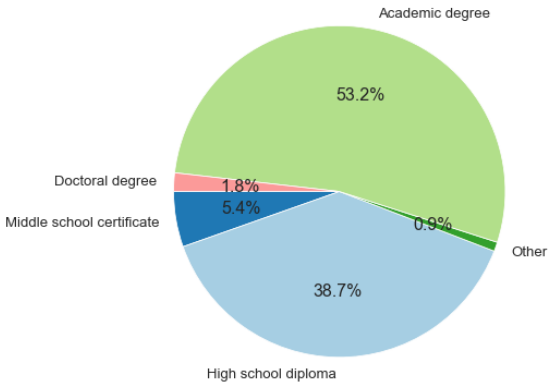

Figure 3: Distribution of annotations with respect to readers' education.

Finally, Figure 3 shows the distribution of submissions with respect to readers' education: 91.9% of annotations 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; 4 (1,8%) held a doctoral degree; the last two indicated a non-specific 'Other'.

3.1 Inter-Annotator Agreement

At this point, we defined a selection function to discard inaccurate annotations and obtain the same number of coherent annotations for each survey. Thus, we firstly built the average vector of every survey 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. We implemented the euclidean metric generalized to the *n*-dimensional space that computes the distance between two vectors as the square root of the sum of their sizes squared difference:

$$\sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} \quad (1)$$

To give relevance to the deviating degree of answers differing from the average, we assigned every pair a weight (w_k) equal to the number of times in which the 'winning' essay was chosen; then, we computed the weighted distance between annotations and average vectors.

$$\sqrt{\sum_{k=1}^{n} w_k(p_k - q_k)^2} \quad (2)$$

Finally, we ranked weighted and unweighted distance values of each survey in ascending order and calculated the Inter-annotator agreement (IAA) of the first 15 and 20 annotations. We implemented 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 submissions ordered by their increasing weighted distance were the highest. Thus, we took them into account (150 total annotations) for the analysis and discarded the remaining 73[4]. 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. Relying on the selected annotations, we established the 'winning' and 'loser' essay of each pair.

4 Data Analysis

We carried out two evaluations: a first one was meant to identify which linguistic features impact more on the human assessment of the writing quality; a second one focused on the impact of students errors on annotators' judgments. In what follows we describe the approach underlying the two perspectives and discuss our most interesting findings.

4.1 Linguistic Profiling and Stylistic Analysis

The first analysis relies on *linguistic profiling*, a NLP-based methodology in which a large set 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;

[4]The corpus of evaluated essays is available at http://www.italianlp.it/EvaluatedEssays.zip

Feature	'Winning'		'Losers'	
	Avg.	SD	Avg.	SD
n_tokens	374.9	127.4	342.7	116.3
ttr_form_chunks_100	0.72	0.06	0.70	0.06
upos_dist_NOUN	16.31	2.49	16.98	2.63
verbs_tense_dist_Fut	2.75	4.37	2.47	6.90
verbs_form_dist_Ger	3.13	3.52	2.32	3.25
aux_mood_dist_Sub	4.41	7.22	2.48	4.51
n_prepositional_chains	10.70	6.28	9.50	5.98

Table 2: Linguistic features whose average varies significantly between the two subsets.

van Halteren, 2004). To perform the analysis, we relied on Profiling-UD[5], a recently introduced tool that allows the extraction of a wide set of lexical, morpho-syntactic and syntactic features from texts linguistically annotated according to the Universal Dependencies (UD)[6] formalism. These features, described in details in Brunato et al. (2020), have been shown to be involved in many tasks, all related to modeling the form rather than the content of a text, such as the assessment of text readability and linguistic complexity and the identification of stylistic traits of an author or groups of authors.

We thus split our annotated corpus into two sections: one comprised all 'winning' essays and the other all 'loser' ones. Using Profiling-UD, we extracted for each text of the two subsets a feature-based vector representation. For each considered feature we calculated the average value, the standard deviation and the coefficient of variation ($\frac{SD}{Avg}$) in the two subsets and we assessed whether the variation between mean values was significant using the Wilcoxon rank sum test.

Table 2 shows the seven linguistic features whose variation turned out to be statistically significant ($p - value < 0.05$), ordered by increasing p-values. It emerges that 'winning' essays are on average longer (32.2 tokens more) than the 'losers' (*n_tokens*), a finding that may suggest that longer compositions are evaluated as more reasoned, structured and content-rich. Interestingly, this also reflects the students' perception of school writing: Barbagli et al. (2015) showed that two of the most frequent suggestions contained in essays that respond to 'common prompts' 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. Secondly, we noticed that a richer vocabulary (*ttr_form_chunks_100*) plays a crucial role in native's judgment. This is in line with another advice of the just mentioned ranking, *Usa un vocabolario ricco ed espressivo* ("Use a rich and expressive vocabulary"), that reflects teachers' encouragement to use synonyms in order to write clearer and more readable compositions. Values related to the third feature (*upos_dist_NOUN*) reveal that 'loser' essays present a slightly higher distribution of nouns. A predominant use of nouns is typical of highly informative texts (e.g. newspaper articles, laws), while genres closer to speech contain more verbs (Montemagni, 2013). Belonging to the second category, a school essay with fewer nouns is probably perceived as more coherent with its genre. Concerning verbal inflection, 'better' productions include, on average, 0.28% more future verbs (*verbs_tense_dist_Fut*), 0.81% more gerund verbs (*verbs_form_dist_Ger*) and 1.93% more subjunctive auxiliary verbs (*aux_mood_dist_Sub*). Verbal tenses differing from present and moods differing from indicative require elevated linguistic skills, which positively influence annotators' choices. The last feature significantly varying between the two groups is the number of prepositional chains (*n_prepositional_chains*): 'winning' compositions have, on average, 1.2 more of them.

