Analysing the Evolution of Students' Writing Skills and the Impact of Neo-standard Italian with the help of Computational Linguistics

Rachele Sprugnoli
FBK, Trento
sprugnoli@fbk.eu

Sara Tonelli
FBK, Trento
satonelli@fbk.eu

Alessio Palmero Aprosio
FBK, Trento
aprosio@fbk.eu

Giovanni Moretti
FBK, Trento
moretti@fbk.eu

Abstract

English. We present a project aimed at studying the evolution of students' writing skills in a temporal span of 15 years (from 2001 to 2016), analysing in particular the impact of neo-standard Italian. More than 2,500 essays have been transcribed and annotated by teachers according to 28 different linguistic traits. We present here the annotation process together with the first data analysis supported by NLP tools.

Italiano. *In questo contributo presentiamo un progetto finalizzato allo studio dell'evoluzione delle abilità di scrittura negli studenti in un arco temporale di 15 anni (dal 2001 al 2016), e in particolare all'analisi dell'impatto dell'italiano neo-standard. In questo contesto, più di 2.500 temi sono stati trascritti e annotati da insegnanti, registrando la presenza di 28 diversi tratti linguistici. Il presente studio illustra il processo di annotazione e le prime analisi dei dati con il supporto di strumenti TAL.*

1 Introduction

In this work, we present an extensive study on the evolution of high-school students' writing skills, taking into account essays spanning 15 years (from 2001 to 2016). In particular, we are interested in tracking the presence of expressions and constructions typical of neo-standard Italian (Berruto, 2012), in the light of the recent public discussion on the 'decline of Italian in schools' [1]. The Italian neo-standard is the current linguistic register in Italy, in which forms previously considered colloquial have become widely accepted in the national language.

We analyse more than 2,500 essays written by students from different high-schools in the Autonomous Province of Trento during the exit exam (the so-called *Maturità*). The study is the outcome of a project comprising different steps: *i)* digital acquisition and transcription of thousands of essays balancing their distribution across school years and school types; *ii)* computer-assisted annotation of some linguistic traits of interest; *iii)* diachronic analysis of the traits. While the first step has been carried out by the Istituto provinciale per la Ricerca e la Sperimentazione educativa (IPRASE), we led steps *ii)* and *iii)*, which are discussed in the next sections. Beside an in-depth and diachronic study of the evolution of students' writing skills, a major contribution of this paper is also the release of the corpus in the form of embeddings and n-grams.

2 Corpus Collection

The staff of IPRASE have digitized and transcribed essays stored in the archives of 21 secondary schools located in different areas of Trentino Province. These areas include both the two major cities, Trento and Rovereto, but also other communities in the valleys (Val di Fiemme, Val di Non, Valsugana) and Riva del Garda. Nine different types of schools were involved: liceo classico, liceo scientifico, liceo artistico, liceo linguistico, liceo musicale e coreutico, liceo delle scienze umane, istituto tecnico tecnologico, istituto tecnico economico and istituto professionale. Six school years were chosen between 2000-2001 and 2015-2016, thus having a temporal span of 15 years for a total of 2,544 essays and almost 1.5 million words. Table 2 shows the distribution of essays per year with the corresponding number of

[1] See the open letter signed by around 600 University professors at http://gruppodifirenze.blogspot.it/2017/02/contro-il-declino-dellitaliano-scuola.html.

words. These essays are of the so-called type B, that requires students to write a short essay or a newspaper article. Students can choose between 4 areas: artistic-literary, socio-economic, technical-scientific, historical-political. For each area, a title is given together with a set of reference materials. For example, students writing an essay of type B with historical-political content in 2014 were asked to comment some excerpts from Hannah Arendt, Ghandi and Martin Luther King about violence and non-violence in the XX Century.

SCHOOL YEAR	**#ESSAYS**	**#WORDS**
2000-2001	417	244,312
2003-2004	439	270,388
2006-2007	430	258,188
2009-2010	429	245,821
2012-2013	421	234,329
2015-2016	408	224,776
TOTAL	2,544	1,477,814

Table 1: Number of essays and words per school year in our corpus.

