A NLP-based stylometric approach for tracking the evolution of L1 written language competence

Alessio Miaschi°*, Dominique Brunato* & Felice Dell'Orletta*

° Dipartimento di Informatica, Università di Pisa | Italia

* Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR), ItaliaNLP Lab, Pisa | Italia

Abstract: In this study we present a Natural Language Processing (NLP)-based stylometric approach for tracking the evolution of written language competence in Italian L1 learners. The approach relies on a wide set of linguistically motivated features capturing stylistic aspects of a text, which were extracted from students' essays contained in CItA (Corpus Italiano di Apprendenti L1), the first longitudinal corpus of texts written by Italian L1 learners enrolled in the first and second year of lower secondary school. We address the problem of modeling written language development as a supervised classification task consisting in predicting the chronological order of essays written by the same student at different temporal spans. The promising results obtained in several classification scenarios allow us to conclude that it is possible to automatically model the highly relevant changes affecting written language evolution across time, as well as identifying which features are more predictive of this process. In the last part of the article, we focus the attention on the possible influence of background variables on language learning and we present preliminary results of a pilot study aiming at understanding how the observed developmental patterns are affected by information related to the school environment of the student.

Keywords: Diachronic Evolution of Written Language Competence, Natural Language Processing, Italian Learner Corpus, Stylometry, Learners' errors, Machine Learning

Miaschi, A., Brunato, D., & Dell'Orletta, F. (2021). A NLP-based stylometric approach for tracking the evolution of L1 written language competence. *Journal of Writing Research, 13*(1), 71-105. https://doi.org/10.17239/jowr-2021.13.01.03

Contact: Alessio Miaschi, Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR), via G. Moruzzi 1, 56124, Pisa | Italia – alessio.miaschi@phd.unipi.it

1. Introduction

Over the last fifteen years, there has been a growing interest to exploit the potential of Natural Language Processing (NLP) tools and machine learning methods in the context of language development, with the aim of characterizing the properties of learners' language and how it evolves over time, across modalities and stages of acquisition. A similar concern has been paid to turn theoretical considerations into educational applications, such as Intelligent Computer-Assisted Language Learning (ICALL) systems (Granger, 2003) and tools for automatically scoring learners' writing with respect to language proficiency and writing quality (McNamara et al., 2015; Deane and Quinlan, 2010). Two main ingredients stand at the core of this research: the availability of large digitized corpora of authentic texts produced by learners, which make it possible to complement theoretical underpinnings with corpus-driven evidence, and the reliability of language analyses generated by computational tools that allow quantifying and evaluating the impact of a large number of linguistically–motivated indices considered in the literature as proxies of language development (Crossley, 2020).

Moving in this framework, this paper introduces a NLP–based stylometric approach to model the evolution of written language competence in Italian L1 learners. According to the core assumptions of computational stylometry, formal properties of a text characterizing its style can reveal underlying traits about the author, e.g. in terms of gender, age, ethnicity, as well as language proficiency (Daelemans, 2013). However, while traditional stylometric techniques are typically based on a close set of ad-hoc linguistic features selected according to a specific task in mind (e.g. authorship attribution, authorship verification, gender classification), our approach relies on a wide set of linguistically motivated features extracted from students' essays, which have already shown to be effectively involved in several scenarios, all related to modeling the 'form' of a text, rather than the content: from the prediction of human judgments of perceived linguistic complexity (Brunato et al., 2018) to the automatic identification of the native language of a speaker based on their productions in a second language (L2) (Malmasi et al., 2017).

The proposed approach is developed and tested on texts contained in the CItA (Corpus Italiano di Apprendenti L1) corpus, the first longitudinal corpus of essays written by Italian L1 learners enrolled in the first and second year of lower secondary school (Barbagli et al., 2016). As stated by their creators, this two-year period is considered as crucial for the development of written language, which undergoes remarkable changes both in the way students write and in how they approach writing, as a consequence of being exposed to a more formal way of teaching from the first to the second year of lower secondary school. The longitudinal nature of the corpus, complemented with the emphasis on the

importance of the learning period under investigation, makes CItA particularly suitable to test the effectiveness of a computational model of writing development in L1 learners. Specifically, we decompose this problem into two main research questions:

- Is it possible to track the individual learning trajectory in writing by automatically predicting the chronological order of two essays written by the same student at different times?
- Which typologies of language phenomena contribute more to the prediction task and how they change according to different temporal spans?

The article is structured as follows. In the next section we present related works in the literature which have approached the problem of modeling language development using NLP techniques and corpus-driven insights. Although our contribution is focused on writing development, in this section we also consider some major studies which have addressed the acquisition of spoken language in preschool children using a similar methodological framework. In Section 3 we provide an overview of the approach devised to answer the two main research questions, while in Sections 4 and 5 we present the corpus of essays and the set of linguistic features on which our study is based. Discussion and results of experiments are reported in Section 6 and Section 7. Finally, in the last part of the article (Section 8) we focus our attention on the possible influence of background variables on language learning and present results of a pilot study in which we try to understand how the observed developmental patterns of writing in high–school age are affected by information related to school environment.

2. Related Works

In this section we take a closer look at studies in the literature which have relied on data–driven approaches, complemented with NLP–based analyses at different degrees of sophistication, to track the process of language development, both in spoken and written language. In the context of child language acquisition, a first line of research has focused on modeling the development of syntactic abilities in preschool children using data from the CHILDES database (MacWhinney, 2000) and a variety of features derived from a semi- or fully-automatic process of linguistic annotation. The CHILDES corpus contains transcripts of spoken interactions in natural settings involving children of different ages for over 25 languages, which makes it a reference corpus for empirical research on language acquisition. Based on a subset of utterances from English-speaking children (age 1-6), which were automatically annotated for syntactic dependency relations, Sagae et al. (2005) demonstrated that the hand-crafted calculation of the Index of Productive Syntax (*IPSyn*)[1] (Scarborough, 1990) can be effectively automated using features extracted from the sentence parse tree, in addition to information related to Part-of-speech

(POS) tagging. In a similar vein, Lu (2009) proposed a heuristic-based approach to automatically assign a score of syntactic complexity to children's utterances according to a revised version of the D-Level Scale (Covington et al., 2006), a seven-step developmental level scale based on empirical observations about the emergence of increasingly more complex constructions from the child speech literature. In this case too, a corpus of utterances from CHILDES was automatically analyzed with a state-of-the-art English parser to allow the extraction of the grammatical structures contained in the reference scale.

The main lesson from these studies was that NLP techniques can be used in a reliable way to help automate the laborious computation of expressive metrics for child language development. However, a more challenging step was tackled by Lubetich and Sagae (2014), which proposed a completely data-driven approach to measure syntactic development without the need of previously designing the sophisticated inventory of grammatical structures associated to a given metric. In this study, a corpus of transcripts of children from 1 to 8 years was syntactically annotated and automatically assigned with its IPSyn score. Then, for each transcript, the IPSyn score was associated with a set of language-independent features extracted from text (e.g. unigrams of parts-of-speech, unigrams of syntactic dependency labels) and deliberately meant to capture information about the syntactic structure of children's sentences. The hypothesis was that if the IPSyn scores could be predicted from these generic vectors, the selected features would be at least informative enough for tracking child language development as the inventory of IPSyn structures. Experiments were performed using a Support Vector Machine (SVM) regression model. The results showed a high correlation between predicted and real IPSyn scores, supporting the hypothesis that simple parse tree features are as indicative of language development as sophisticated language-dependent metrics. The authors also tested the data–driven approach on an age prediction task in which the regression model was trained to predict the age at which an unseen child transcript was produced, using the feature vector extracted from their other transcripts available in training. The underlying idea is that child language development could be better approached with age, rather than with a metric score, on the assumption that language acquisition (at least in a typical setting) evolves monotonically over time.