Feature	'Winning'		'Losers'	
	Avg.	SD	Avg.	SD
verbs_tense_dist_Fut	2.75	4.37	2.47	6.90
dep_dist_cop	1.85	0.98	1.93	1.24
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_det:predet	0.27	0.26	0.24	0.30
dep_dist_parataxis	0.13	0.21	0.15	0.31
obj_pre	31.35	13.02	30.02	15.87
verb_edges_dist_0	1.23	1.62	1.06	1.74
verb_edges_dist_1	13.45	5.44	12.48	6.30
upos_dist_CCONJ	4.17	1.28	4.51	1.61

Table 3: The 10 features that, maximally varying in 'loser' essays, are more uniform in the 'winning' ones.

A further study was focused on the variability degree of linguistic features in the two essay groups. For each subset, we ordered the features by their increasingly coefficients of variation; then, we calculated the difference between the two rankings in order to identify the features that were maximally uniformly distributed in 'better' essays as compared to the 'worse' ones (Table 3). It can be noticed that future verbs (*verbs_tense_dist_Fut*) are very uniformly distributed among 'better' essays. We

[5]http://linguistic-profiling.italianlp.it/

[6]https://universaldependencies.org/

have previously commented that their frequency is higher in the 'winners'; it proves again that natives interpret the use of complex verbal forms as an indicator of higher skills. Also parataxis distribution (*dep_dist_parataxis*) is quite uniform in 'winning' essays; however, its average value is higher in the 'loser' ones. It can be deduced that annotators prefer hypotaxis but this is not surprising: hypotactic periods are more structured and elegant and require refined abilities to be built. The same evidence is given on the morphosyntactic level (*upos_dist_CCONJ*), since 'better' compositions include 0.34% less coordinating conjunctions. Curiously, 'better' essays have, on average, 0.1% more foreign terms (*dep_dist_flat:foreign*); this may suggest that annotators appreciate these expressions. Finally, it is worth highlighting a higher and more uniform percentage of verbs with few modifiers in the 'winning' essays (*verb_edges_dist_0*, *verb_edges_dist_1*).

4.2 Students Errors Impact

The last analysis was aimed at assessing whether and in what measure students errors impact on human judgments. We counted the pairs of essays whose 'winning' composition had a lower number of errors, those in which the 'loser' one had more mistakes and those with an equal number of errors. We noticed that essays with fewer errors had won in 56% cases, reaching the 79% if including pairs with the same number of errors. This procedure gave a first empirical answer to our starting question: errors substantially affect human assessment.

At this point, we focused on error categories to identify which ones affect more the perception of writing quality. For each category, we calculated the average number of errors and their standard deviation in both subsets; then, relying on Wilcoxon rank sum test, we found out that grammatical and orthographic mistakes vary significantly between the two groups (Table 4). As expected, 'loser' essays have, on average, 1.29 more grammatical errors and 0.85 more orthographic errors. It is worth to add that orthographic mistakes variation ($p - value = 0.007$) is more significant than the other ($p - value = 0.029$). This could mean that natives judge deviations in orthography worse than those in grammar. Once again, our findings are in line with Barbagli et al. (2015): *Usa una corretta ortografia* ("Use correct orthography") is the 2nd 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 ($p - value = 0.581$) and punctuation errors ($p - value = 0.617$) are probably due to their scarce amount in the analysed essays.

Category	'Winning' essays		'Loser' essays	
	Avg.	SD	Avg.	SD
Grammar	3.28	5.516	4.57	6.126
Orthography	3.18	4.517	4.03	4.826

Table 4: Error categories whose average varies significantly between the two subsets.

5 Conclusions

We presented a pilot study towards the identification of the linguistic features that are own of well written perceived essays. We collected Italian natives' preferences on 100 pairs of essays written by L1 students, that we analysed in terms of linguistic profiling and errors distribution. Our results reveal an interesting correspondence between annotators' judging criteria and writing instructions that L1 learners receive by teachers. Our findings could be interpreted as an indicator of the reliability of our data and, more in general, could suggest the effectiveness of crowdsourcing methods to quickly build large and reliable datasets. Considering the lack of Italian corpora of graded essays, such datasets could be valuable resources for the development of Computer-Assisted Learning Systems.

The limited size of our dataset certainly reduced the amount of results. Thus, we have to expand it (i) by collecting more annotations for the already existing surveys and (ii) by creating and distributing new surveys in order to gather judgments on new pairs of essays. Analysis on the enlarged dataset could provide more features that are own of good essays. Following the model of Miaschi et al. (2021), we could use the results to train a classifier that, given a pair of essays, recognizes the best written one.

The tool would not presume to replace teachers, but it could be a valuable teaching aid. Students could use it to get an immediate and preliminary self-assessment on their written productions so as to better understand their mistakes and hopefully avoid repeating them. Such tools can be very useful if integrated into educational processes based on distance learning paradigms, which need adequate technological infrastructures to be really efficient.

References

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

Alessia Barbagli, 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.

Alessia Barbagli, 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. European Language Resources Association (ELRA).

Dominique Brunato, 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. European Language Resources Association (ELRA).

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

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics*, pages 200–207.

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

Alessio Miaschi, 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.

Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, pages 145–172.

Hwee Tou Ng, 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. Association for Computational Linguistics.