Due to privacy reasons, we are not allowed to distribute the full texts of the corpus. However, we release both word vectors and n-grams of the essays. We build three types of embeddings with 300 dimensions: the GloVe embeddings based on linear bag-of-words contexts (Pennington et al., 2014), Levy and Goldberg's ones using dependency parse-trees (Levy and Goldberg, 2014), and fastText embeddings with bag of character n-grams (Bojanowski et al., 2017). As for the n-grams, we generated both case-sensitive and case-insensitive sequences per school year, considering the range [1,5]. N-grams and pre-trained word embeddings in text format are available for download on our website[2]. In addition, word vectors are visualized through a dedicated stand-alone version of the TensorFlow embedding projector (Smilkov et al., 2016)[3].

3 Description of Linguistic Traits

Around 20 teachers have been involved in the annotation of essays using the CAT platform (Bartalesi Lenzi et al., 2012), through which they had to annotate between 100 and 150 essays each. We also organised 2 preliminary training sessions with the teachers to show the tool functionalties, explain the annotation process and make sure that everyone followed the guidelines[4]. Note that the teachers knew neither the name of the student writing the essay nor his/her school. Moreover, for all of them, it was the first time using an electronic platform for text annotation.

We briefly present in Table 2 the traits that the teachers had to mark on each essay. The goal of the annotation is to detect the presence of linguistic traits that were deemed relevant to diachronically study style and complexity evolution by IPRASE experts and teachers. This approach is therefore rather different from the standard essay correction that is usually performed by teachers, and for this reason the training phase was particularly relevant.

The list of traits to include in the project was mainly inspired by the work of (D'Achille, 2003) and (Boscolo and Zuin, 2015). The goal of this annotation was to cover all levels of linguistic analysis, including lexical choices (e.g. trait 8 and 20), grammar (e.g. trait 1 and 2), semantics (e.g. trait 15) and discourse structure (e.g. trait 24 and 25).

In the first Table column, we mark traits that were identified in a fully automatic way (A), those that were annotated semi-automatically (S), and the manual ones (M). For those marked with S, we pre-processed the essays using the Tint NLP tool (Aprosio and Moretti, 2018) enriched with a set of new modules developed to add all information needed to speed up annotation. For example, for traits 21 and 23 we matched the essay n-grams with pre-defined lists of politically correct expressions and cliché expressions provided by IPRASE, so that teachers could see in the CAT interface the corresponding markables already highlighted, and they just had to validate them. For other traits, for example 10 and 11, they had to add attributes to the markables. For some traits, we performed pre-annotation using available external resources, for example the list of affixes included in the *derIvaTario*[5] for trait 13 (Talamo et al., 2016).

After the initial training phase, the average annotation time for each essay through the web interface was 30 minutes. We roughly estimate that the same task would take at least one hour on a standard Word document. Another advantage of using

[2] `https://dh.fbk.eu/technologies/students-essays`

[3] `http://dhlab.fbk.eu/TemiVectors/`

[4] A complete version of the annotation guidelines (in Italian) is available at this link: `http://bit.do/erd9P`