The rapid and remarkable changes child language undergoes before age five justify the amount of research for the earliest stages of acquisition, which is the framework underlying all the aforementioned studies. However, under the assumption that linguistic competence keeps growing during the school years as a result of explicit literacy instruction (Karmiloff-Smith, 1986; Kellogg, 2008; Durrant et al., 2020) and memory-related constraints (McCutchen, 2011), research on "later language acquisition" has gained increased attention prompted by the awareness that "becoming a native speaker is a rapid and highly efficient process but becoming a proficient speaker takes a long time" (Berman, 2004). Also in this

scenario, which is the focus of our study, corpus-based approaches complemented with linguistically-informed indices (semi)-automatically extracted from text have started being applied to track the development of writing skills throughout the school years. Note that if, in the case of spoken language, the growth is tracked as a function of age, the development of writing skills is typically addressed as a function of increasing grade level, both in elementary and middle school children and in high school and college level students (Crossley et al., 2011). Inspired by the Multi-Dimensional Analysis (MDA) pioneered by Douglas Biber, which assumes that "linguistic features from all levels function together as underlying dimensions of variation" (Biber, 1993), Chipere et al. (2001) applied this framework to investigate first language development during the school years. This study examined a large corpus of 899 graded essays written by school children (aged 8 to 15) with the aim of assessing the relationship between vocabulary diversity and age and level of linguistic ability. The former was operationalized in terms of a normalized version of type–token ratio (TTR) to account for the effect of text length. Results showed that vocabulary diversity is, in fact, correlated with age and ability level, although with few exceptions involving the transitions between middle and high school grades (i.e. 11 and 14 years). With this respect, the authors recognized that vocabulary diversity is only one of the factors qualifying writing ability and that an index like TTR could attribute lower scores to essays in which pupils intentionally use repeated words not because they don't have enough lexical knowledge but to produce a more coherent discourse.

Recent developments in computational linguistics methods and machine learning techniques have granted researchers the opportunity to assess large corpora of graded essays to examine overall writing ability and its development. With the aid of the automatic tool *Coh-Metrix*[2], Crossley et al. (2011) enlarged the analysis to several linguistic domains and examined to what extent essays written at various grade levels can be distinguished from one another using a number of linguistic features related to lexical sophistication (i.e., word frequency, word concreteness), syntactic complexity (i.e., the number of modifiers per noun phrase), and cohesion (i.e., word overlap, incidence of connectives). The main findings show that high school and college writers develop different linguistic strategies as a function of grade level and that even in advanced writers, lexical and syntactic constructions continue to develop. In contrast, as the grade increases, writers tend to produce fewer cohesive devices, which is interpreted as a tendency towards a more elaborate and complex discourse composition. Similar conclusions are reported by McNamara et al. (2010; 2015), which relied on the same tool to examine the degree to which high- and low-proficiency essays rated by experts can be predicted by linguistic indices of cohesion, syntactic complexity, the diversity of words used by the writer, and characteristics of words. The study showed that the three most predictive indices of essay quality were syntactic complexity, lexical

diversity and word frequency but, interestingly, no indices of cohesion correlated with essay ratings.

As expected, a large part of empirical studies based on NLP approaches and machine learning techniques has been carried out with respect to the English language and focused on high school and college learners. However, more recently other L1s and age samples have been addressed. In this respect, the recent study by Weiss and Meurers (2019) is particularly relevant for our research. This study is concerned with writing development in German speaking students across elementary and secondary school through a linguistically-informed classification approach. Using a wide set of linguistic measures modeling text complexity and accuracy, together with error rate and background information on topic essay and school tracks, their best performing model was able to reach an accuracy of 72.68% in predicting the correct grades of students according to a fourth-level classification; notably, the model using only linguistically informed features, without any metadata information, performs almost at the same level. A fine-grained analysis of the contribution of the individual features also revealed that writing acquisition in initial grades is best characterized in terms of accuracy development, while the upper stages of secondary school exhibit an increased linguistic complexity, in particular in the domains of lexis and syntactic complexity at the phrasal level. These findings have been further confirmed by a similar study by Kerz et al. (2020) carried out on the same corpus, which still focused on the predictive role of language complexity features to tracking writing development but obtained through a sliding window technique, in order to monitor the progression of complexity within a text.

While our study shares some characteristics with the approach and goals presented in these latter works, it has the main novelty of using genuine longitudinal data rather than an approximation of it as a function of grade level. Given the greater efforts in terms of time of collecting longitudinal corpora, this perspective is much less investigated than the one based on cross-sectional data, yet it offers the possibility of studying differences and learning trajectories which pertain to individuals rather than to the overall group characteristics.

3. Our Approach

In order to track how written language competence evolves in the two considered school grades, we ask whether the writing development curve of a student can be automatically learned. We model the problem as a binary classification task in which a machine learning classifier has to predict the relative order of two essays using a wide set of linguistically motivated properties automatically extracted from the L1 learner's essays contained in the CItA corpus.

The classifier uses a Linear Support Vector Machine (LinearSVM) as machine learning algorithm, i.e. a discriminative algorithm that, given labeled training data,

outputs an optimal hyperplane which categorizes new examples. We rely on LinearSVM rather than more powerful learning algorithms, such as Recurrent Neural Networks (RNNs), in order to obtain meaningful explanations when the classifier outputs its predictions, so as to anchor the observed patterns of language development to explicit linguistic evidence. To prevent overfitting, we train and test our model in a cross-domain manner, using essays of students from different schools during the training and testing phase. Doing so, the algorithm is tested not only on essays written by different students, but also on students coming from different schools.

We further extract and rank the feature weights assigned by the LinearSVM in order to understand which typology of linguistic features contributes more to the classification task. The underlying assumption is that the higher will be the weight associated with a specific feature, the higher will be its importance in solving the classification task and, consequently, in tracking the students written language evolution.

Finally, to provide first insights into the possible influence of background variables on predicting writing development, we ran the same binary classification task distinguishing students enrolled in the center and suburban schools and we assessed the confidence of our classifier in the two scenarios. Since the confidence reflects the uncertainty of the model estimates (i.e. the higher the confidence the easier the prediction was for the classifier), this measure can be viewed as a mean to approximate the degree of changes in the learning curve of each student. That is, we can assume that the classifier is more confident when the two essays for which the relative order has to be predicted show greater differences with respect to the considered features.

In what follows, we first introduce the two main ingredients of our approach, namely the corpus and the set of linguistic features. We then describe the set-up of the experiments and discuss the obtained results in light of the main research questions of the study.

4. The CItA Corpus

As shown by studies presented in Section 2, the availability of authentic texts produced by language learners is of pivotal importance. Such resources can differ according to the modality (i.e. written texts or speech transcriptions), the typologies of learners considered (e.g. preschool children, first or second language students), the goals of analysis (e.g. theoretical studies or development of educational applications). For the purpose of our study, we relied on CItA (Corpus Italiano di Apprendenti L1), a longitudinal corpus of essays written by the same students in the first and second year of lower secondary school (Barbagli et al., 2016). This makes the corpus particularly suitable to track the evolution of L1 written language competence over the time. The corpus 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[3]-IPS (Association for the Evaluation of Educational Achievement) activities (Lucisano, 1984). As stated by their creators, the collection of essays in CItA was motivated by two underlying hypotheses. The former is that students' competence in writing undergoes a variety of relevant changes from the first to the second year of lower secondary school, as a consequence of being exposed to a more formal teaching. The latter is that the development of written language competence could be related to background variables of students, such as the city area where the school is located (historical center or suburbs), the language(s) the students speak at home or their parents' employment. To make it possible to explore the effects of these variables, the CItA essays were collected from seven different schools located in Rome, three of which in the historical center and four in suburbs. In addition, all students whose essays are included in the final corpus were asked to answer a questionnaire of 34 questions to obtain information about their biographical, socio-cultural and sociolinguistic background. For example, they were asked to provide biographical information such as the language(s) the students usually speak at home, when and where they were born, their parents' education, etc.

4.1 Corpus Description

The corpus contains a total of 1,352 essays written by 156 students (see Table 1). The essays belong to five textual typologies, which reflect the different writing prompts students were asked to respond: reflexive, narrative, descriptive, expository and argumentative.

In addition, a prompt common to all schools was also assigned at the end of each year. Specifically, at the end of second year, students were asked to respond to the Italian version of Task 9 of the IEA-IPS study (Lucisano, 1984; Visalberghi and Costa, 1995), i.e. a letter of advice to a younger student on how they should write in order to get good grades at high school; at the end of the first year, they were presented with a modified version of Task 9 still with the same aim. Table 2 shows examples of prompts given to the students according to the different typologies.

As shown in Table 3, there are some differences over the two years and the seven schools. First of all, it can be noted that the number of prompts differs among the seven schools: teachers of the schools located in the city center tend to give a higher number of prompts than their colleagues in the suburban schools. Secondly, if reflexive prompts are the most frequent textual type in the two years, from the first to the second year the distribution of narrative prompts are halved while the expository and argumentative ones are doubled. This different distribution is a consequence of the approach adopted by teachers to teach writing: writing a narrative essay is considered as a simpler 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 competencies are required (Kellogg, 2008; Barbagli, 2016).

Table 1. Composition of the corpus.

	First year			Second year		
	School	Students	Essays	School	Students	Essays
Center	1	25	123	1	25	108
	2	27	143	2	28	130
	3	24	138	3	23	117
Suburbs	4	21	58	4	22	62
	5	19	77	5	19	64
	6	24	66	6	24	146
	7	13	64	7	14	56
Total	7	153	669	7	155	683

Table 2. Prompt examples according to the different typologies.

Typology	Prompt example
Reflexive	What's your attitude regarding the reading activity?
Narrative	Narrative essay in which you describe an episode of bullying
Descriptive	Describe a primary school teacher you are particularly close to
Expository	Write a news story that the media has been dealing with recently
Argumentative	Mobile phones in class: what do you think about it and how do you think it could be solved?
Common Prompt	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.

Table 3. Distribution of typologies of prompts.

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

4.2 Error Annotation

One of the characteristics that mostly distinguishes CItA from other corpora of L1 Italian learners, such as those described in Marconi (1994), is that the essays were annotated according to different types of linguistic errors occurring in text. Error annotation is a challenging issue since it assumes that a deviation from a linguistic norm is occurring. However, there is no agreed-upon consensus about how to interpret a particular error in the development of writing ability since an error can express a real mistake, a deviation from convention, or a developmentally appropriate construction (Wilcox et al., 2014). Moreover, the annotation of errors in L1 corpora is a much less common practice than in L2 corpora, where this level of information is typically used to investigate the properties of interlanguage (Brooke and Hirst, 2012) or as a reference resource for automatic error detection and correction tasks (Dahlmeier et al., 2013). In the absence of a L1 error taxonomy already available for the Italian language, for the annotation of CItA a new annotation scheme was introduced to mark CItA essays with learners' errors. This is inspired by Berruto's definition of "neo-standard Italian" as linguistic norm (Berruto, 1987) following the literature on the evaluation of written skills of L1 Italian learners (Visalberghi and Costa, 1995; De Mauro; 1983; Colombo, 2010).

As shown in Table 4, it is a three-level schema including grammatical, orthographic and lexical errors, which makes it also similar to already existing schemes in other languages (e.g. Granger (2003) for French as a second language). Following the annotation format proposed by Ng et al. (2013), CItA texts were annotated as follows:

[...] scapparono al piano di sopra e dal <M t="200" c="buio">buglio</M> <M t="113" c="spuntò">spuntarono</M> un esercito [...]

([...] they ran away upstairs and from the darkness an army appeared [...])

where the textual span of error is marked by <M> and </M> (Mistake), the attribute t (type) is the macro-class and subclass of error, and c (correction) reports the corrected form. In the reported example, there is a generic orthographic error (the word *buglio* instead of the correct one *buio* (`darkness')) and a grammatical mistake concerning Subject-Verb agreement (the third person plural of the verb *spuntare* (`to appear') instead of the required third person singular).

The annotation was manually performed by a teacher of lower secondary school and revised by two undergraduate students in digital humanities, who were adequately trained on the task annotation guidelines. Inspecting the statistical distribution reported in Table 4, it can be noted that the majority of errors has a statistically significant variation over the two years thus showing that several common trends in the development of writing competence occur during the transition from the first to the second year. In both years (rows Total) orthographic and grammatical errors are the most frequent ones (47.63/44.72% and 46.41/48.7% respectively) while lexical errors are far less (about 6%). More specifically, the most frequent errors affect the area of orthography without distinction into specific typologies (i.e. the class Other) (22.32%) followed by the erroneous use of verb tenses (11.26%), the unclassified grammatical errors (6.37%) and the erroneous use of prepositions (6.6%). Interestingly, while almost all categories of errors are similarly distributed in the two years, essays by II-year students exhibit a significantly higher percentage of errors affecting verb morphology, and in particular incorrect tense inflection. This kind of error could be related to an increased awareness by older students that "good" writing also involves organizing their ideas in a coherent way within sentences and larger passages, using the available lexical and morpho-syntactic devices of the language. In this respect, appropriate shifts in verb tenses and moods are one of the means that allows the writer to convey temporal relationships between expressed events and actions. However, this is an ability that develops across school-age years and previous studies in the literature indicate that incorrect verb inflection is still one of the most common errors in adolescent student writing (Wilcox et al., 2014).

Table 4. Error annotation schema.

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

Note. Errors varying significantly over the two years (i.e. $p < 0.05$) are marked with an asterisk.

4.3 Linguistic Annotation

To allow the extraction of linguistic features used as predictors of writing development in the classification experiments, the CItA corpus was firstly automatically annotated using UDPipe (Straka et al., 2016), a NLP pipeline carrying out basic pre-processing steps, i.e. sentence splitting and tokenization, POS tagging, lemmatization and syntactic parsing, according to the Universal Dependencies (UD) annotation framework (Nivre et al., 2016). Although we used a state–of–the art pipeline, it is well acknowledged that the accuracy of statistical parsers decreases when tested against texts of a different typology from that used in training (Gildea, 2001). In this respect, learners' data are particularly challenging for general–purpose text analysis tools since they can exhibit high deviations from correct and standard language (Berzak et al., 2016). For instance, missing or anomalous use of punctuation (especially in 1st grade prompts) could already impact on the coarsest levels of text processing, i.e. sentence splitting, and thus may affect all subsequent levels of annotation. Nonetheless, if we can expect that the predicted value of a given feature might be different from the real one (especially for features extracted from more complex levels of annotation such as syntax), we can also assume that results will be consistent, at least when parsing texts of the same domain. The validity of this claim has been shown in other studies relying on engineered features similar to ours for classification or linear regression analyses. For instance, Dell'Orletta et al. (2011) proved that the values of a set of morpho-syntactic and syntactic dependency features are comparable when extracted from a gold (i.e. manually annotated) and an automatically annotated corpus of the same domain (i.e. biomedical language). In a study aimed at investigating dependency distance minimization in English using a large diachronic corpus, Lei and Wen (2020) checked whether any possible errors from the parser significantly affected the results of their analysis. To this end, they manually revised the annotation of a subset of the automatically parsed corpus under investigation and correlated the values of their examined features (i.e. mean and normalized dependency distance) extracted from the automatically and the manually revised portion, obtaining very high correlation scores.

We applied a similar approach to our corpus in order to observe the impact of possible parsing errors on the reliability of the feature extraction process with respect to learner data. Specifically, we randomly extracted a few parsed sentences from both I and II-year CItA essays for a total of 800 tokens and we manually revised the output of the automatic annotation in every step. We then extracted all monitored features from the manually revised sentences and compared these values to the corresponding ones extracted from the automatically parsed sentences. The resulting Spearman's rank correlation coefficient between the two samples shows that, with the only exception of the distribution of parataxis relations (dep_dist_parataxis), linguistic features extracted from automatically annotated and manually revised sentences are extremely highly correlated (average = 0.93).

5. Linguistic Features

To extract the linguistic features from the automatically parsed corpus, we relied on Profiling–UD (Brunato et al., 2020), a multilingual tool specifically conceived to carry out linguistic profiling on corpora annotated in UD–style. UD is an ongoing project aimed at developing corpora with a cross-linguistically consistent annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective[4]. The choice of relying on UD–style annotation makes the process of feature extraction language–independent, since similar phenomena are annotated according to a common annotation scheme at morpho–syntactic and syntactic level of analysis.