[5] `http://derivatario.sns.it/`

Type	ID	Trait	Description
S	1	Monosyllables	Annotate monosyllabic terms with a wrong accent
A	2	Apostrohpes	Annotate the wrong use of apostrophes for the article ‘un’
S	3	Capitalized words	Annotate wrong capitalisations inside a sentence
A	4	“il”	Annotate the wrong use of “il”
S	5	Personal pronouns	Annotate personal pronouns and mark when ‘loro’ is used to mean ‘a loro’
S	6	“Gli”	Annotate different uses of ‘gli’ including mistakes
S	7	“Questo”	Annotate when ‘quest*’ is used to refer generically to the discourse context
A	8	Generic words	Annotate generic words such as ‘bello’, ‘brutto’, ‘fare’, ‘dire’, ‘cosa’
S	9	Indicativo imperfetto	Annotate different types of imperfetto (e.g. in place of conjunctive, in hypothetical clauses)
S	10	Gerund	Annotate different types of gerundio
S	11	Indicativo presente	Annotate different types of indicativo presente
A	12	‘stare / andare’	Annotate when ‘stare’ / ‘andare’ are used properly or in phrasal constructions
S	13	Affixes	Annotate words created using specific affixes such as -anti, ‘-dopo’, ‘-trans’, ‘-ismo’, ‘-izzare’, ...
S	14	Number of words, clauses, sentences	Count the number of words, clauses and sentences. Annotate verbless clauses when not in the title
S	15	Connectives 1	Annotate the use of very generic connectives (‘che / dove / allora’) and their correct or improper use
S	16	Connectives 2	Count complex connectives such as ‘nondimeno’, ‘sebbene’, ‘qualora’ and annotate their use
S	17	Punctuation	Count punctuation marks: [; : ! ” ... , .] and annotate their correct or improper use
S	18	Connectives beginning a sentence	Identify connectives such as ‘perché’ and ‘quando’ at the beginning of a sentence and annotate their use
S	19	Informal register	Annotate a set of expressions belonging to an informal register (‘della serie’, ‘tipo’, ‘troppo forte’, etc.)
S	20	Anglicisms	Annotate adapted and not adapted anglicisms
S	21	Politically correct terms	Annotate politically correct terms such as ‘ministra’, ‘sindaca’, ‘non vedente’, etc.
S	22	Multiwords	Annotate multiword expressions (*polirematiche*)
S	23	Cliché expressions	Annotate cliché expressions from a predefined list
M	24	Dislocated clauses	Annotate left or right dislocated sentences
S	25	Cleft sentences	Annotate cleft sentences
S	26	‘li’	Annotate ‘li’ when it is mistakenly used instead of ‘gli’
A	27	Euphonic ‘d’	Annotate when ‘d’ is added before a word starting with a vowel
M	28	Other traits	Add other relevant linguistic phenomena that are not captured by previous traits

Table 2: List of annotated traits with a label for Automatic (A), Semi-automatic (S) or Manual (M)

the CAT interface was the possibility to have all annotations in a consistent format, easily export them to compute statistics and make comparisons.

4 Linguistic Analysis

We present here an analysis of some traits of interest. We focus in particular on traits that are, at least in part, automatically annotated and counted (marked with A or S in Table 2), because the work of those requiring a manual annotation is still in progress. For each trait we compute the observed relative frequency per 10,000 words. This normalization has allowed us to have more easily comparable and legible numbers. Furthermore, we calculate the Gulpease index to monitor writing complexity (Lucisano and Piemontese, 1988). This score has been specifically defined for measuring the readability of Italian texts based on proficiency level and it combines two linguistic variables: the average length of the words and of the sentences in a document. Its value determines the level of readability of a text: the higher the score, the easier the text is to understand.

To extract reliable measures of students' language use, we removed from the texts the quotations present in the essays citing the reference material provided together with the topic. This preprocessing step was performed by adopting the FuzzyWuzzy package[6], a Java fuzzy string matching implementation, and the Stanford CoreNLP quote annotator[7]. These tools allow us to recognize text reuse both when it is explicitly signaled by quotes and when there is no overt signal. The average percentage of quotations within the corpus is 1.9% but it varies a lot among the essays, reaching up to 46% of the content in some cases. The following is an example taken from an essay about the pursuit of happiness in 2010 for the socio-economic area. The snippet in bold, containing one of the complex connectives of trait 16, was automatically removed: *La riflessione di Zygmunt Bauman sembra essere una risposta:* ***“L'incertezza è l'habitat naturale della vita umana, sebbene la speranza di sfuggire ad essa sia il motore delle attività umane.”***

After removing quotations, we obtain the following results for the automatically annotated traits:

Trait 8 - Generic Words. We trace the presence of semantically generic and polysemic words, which are frequently used in neo-standard Italian (Fig. 1). In particular, lemmas ‘fare’, ‘dire’, and ‘cosa’ (*to make*, *to say*, *thing*) show a decrease in occurrence in the last two school

[6] https://github.com/xdrop/fuzzywuzzy

[7] https://stanfordnlp.github.io/CoreNLP/quote.html

years considered (2012-2013 and 2015-2016). For example, the relative incidence of 'fare' every 10,000 tokens goes from 42.013 in 2000-2001 to 26.857 in 2015-2016 indicating an effort to use more specific and differentiated expressions. Liceo classico has the lowest ratio for 'fare' and 'dire', whereas istituto professionale has an occurrence above the average for 'fare' and 'cosa'.