Profiling–UD allows the computation of a wide set of features encoding a variety of lexical and grammatical properties of a text informed by the literature on linguistic complexity and language development. They range from superficial ones, such as the average length of words and sentences, to morpho–syntactic information concerning the distribution of parts-of-speech (POS) and the inflectional properties of verbs, to more complex aspects of syntactic structure deriving from the whole parse tree and from specific sub-trees (e.g. subordinate clauses.). The set of features is reported in Table 5 according to the level of annotation from which they derive.

By looking at the statistical distribution of features which turned out to vary in significant way[5], it can be noted, for example, that essays written in the second year are on average longer both in terms of tokens (i.e. I-year: 293 tokens; II-year: 345 tokens) and in terms of number of sentences for document (i.e. I-year: 13 sent/doc; II-year: 16 sent/doc). II-year essays also contain a lower percentage of conjunctions, pronouns (especially clitic and personal ones), and a higher percentage of prepositions and nouns with respect to the essays of the first year (Table 6). These statistically significant differences suggest that II–year students possibly exploit more the pro-drop potentiality of the Italian language in their writing, thus making less use of overt pronouns. At syntactic level (Table 7), this speculation seems to be corroborated by the lower distribution in second year's essays of syntactic relations linking a nominal subject (either headed by a noun phrase or realized as a pronoun) to its verbal head (dep_nsubj). Moreover, when the subject is overtly expressed, it tends to be placed in the canonical position (i.e. left to the verb since Italian is a SVO language), especially by younger writers. We also observe an increase in the usage of complex sentences, i.e. sentences characterized by deeper syntactic trees, as well as in the use of subordination, as shown by the higher distribution of adnominal clause modifiers such as relative clauses, across the two years.

Table 5. Linguistic features used in the experiments.

Level of Annotation	Linguistic Feature	Label
Raw Text	Sentence Length	tokens_per_sent
	Word Length	char_per_tok
	Document Length	n_sentences
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	Distribution of UD and language–specific POS	upos_*, xpos_*
	Lexical density	lexical_density
	Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Dependency Parsing	Depth of the whole syntactic tree	parse_depth
	Length of dependency links and of the longest link	links_len, max_links_len
	Average length of prepositional chains and distribution by depth	prepositional_chain_len, prep_*
	Clause length (n. tokens/verbal heads)	token_per_clause
	Order of subject and object	subj_pre, subj_post, obj_pre, obj_post
	Verb arity and distribution of verbs by arity	verb_edges, verb_edges_*
	Distribution of verbal heads per sentence	verbal_head_sent
	Distribution of verbal roots	verbal_root_perc
	Distribution of dependency relations	dep_*
	Distribution of subordinate and principal clauses	principal_prop, subord_prop
	Length of subordination chains and distribution by depth	subord_chain_len, subord_*
	Relative order of subordinate clauses	subord_post, subord_prep

These features reveal that II-year essays contain more complex syntactic constructions, thus anticipating a tendency towards an increased linguistic complexity which has been observed in texts written by high school adolescents across several languages (Berman, 2017).

Table 6. Distribution of major morpho–syntactic features varying significantly between the two school years.

Feature	I year (%)	II year (%)
Conjunctions	6.88	6.38
Determiners	13.86	14.12
Preposition	10.53	11.21
Pronouns	8.97	8.04
Clitic pronouns	4.58	4.08
Personal pronouns	1.58	1.2
Nouns	16.02	16.38

Table 7. A subset of syntactic features varying significantly between the two school years.

Features	I year (%)	II year (%)
preverbal subjects	84.19	82.57
postverbal subjects	15.81	17.14
preverbal objects	35.69	30.39
postverbal objects	64.31	69.61
nominal subjects (dep_nsubj)	5.59	5.04
passive subjects (dep_nsubj:pass)	0.19	0.28
adnominal clause modifiers (dep_acl)	0.53	0.62
copular constructions (dep_cop)	2.13	1.89
coordination (dep_cc)	4.38	4.14
parse tree depth	4.589	4.716

While the distribution of verbs is almost similar between the two years (i.e. around 13%, without significant variation), the use of verbal morphology changes from the first to the second year (Table 8). As could be expected, the indicative mood is predominant in all essays, although in the second year, students start using in a slightly higher percentage also more complex moods, such as the subjunctive. Instead, a greater variation affects the use of tenses, especially the imperfect one, which decreases significantly in the second year. On the one hand, this could be expected since imperfect indicative verbs are easier than other past tenses of the Italian verbal morphology. On the other hand, this variation might be related to the different type of essays assigned in the two years. In fact, in the second year the typology of narrative essays, for which is commonly required the use of imperfect tenses, is less predominant across prompts. In this regard, also the more extensive use of first singular and plural person verbs in essays written by younger students is indicative of a more subjective writing style.

Table 8. Distribution of verbal morphology features (mood, tense and person) varying significantly between the two school years.

Features	I year (%)	II year (%)
Indicative mood	94.83	92.60
Subjunctive mood	2.61	3.31
Imperfect tense	16.48	10.99
Present tense	42.36	49.28
Verbs-1PerSing	15.22	13.55
Verbs-1PerPlu	6.96	5.25

6. Tracking the evolution of written language competence

Our first research question was aimed to explore whether it is possible to automatically track the development of students' writing competence across time. We model this problem as a classification task, starting from the assumption described in Richter et al. (2015): given a set of chronologically ordered essays written by the same student, a document d_j should show a higher quality level with respect to the ones written previously (d_i). Thus, given two essays d_i and d_j written by the same student, we want to classify whether $t(d_j) > t(d_i)$, where $t(d_i)$ is the time in which the document d_i was written.

For this purpose, we built a classifier operating on morpho-syntactically tagged and dependency parsed essays which assigns to each pair of documents (d_i, d_j) a score expressing its probability of belonging to a given class:

1 if $t(d_j) > t(d_i)$, 0 otherwise.

For each pair of essays, we built an E event:

$E = V_i + V_j + (V_i - V_j)$

where V_i and V_j are, respectively, the feature vectors of the first and second essays, and V_i V_j is the vector difference between them.

Vectors are composed by the values of multi-level linguistic features both automatically extracted, as shown in Sec. 3, and manually annotated (i.e. features related to the error annotation) in CItA. As previously mentioned, the classifier uses linear Support Vector Machines (SVM) as the machine learning algorithm.

We split all texts of the CItA corpus into four sets, pairing essays written by the same students considering all the possible temporal spans at the same time (All essays) and considering only essays written at a distance of one month (1 month), one year (1 year) and two years (2 years). Table 9 summarizes the statistics of the four datasets. We evaluated the system with a 7-fold cross validation in which every fold is represented by a different school. It follows that in each experiment the test set is composed by documents which are not included in the corresponding training set. Each line of the training and test sets follows this structure:

Student code, Label, E event

where Student code is an identifier assigned to the student, Label could be 1 or 0 depending on the two essays' order and E event is the feature event associated with the two essays.

Table 9. Number of samples/E events within each dataset.

Temporal span	Number of samples
All essays	7,228
1 month distance	1,308
1 year distance	348
2 years distance	208

Three different sets of experiments were devised to test the performance of our system, which differ with respect to the number and type of linguistic features extracted for each essay. In the first set (#1) we used only the lexical, morpho-syntactic and syntactic features extracted from the parsed corpus. In the second set of experiments (#2) we added to them a set of features related to word frequency (word frequency class), which was measured as the average class frequency of all lemmas in the document. The class frequency was computed for each lemma and form exploiting the itWAC (Italian Web as Corpus) corpus[6] as follows:

$$C_l = \lfloor log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$$

$$C_{wf} = \lfloor log_2 \frac{freq(MFF)}{freq(CF)} \rfloor$$

where MFL and MFF are, respectively, the most frequent lemma and word form of the corpus, and CL and CF are the considered lemma and word form. In the third experiment (#3) we expanded our set of linguistic features with those related to the distribution of the different kinds of errors annotated in CItA (Section 4.2): grammatical errors; orthographic errors; lexical errors and punctuation errors.

In order to verify the effectiveness of our model, we compared our classification results with the ones obtained with a baseline computed with a LinearSVM that takes as input the average sentence length of the essays for each sample pairs. Classification results are reported in Table 10.

Table 10. Cross-school results (in terms of weighted accuracy standard deviation) for the three sets of experiments.

	Samples	#1	#2	#3	Baseline
All essays	7,228	0.53±0.08	0.55±0.09	0.58±0.09	0.45±0.06
1 month	1,308	0.49±0.03	0.50±0.05	0.54±0.04	0.50±0.03
1 year	348	0.54±0.15	0.63±0.09	0.65±0.15	0.61±0.12
2 years	208	0.66±0.16	0.71±0.15	0.75±0.14	0.40±0.14

As a general remark, we observe that the larger the temporal span between the tested documents, the higher the achieved accuracy. Not only does this suggest that the pairs of essays written by each student at more distant times exhibit a quite divergent linguistic profile – which makes the classification task easier –, but also that linguistic patterns underlying writing development are consistent across students and schools. Remember indeed that in all the experiments the classifier is tested not only on essays written by different students, but also on students coming from different schools. If we compare the results obtained considering the 1 month and 2 years time intervals we can notice an improvement of 20% in terms of accuracy scores. As expected, accuracy scores in the 1-month temporal span are comparable with those obtained with the simple baseline, proving that the complexity of this task does not allow to obtain reliable results when considering excessively short time intervals.

Focusing on the three different set of experiments, we can see that the results tend to improve as more features are used for classification. In particular, the contribution of vocabulary–related features operationalized in terms of word frequency is particularly effective when considering essays written at a long-term distance, such as 1 year or 2 years. In addition, differently from what is reported in Richter et al. (2015), we observe a general improvement when lexical, morpho–syntactic and syntactic features are complemented with features related to the distribution of errors made by students.

Table 11. Classification results using different sets of annotated error features.

Features	Two years distance
Grammatical errors	0.74
Orthographic errors	0.72
Lexical errors	0.70
Punctuation errors	0.68

In this respect, to have a better understanding of their contribution in the automatic classification, we repeat our experiments using the four sets of error-related features (i.e., grammatical, orthographic, lexical and punctuation) in a separate way. As shown in Table 11, the improvement of classification accuracy mainly depends on the presence of grammatical errors. Indeed, the accuracy obtained using only this typology of errors, in addition to general linguistic features, is even higher than the one obtained using the four sets of errors together (from 0.73% to 0.74%). These data are in line with the qualitative observations reported in Section 4.3 since grammatical errors, as well as orthographic errors, undergo a significant variation over the two school years, thus allowing the classifier to obtain better results.

7. Cross-Prompt Testing

As reported in Section 4, the assigned prompts are differently distributed over the two years. This observation may cast doubts on the effectiveness of our features to serve as real proxies of writing development rather than as prompt–related characteristics. To discard this hypothesis and verify whether the results we obtained generalize across prompts, we replicate the experiments in a cross-prompt scenario. In particular, we used the four datasets previously described (All essays, 1 month, 1 year and 2 years) and we performed the experiments with a cross–prompt validation strategy, i.e. testing the resulting model only on pairs of essays that have the same prompt. The new classification results are reported in Table 12. As we can notice, our model achieved better results with respect to the length baseline for all the datasets and according to the three sets of experiments, thus

allowing us to confirm that the system clearly generalizes across prompts and is actually modeling written language evolution rather than prompt-dependent characteristics.

Table 12: Cross-prompt results (in terms of weighted accuracy standard deviation) for the three sets of experiments along with total test samples size (Samples).

	Samples	**#1**	**#2**	**#3**	**Baseline**
All essays	2,662	0.64±0.04	0.64±0.04	**0.67±0.04**	0.52±0.01
1 month	532	0.47±0.02	0.46±0.05	**0.50±0.01**	0.48±0.04
1 year	128	0.53±0.05	0.53±0.05	**0.68±0.10**	0.65±0.16
2 years	119	0.82±0.04	0.84±0.05	**0.85±0.05**	0.48±0.01

8. Studying linguistic phenomena

The results obtained in the previous experiments showed that it is possible to predict the chronological order of two essays written by the same student by using features of different nature. This confirms that relevant transformations occur in L1 writing during the transition from the first to the second year of lower secondary school. However, very little has been said about the contribution of each single feature in the classification tasks. Since we showed that not all the linguistic features vary significantly during the 2-year temporal span, we can reasonably assume that within the set of our features, some of them are also more predictive than others for the classification. To better explore this question, we established a ranking of the most important features according to the different classification scenarios. To do this, we evaluated the importance of each linguistic property by extracting and ranking the feature weights assigned by the LinearSVM model that uses features of all categories (i.e. linguistic features, word class features and error–related ones).

Table 13 shows the rankings of the 20 most important features according to three considered temporal spans[7]. As we can see, error–related features acquire relevance as the temporal span increases: in the second classification experiment, where the task was to predict the chronological order of essays written at a distance of one year, three of the ten most significant features derive from error annotation. Similarly, in the third classification scenario, error-related features occur three times in top-ranked positions and one of them, i.e. omission of pronouns, is the first ranked one. The omission of pronouns in required contexts, complemented with their unnecessary use (i.e. error_pronouns_redundancy, 12th-ranked), could be indicative of the influence of spoken language phenomena on written texts by middle–school students, which is still pervasive even at longer temporal spans. At syntactic level, this seems to be confirmed by the occurrence of dislocated

dependencies (dep_dislocated) in the first position of the ranking derived by classifying the order of essays written at a distance of one year.

Table 13. Ranking of the first 20 features for three different temporal spans.

1 month distance	1 year distance	2 years distance
dep_punct	dep_dislocated	error_pronouns_omission
upos_AUX	xpos_PP	n_tokens
upos_X	xpos_BN	wfc-verbs-lemma
dep_aux	xpos_PD	n_prepositional_chains
upos_PUNCT	xpos_DD	aux_tense_Past
verbs_form_Part	aux_form_Fin	xpos_RI
dep_cop	error_preposition_omission_redundancy	obj_pre
upos_VERB	avg_lexical_errors	obj_post
verbs_form_Fin	error_vocabulary-erroneous-use	n_sentences
xpos_AP	n_sentences	dep_aux
dep_det:poss	dep_vocative	aux_1PerPl
xpos_SP	xpos_RI	error_pronouns_redundancy
dep_conj	wfc-nouns-lemma	verbs_num_pers_2PerSing
wfc-adjectives-lemma	n_prepositional_chains	aux_form_Ger
wfc-nouns-word	n_tokens	xpos_FB
verbs_3PerSing	aux_tense_Imp	dep_conj
dep_root	verb_edges_3	aux_tense_Imp
dep_acl:relcl	error_conjunctions-misuse	wfc-adjectives-word
dep_nsubj	error_full-stop-omission	error_monosyllabes-misuse-po'
dep_advcl	avg_punctuation_errors	wfc-nouns-word

According to the UD annotation tagset, this syntactic relation has the specific function of indicating fronted or postposed elements that do not fulfill the usual

core grammatical relations of a sentence, which is quite typical in speech. In addition to these features, what helped more the classifier in the same classification scenario is the different use of functional categories, specifically pronouns (xpos_PP, xpos_PD), negative adverbs (xpos_BN) and determiners (xpos_DD, xpos_RI).

Beyond error-related features, morpho-syntactic information still has a relevant role in classifying essays when the longest temporal span is considered. However, in this case, features related to verbal inflectional morphology (tense, mood and person) are more highly ranked than those concerning the distribution of core grammatical categories (see, e.g. aux_tense_Past, aux_form_Ger, aux_tense_Imp). This is in line with what we observed in the linguistic profiling section (Table 8), where differences concerning the use of verbal features in the two years were found to be statistically significant. Interestingly, with the exception of the words frequency class, lexical features do not seem to be particularly relevant and this allows us to confirm what already reported in Barbagli (2016), namely that vocabulary distribution, lexical density and TTR (Type Token Ratio) do not change significantly over the two school years.

9. Investigating relationships between writing competence and background information

The last part of this article presents the first results of a pilot study that we performed in order to explore the hypothesis put forth in Barbagli (2016) that there could be a relationship between the observed trends in the evolution of writing competence and the school environment of the students. This information was explicitly collected as one of the background variables of each student included in the corpus.