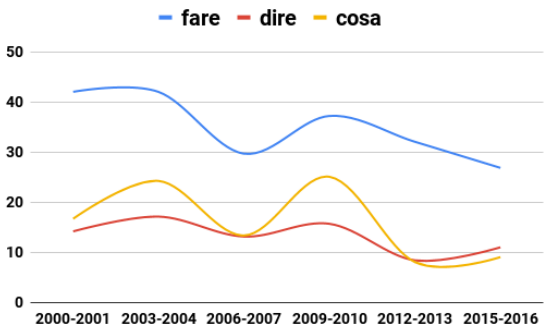

Figure 1: Observed relative frequency of three generic words per 10,000 tokens.

Trait 14 - Nominal Sentences. Sentences without a verbal predicate are a typical feature of news style and juvenile writing, to make the text dramatic and concise (Dardano, 1986; Ardrizzo and Gambarara, 2003). This tendency is present also in our corpus with an impact of 6.1% over the total amount of sentences, after removing the title of the essays. The trait is particularly relevant in liceo classico with an above-average percentage of 7.7%.

Trait 16 - Complex Connectives. The lack of complex connectives is another indicator of neo-standard Italian. As shown in Figure 2, 'nondimeno' is never used by students and also 'qualora' and 'giacché', used mostly in liceo classico, disappear in the last two school years from all the essays. 'Affinché' is adopted in all school types with the only exception of liceo artistico, in which complex connectives are barely used.

Trait 17 - Punctuation. Over the last two school years considered in our analysis, there has been an overall decline in the use of punctuation with the exception of question marks (see Figure 3). The frequent use of question marks is inherited from the style of news (Buroni, 2009); however, the peak in 2009-2010 is also due to the presence of a question in the title of an essay (*Siamo soli?*), which led students use the same rhetorical device in their texts. The presence of punctuation not suitable for medium-high style such as multiple exclamation marks and suspension points is also decreasing.

Trait 27 - Euphonic 'd'. Following a recent grammatical rule[8], the euphonic 'd' should be introduced only when the conjunction 'e' or the preposition 'a' are followed by a word starting with the same vowel: e.g., *ed ecco*, *ad andare*. However, this rule is not followed in the essays and the presence of 'd' between two different vowels is higher than the one between the same vowels (33.8 versus 17.6 of relative frequency). Besides, while the disappearance of this trait is considered a characteristic of neo-standard Italian (D'Achille, 2003), this trend is not found in our corpus, where the relative frequency of euphonic 'd' is only 6 points lower than the same conjunction without 'd' preceding a vowel.

Gulpease. We computed the Gulpease index to see whether there has been a decrease of complexity, i.e. an increase in readability, over time. Contrary to our expectations, the average readability of essays has slightly decreased in the last two years considered, with a drop of 1.8 points, bringing it below 50. This corresponds to texts that are quite difficult to read for a person with a medium school degree (*diploma di scuola media* in the Italian school system), but not too challenging for a person with a high school degree. Moreover, values do not change much across different school types.

These preliminary analyses show that the impact of neo-standard Italian is multi-faceted and, while some traits confirm that students' language is getting simpler and less formal (e.g. overall decline of punctuation), some others seem to contradict this finding (e.g. decline in the use of 'fare', 'dire', 'cosa'). Also the differences across school types are not clear-cut and consistent.