To this end, we inspect again the classification results by computing the confidence of our model (C_m), i.e., the measure that, as mentioned in Sec. 3, depicts the uncertainty of the classifier estimates. In particular, C_m can be defined as the variation between the two probabilities assigned by our classifier to each label (1 if $t(d_j) > t(d_i)$, 0 otherwise). On the assumption that the more confident the model in predicting the chronological order of essays written by a given student, the easier the classification task for that student, we can state that higher C_m values could be indicative of a greater evolution in student's writing competence.

Table 14. C_m values according to the two urban areas.

Urban area	1 month distance	Two years distance
Center	0.579	0.629
Suburbs	0.513	0.670

On the contrary, if we consider essays for which our classifier is less confident with its predictions, we can infer that the two essays do not present noticeable variations in their linguistic profile, although they were written in two different periods.

Specifically, we performed an experiment by computing the C_m values of our classifier for two different temporal spans (1 month distance and Two years distance) and then dividing the students according to the two different areas of Rome: historical center and the suburbs. As we can see in Table 14 there is no particular difference between the results. However, as the temporal span increases the C_m values for both urban areas show a slight improvement, in particular for the students of the suburban schools. This allows us, partly, to confirm that the evolution of writing competence is more evident for those students attending schools in suburbs, possibly because their entry level is lower, as suggested by the answers obtained from the questionnaires.

10. Discussion and Conclusion

The longitudinal nature of the CItA corpus allowed us to define a computational model to track the evolution of the written language competence in Italian as a first language, as well as to identify which linguistic features are more predictive of this evolution and how they change according to the considered temporal span.

As regards the first research question, the results obtained in the three experiments have demonstrated that linguistic features automatically extracted from text not only allow making explicit the relevant transformations occurring in L1 learners' writing competence but can be exploited as effective predictors in the automatic classification of the chronological order of essays written by the same student, especially at more distant temporal spans. Moreover, by testing our approach on a cross-prompt scenario, we show that the considered features capture markers of language evolution which are not related to the textual typology of the essay.

When training our model using also the twenty-six features related to error annotation, we obtained a general improvement in almost all cases. These results demonstrate that analyzing the diverse typologies of errors made by students in their texts is effective to capture aspects of the written language competence evolution. In this regard, we also noticed that the errors which allow the classifier to achieve a better accuracy are the grammatical ones. This could be due both to the larger amount of errors of this category (46.41% and 48.7% of the total in the first and second school year) and by the fact that grammatical errors, as well as orthographic errors, have a significant variation over the two school years, and thus they probably allow the classifier to obtain better results. Interestingly, we observed that, when significant, this variation does not always follow the expected developmental trend. That is to say, the total amount of errors of some categories

is higher in essays written by older students compared to younger students' essays. This is especially true for grammatical errors concerning the correct use of verb tenses and moods within text, suggesting that this is an ability that continues to develop across later school-age years. In fact, errors affecting the use of inflectional morphology have been reported among the most common ones also in high-school student writing (Wilcox et al., 2014).

Regarding the second research question, extracting the feature weights assigned by the linear model we were able to establish a ranking of the most important features according to different temporal spans. Changes of the resulting rankings in the different classification scenarios suggest that both linguistic and error-related features contribute in a different way according to time intervals. For instance, it was shown that features related to the error annotation acquire much more relevance as the temporal span increases, and this allows us to confirm that the errors made by the students are an indicative proxy to track the writing competence evolution, especially in the transition from the first to the second year.

In a similar fashion, we observed that the classifier is sensitive to changes affecting morpho–syntactic features, especially those related to the use of grammatical categories and to the inflectional properties of verbs: the latter were also found to change in a significant way when comparing the whole subcorpus of essays written in the first and in the second year. This gives additional evidence that mastering verbal morphology in a morphologically-rich language like Italian is an important skill that evolves in writing during the considered school years. This is also in line with the Weiss and Meurers (2019) study on German cross-sectional data, which showed that features belonging to morphological complexity play an important role especially in the development of secondary school writing. However, unlike Weiss and Meurers (2019) and Kerz et al. (2020), our analysis showed that features related to lexical sophistication do not seem to be particularly relevant for identifying the evolution of writing competence.

Lastly, we presented a pilot study in which we try to explore the relationships between the developmental patterns in writing and information about students' background variables. Although preliminary, the obtained results suggested that the student's learning curve varies according at least to the geographical area where the school is located. In fact, we saw that, when a higher temporal span is considered (e.g. Two years distance), the classifier is more confident about its decision for essays written by students who belong to suburban schools. These results go in the direction of what suggested in Barbagli (2016), namely that the evolution of writing skills is strictly related to the socio-cultural context inferred from background variables, and that these aspects affect the linguistic entry level of the students.

To conclude, we would like to draw attention to some of the perspectives that the presented study could enable, which are especially relevant in the field of NLP-based educational applications. Finding theoretically motivated methods to

monitor the learning growth of each student can support the learning assessment process by teachers, which could be a very demanding task especially in distance learning paradigms. Similarly, we believe that the new educational frameworks pose new challenges concerning students' engagement in virtual classes. As shown by Slater et al. (2017), a variety of linguistic features identified by NLP tools can be used as reliable predictors of affective states experienced by students, such as boredom, confusion, frustration, engaged concentration. With this respect, it would be interesting to explore potential correlations between the motivation and level of engagement shown by students and the linguistic properties turned out to be involved in modeling language learning so as to promote personalized teaching and learning strategies.

Last but not least, the proposed approach can enable comparative studies on the evolution of the written language competence from a cross-linguistic perspective. In fact, one of the main novelties of the proposed approach is that the linguistic features used as predictors of language learning were extracted from corpora annotated according to the Universal Dependencies (UD) framework. Since this annotation is inspired by 'universal' principles aiming at annotating similar constructions across languages in a consistent way, the process of feature extraction can be applicable to other learner corpora for all languages included in the UD project.

Notes

1. IPSyn is a sophisticated metric of child language acquisition, which scores children's utterances according to the distribution of more than 50 syntactic constructions (e.g. relative clauses, wh-questions with auxiliary inversion, propositional complements).
2. Coh-Metrix is a computational system for computing cohesion and coherence metrics in written and spoken texts (http://cohmetrix.com).
3. http://www.iea.nl
4. At the time of this manuscript, 183 UD treebanks for over 100 languages have been released.
5. The statistical significance for all features discussed in this Section was assessed using the Wilcoxon-rank-sum-test.
6. A 1.5 billion words corpus made up of texts collected from the Web (Baroni et al., 2009).
7. Rankings of the top 100 features, along with their corresponding weights, are reported in Appendix A.

References

Barbagli, A., (2016). Quanto e come si impara a scrivere nel corso del primo biennio della scuola secondaria di primo grado. Nuova Cultura.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., Venturi, G., (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), 88-95.

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E., (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation, 43, 209-226. doi: 10.1007/s10579-009-9081-4

Berman, R.A., (2004). Between emergence and mastery. The long development route of language acquisition. Trends in language acquisition research vol. 3, Benjamins, Amsterdam Philadelphia. doi: https://doi.org/10.1075/tilar.3.05ber

Berman R.A. (2017) Language Development and Literacy. In: Levesque R. (eds) Encyclopedia of Adolescence. Springer, Cham. https://doi.org/10.1007/978-3-319-32132-5_19-2

Berruto, G., (1987). Sociolinguistica dell'italiano contemporaneo. Volume 33. Roma, Carocci.

Berzak, Y., Kenney, J., Spadine, C., Wang, J.X., Lam, L., Mori, K.S., Garza, S., Katz, B., (2016). Universal Dependencies for Learner English, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 7-12, 2016, 737–746. doi: 10.18653/v1/P16-1070

Biber, D., (1993). Using register-diversified corpora for general language studies. Computational Linguistics, MIT Press, Cambridge, MA, USA, 19, 219–242.

Brooke, J., Hirst, G., (2012). Measuring interlanguage: Native language identification with L1-influence metrics., in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), 779–784.

Brunato, D., Cimino, A., Dell'Orletta, F., Montemagni, S., Venturi, G., (2020). Profiling-UD: a tool for linguistic profiling of texts, in Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), Marsseille, France, 7145–7151.

Brunato, D., De Mattei, L., Dell'Orletta, F., Iavarone, B., Venturi, G., (2018). Is this sentence difficult? Do you agree?, in: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). doi: 10.18653/v1/D18-1289

Chipere, N., Malvern, D., Richards, B., Duran, P., (2001). Using a corpus of school children's writing to investigate the development of vocabulary diversity, in: Technical Papers. Volume 13. Special Issue. Proceedings of the Corpus Linguistics 2001 Conference, Citeseer. pp. 126–133.

Colombo, A., (2010). A me mi. Dubbi, errori, correzioni nell'italiano scritto: Dubbi, errori, correzioni nell'italiano scritto. FrancoAngeli.

Covington, M.A., He, C., Brown, C., Naci, L., Brown, J., (2006). How complex is that sentence? a proposed revision of the Rosenberg and Abbeduto d-level scale. CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. Journal of Writing Research, 11(3), 415-443. doi: https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, A.S., Weston, J., McLain Sullivan, S., McNamara, D.S., (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. Written Communication, Written Communication, 28(3), 282-311. doi: https://doi.org/10.1177/0741088311410188

Daelemans, W., (2013). Explanation in computational stylometry. Computational Linguistics and Intelligent Text Processing 7817, 451-462. doi: 10.1007/978-3-642-37256-8_37

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills Journal of Writing Research, 2 (2), 151-177. doi: 10.17239/jowr-2010.02.02.4

Dahlmeier, D., Ng, H.T., Wu, S.M., (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, pp. 22–31.

De Mauro T. (1983). Per una nuova alfabetizzazione, in Gensini S. e Vedovelli M. (edited by), Teoria e pratica del glotto-kit. Una carta di identità per l'educazione linguistica, Milano, FrancoAngeli, 19-29.

Dell'Orletta, F., Venturi, G., Montemagni, S., (2011). Ulisse: an unsupervised algorithm for detecting reliable dependency parses, in Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, 115-124.

Durrant, P., Brenchley, M., & Clarkson, R. (2020). Syntactic development across genres in children's writing: The case of adverbial clauses. Journal of Writing Research, 12 (2), 419-452. doi: https://doi.org/10.17239/jowr-2020.12.02.04

Gildea, D. (2001). Corpus variation and parser performance, in Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. URL: https://www.aclweb.org/anthology/W01-0521.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. CALICO Journal, 20(3), 465-480. Retrieved February 9, 2021, from http://www.jstor.org/stable/24157525.

Karmiloff-Smith, A. (1986). Some fundamental aspects of language development after age 5. In P. Fletcher & M. Garman (Eds.), Language Acquisition: Studies in First Language Development (pp. 455-474). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620683.026

Kellogg, R.T. (2008). Training writing skills: A cognitive developmental perspective. Journal of writing research, 1(1), 1-26, doi:https://doi.org/10.17239/jowr-2008.01.01.1

Kerz, E., Qiao, Y., Wiechmann, D., Ströbel, M., (2020). Becoming linguistically mature: Modeling English and German children's writing development across school grades, in: Proceedings of the Fifteenth Work-shop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA Online, 65-74, doi:10.18653/v1/2020.bea-1.6.

Lei, L., Wen, J., (2020). Is dependency distance experiencing a process of min-imization? a diachronic study based on the state of the union addresses. Lingua 239, 102762, https://doi.org/10.1016/j.lingua.2019.102762.

Lu, X., (2009). Automatic measurement of syntactic complexity in child language acquisition. International Journal of Corpus Linguistics, 14(1):3-28. doi: https://doi.org/10.1075/ijcl.14.1.02lu

Lubetich, S., Sagae, K., (2014). Data-driven measurement of child language development with simple syntactic templates, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2151-2160.

Lucisano, P., (1984). L'indagine IEA sulla produzione scritta. Ricerca educativa 5, 41-61.

MacWhinney, B., (2000). The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associate. doi: 10.1177/026565909200800211

Malmasi, S., Keelan, E., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., Qian, Y., (2017). A report on the 2017 Native Language Identification shared task. Proceedings of the 12th Workshop on Building Educational Applications Using NLP, doi: 10.18653/v1/W17-5007.

Marconi, L., (1994). Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementari. Zanichelli.

McCutchen, D. (2011). From Novice to Expert: Implications of Language Skills and WritingRelevant Knowledge for Memory during the Development of Writing Skill. Journal of Writing Research, 3(1), 51-68. http://dx.doi.org/10.17239/jowr-2011.03.01.3

McNamara, D.S., Crossley, S.A., McCarthy, P.M. (2010). Linguistic features of writing quality. Written communication 27, 57-86. doi: https://doi.org/10.1177/0741088309351547

McNamara, D.S., Crossley, S.A., Roscoe, R.D., Allen, L.K., Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing 23, 35–59, doi: https://doi.org/10.1016/j.asw.2014.09.002

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

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al., (2016). Universal dependencies v1: A multilingual treebank collection, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1659–1666.

Richter, S., Cimino, A., Dell'Orletta, F., Venturi, G., (2015). Tracking the evolution of written language competence: an NLP–based approach, in: Proceedings of 2nd Italian Conference on Computational Linguistics (CLiC-it), 3-4 December, Trento, Italy, 236-240, doi: 10.4000/books.aaccademia.1277.

Sagae, K., Lavie, A., MacWhinney, B., (2005). Automatic measurement of syntactic development in child language, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 197-204. doi: https://doi.org/10.3115/1219840.1219865

Scarborough, H.S., (1990). Index of productive syntax. Applied psycholinguistics, Applied Psycholinguistics, Volume 11, Issue 1, March 1990, 1-22,
doi: https://doi.org/10.1017/S0142716400008262

Slater, S., Ocumpaugh, J., Baker, R., Almeda, M.V., Allen, L., Heffernan, N., (2017). Using natural language processing tools to develop complex models of student engagement, in: Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on, IEEE. pp. 542–547, doi: 10.1109/ACII.2017.8273652

Straka, M., Hajic, J., Strakova, J., (2016). UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC).

Visalberghi, A., Costa, M.C., (1995). Misurare e valutare le competenze linguistiche: guida scientifico-pratica per gli insegnanti. La nuova Italia.

Weiss, Z., Meurers, D., (2019). Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school, in: Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) at ACL. doi: 380–393.10.18653/v1/W19-4440

Wilcox, K.C., Yagelski, R. & Yu, F. (2014). The nature of error in adolescent student writing. Reading and Writing, 27(6), 1073-1094. doi: 10.1007/s11145-013-9492-x

Appendix

Table 15. Ranking of the first 80 features for the one-month temporal span. LinearSVR weights for each feature are also reported

Features	**Weights**	**Features**	**Weights**
dep_dist_punct	0.8586	verbs_num_pers_dist_Sing+1	0.1638
upos_dist_AUX	0.7886	upos_dist_PRON	0.1611
upos_dist_X	0.7414	wfc-S-lemma	0.1604
dep_dist_aux	0.7269	subj_post	0.1520
verbs_num_pers_dist_Sing+	0.5791	xpos_dist_VA	0.1495
upos_dist_PUNCT	0.5398	xpos_dist_PR	0.1433
verbs_form_dist_Part	0.5242	wfc-V-lemma	0.1429
dep_dist_cop	0.5238	verbs_num_pers_dist_Plur+1	0.1407
upos_dist_VERB	0.5094	upos_dist_NOUN	0.1363
xpos_dist_SW	0.4775	xpos_dist_S	0.1363
xpos_dist_FS	0.4615	dep_dist_xcomp	0.1322
verbs_form_dist_Fin	0.4285	xpos_dist_FB	0.1299
xpos_dist_AP	0.3745	dep_dist_flat:name	0.1225
dep_dist_det:poss	0.3360	ttr_form_chunks_200	0.1218
xpos_dist_SP	0.3145	subj_pre	0.1183
dep_dist_conj	0.2963	xpos_dist_A	0.1182
wfc-A-lemma	0.2739	verbal_head_per_sent	0.1179
wfc-S-word	0.2727	dep_dist_det	0.1143
verbs_num_pers_dist_Sing+3	0.2711	dep_dist_advmod	0.1121
wfc-A-word	0.2647	upos_dist_ADJ	0.1120
dep_dist_root	0.2581	ttr_lemma_chunks_100	0.1054
dep_dist_acl:relcl	0.2422	upos_dist_ADP	0.1038
dep_dist_nsubj	0.2391	xpos_dist_E	0.1038
dep_dist_advcl	0.2391	xpos_dist_PC	0.1013
wfc-TOT-lemma	0.2312	n_sentences	0.0997

dep_dist_mark	0.2239	n_tokens	0.0988
dep_dist_amod	0.2236	xpos_dist_V	0.0988
dep_dist_iobj	0.2123	n_ prepositional_chains	0.0903
verbs_num_pers_dist_Plur+	0.2100	dep_dist_acl	0.0865
dep_dist_nmod	0.2099	xpos_dist_FC	0.0864
dep_dist_flat:foreign	0.2074	dep_dist_ccomp	0.0843
xpos_dist_FF	0.1929	upos_dist_CCONJ	0.0840
dep_dist_expl	0.1907	xpos_dist_CC	0.0840
dep_dist_obl	0.1834	verbs_tense_dist_Past	0.0837
xpos_dist_X	0.1817	avg_token_per_clause	0.0822
dep_dist_aux:pass	0.1801	ttr_form_chunks	0.0773
dep_dist_cc	0.1756	xpos_dist_RD	0.0764
dep_dist_case	0.1753	xpos_dist_B	0.0761
dep_dist_obj	0.1743	ttr_lemma_chunks_200	0.0759
verbs_num_pers_dist_Plur+3	0.1641	ttr_lemma_chunks	0.0709

Table 16. Ranking of the first 80 features for the One year temporal span. LinearSVR weights for each feature are also reported

Features	**Weights**	**Features**	**Weights**
dep_dist_dislocated	0.1201	subj_pre	0.0541
xpos_dist_PP	0.1072	xpos_dist_FC	0.0533
xpos_dist_BN	0.1065	wfc-V-lemma	0.0517
xpos_dist_PD	0.0938	subordinate_pre	0.0516
xpos_dist_DD	0.0887	subordinate_post	0.0516
aux_form_dist_Fin	0.0872	xpos_dist_PE	0.0499
error_prep_omission	0.0814	error_capital_letter	0.0498
tot-error-lexicon	0.0813	xpos_dist_FF	0.0480
error_vocabulary	0.0813	wfc-S-word	0.0476
n_sentences	0.0802	dep_dist_fixed	0.0469
dep_dist_vocative	0.0800	verbs_form_dist_Fin	0.0464
xpos_dist_RI	0.0789	dep_dist_nsubj	0.0450
wfc-S-lemma	0.0781	aux_num_pers_dist_Plur+	0.0450
n_ prepositional_chains	0.0778	error_subj_verb_agreement	0.0446
n_tokens	0.0777	dep_dist_xcomp	0.0445
aux_tense_dist_Imp	0.0763	verbs_tense_dist_Fut	0.0441
verb_edges_dist_3	0.0760	dep_dist_conj	0.0441
error_conjunctions	0.0741	dep_dist_amod	0.0439
error_full_stop	0.0737	error_orthography_other	0.0438
tot-error-punctuation	0.0737	upos_dist_NOUN	0.0436
error_number_agreement	0.0731	xpos_dist_S	0.0436
error_prepositions	0.0724	tokens_per_sent	0.0424
char_per_tok	0.0710	verb_edges_dist_11	0.0422
verb_edges_dist_9	0.0707	aux_mood_dist_Sub	0.0419
verbs_num_pers_dist_Sing+	0.0673	dep_dist_aux	0.0417
aux_tense_dist_Fut	0.0672	aux_form_dist_Part	0.0414

verbs_num_pers_dist_Sing+3	0.0662	dep_dist_cop	0.0410
aux_num_pers_dist_Plur+2	0.0661	aux_tense_dist_Pres	0.0410
aux_num_pers_dist_Sing+	0.0657	obj_post	0.0405
aux_tense_dist_Past	0.0648	obj_pre	0.0405
avg_max_depth	0.0643	dep_dist_flat:name	0.0400
dep_dist_acl	0.0639	avg_prepositional_chain_len	0.0392
xpos_dist_DI	0.0624	verb_edges_dist_7	0.0386
verbs_form_dist_Part	0.0619	aux_mood_dist_Imp	0.0383
verb_edges_dist_4	0.0600	dep_dist_ccomp	0.0381
dep_dist_mark	0.0578	prep_dist_2	0.0379
subj_post	0.0569	xpos_dist_FS	0.0376
error_po’	0.0552	error_pronouns_omission	0.0366
prep_dist_3	0.0552	dep_dist_compound	0.0360
verbs_tense_dist_Imp	0.0548	tot-error-grammar	0.0357

Table 17. Ranking of the first 100 features for the Two years temporal span. LinearSVR weights for each feature are also reported

Features	**Weights**	**Features**	**Weights**
error_pronouns_omission	0.0464	upos_dist_AUX	0.0249
n_tokens	0.0441	dep_dist_mark	0.0248
wfc-V-lemma	0.0427	verbs_num_pers_dist_Plur+	0.0246
n_ prepositional_chains	0.0425	dep_dist_acl	0.0235
aux_tense_dist_Past	0.0414	verbs_num_pers_dist_Sing+3	0.0231
xpos_dist_RI	0.0405	upos_dist_NOUN	0.0230
obj_pre	0.0399	xpos_dist_S	0.0230
obj_post	0.0399	dep_dist_flat	0.0228
n_sentences	0.0370	verbs_form_dist_Part	0.0227
aux_num_pers_dist_Plur+1	0.0351	error_use_of_tense	0.0226
dep_dist_aux	0.0345	verbs_tense_dist_Imp	0.0226
error_pronouns_redundancy	0.0335	xpos_dist_DQ	0.0224
verbs_num_pers_dist_Sing+2	0.0333	aux_num_pers_dist_Sing+3	0.0223
aux_form_dist_Ger	0.0312	verbs_form_dist_Fin	0.0219
xpos_dist_FB	0.0311	wfc-A-lemma	0.0205
dep_dist_conj	0.0309	verb_edges_dist_0	0.0204
aux_tense_dist_Imp	0.0307	aux_num_pers_dist_Plur+	0.0201
wfc-A-word	0.0305	aux_num_pers_dist_+	0.0200
error_po'	0.0297	dep_dist_root	0.0198
wfc-S-word	0.0296	xpos_dist_FS	0.0193
verbs_num_pers_dist_Plur+1	0.0295	upos_dist_ADP	0.0190
dep_dist_aux:pass	0.0292	xpos_dist_E	0.0190
verbs_gender_dist_Masc	0.0291	xpos_dist_FF	0.0189
wfc-V-word	0.0289	error_use_of_h_redundancy	0.0182
verb_edges_dist_8	0.0288	subj_post	0.0182
upos_dist_NUM	0.0287	subj_pre	0.0182

xpos_dist_N	0.0287	prep_dist_1	0.0182
wfc-S-lemma	0.0286	verbs_tense_dist_Pres	0.0182
xpos_dist_PQ	0.0283	verb_edges_dist_2	0.0180
aux_form_dist_Inf	0.0275	verbs_num_pers_dist_Plur+3	0.0177
aux_num_pers_dist_Sing+1	0.0274	xpos_dist_DD	0.0177
max_links_len	0.0272	tot-error-lexicon	0.0176
verbs_num_pers_dist_Plur+2	0.0272	error_vocabulary	0.0176
xpos_dist_PD	0.0265	xpos_dist_X	0.0175
error_gender_agreement	0.0265	aux_form_dist_Fin	0.0172
xpos_dist_BN	0.0265	verbs_num_pers_dist_Sing+	0.0171
error_relative_pronouns	0.0265	prep_dist_3	0.0170
verbs_gender_dist_Fem	0.0261	xpos_dist_PI	0.0167
xpos_dist_VA	0.0261	dep_dist_obj	0.0167
dep_dist_nsubj:pass	0.0252	subordinate_dist_7	0.0163