5 Related Work

While several works in the past have focused on the creation and analysis of corpora to study students' mistakes, their writing quality and their rate of progress over the year (Parr, 2010; McNamara et al., 2010), they have mainly dealt with English essays. A notable exception are two corpora in

[8] http://www.accademiadellacrusca.it/it/lingua-italiana/consulenza-linguistica/domande-risposte/d-eufonica

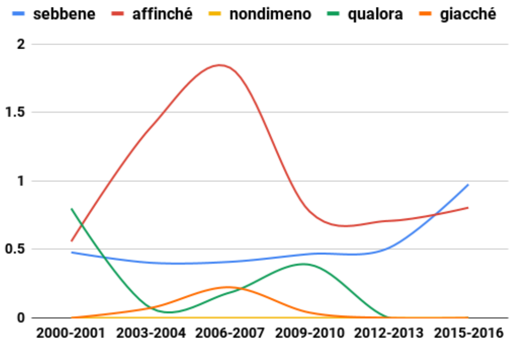

Figure 2: Observed relative frequency of complex connectives per 10,000 words.

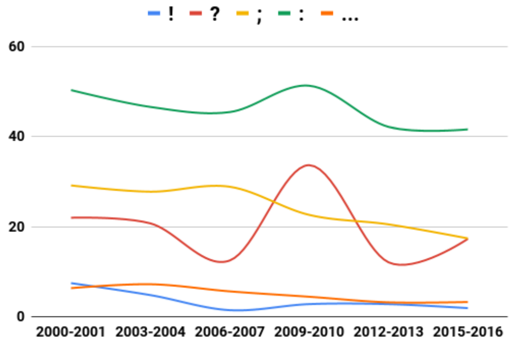

Figure 3: Observed relative frequency of punctuation per 10,000 words.

German, the KoKo corpus of argumentative essays to study pupils' writing competences (Abel et al., 2016) and the corpus collected by Berkling et al. (2014) to study different error categories.

As for Italian, a relatively small number of studies has been carried out with various goals. The projects TIscrivo (2011-2014) and TIscrivo 2.0 (2014-2017)[9] have been launched to investigate the writing skills of primary schools and lower secondary schools in Southern Switzerland (Cignetti et al., 2016), and have led to the creation of a corpus of 1,735 essays. Another research deals with the analysis of oral and written productions of Italian children in primary schools, and 200 texts have been collected in the ISACCO corpus (Brunato and dell'Orletta, 2015). Another corpus, called CItA (Barbagli et al., 2016), includes texts written in the first and second year of lower secondary school, tracking L1 writing competence of the same group of students over two school years.

Compared to previous works, our analysis is different in several ways. First, none of the previous studies considers a text span of 15 years. Then, the traits to be annotated are different: we do not focus on mistakes, but on indicators of neo-standard Italian. Finally, our interest lies also in the annotation workflow, studying how NLP can support the identification of such traits and implementing the necessary processing modules to speed up annotation.

6 Conclusions

In this work, we have presented a project aimed at tracking the evolution of students' writing skills over time. The goal of this work was not only to introduce the corpus collection and annotation activities, but also to show how this kind of projects can benefit from NLP by speeding up annotation and increasing data consistency. In the future we will complete the analysis of all the traits for a more comprehensive view of the role of neo-standard Italian in students' essays. We will also use some of the manual annotations to train new NLP modules performing the same task automatically.

Acknowledgments

We would like to thank Chiara Motter from IPRASE for coordinating the corpus transcription, and the high-school teachers for annotating the essays.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2016. An extended version of the koko german L1 learner corpus. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: An all-inclusive Suite for NLP in Italian. In *Proceedings of CLIC-it.*

Giuseppe Ardrizzo and Daniele Gambarara. 2003. *La comunicazione giovane.* Rubbettino Editore.

[9] `http://dfa-blog.supsi.ch/tiscrivo/`

Alessia Barbagli, Pietro Lucisano, 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 2016, Portorož, Slovenia, May 23-28, 2016.*

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *In Proceedings of LREC 2012*, pages 333–338.

Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Hein, Rémi Lavalley, Ludwig Linhuber, and Sebastian Stüker. 2014. A database of freely written texts of german school students for the purpose of automatic spelling error classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Gateano Berruto. 2012. *Sociolinguistica dell'italiano contemporaneo*. Carocci.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Pietro Boscolo and Elvira Zuin, editors. 2015. *Come scrivono gli adolescenti. Un'indagine sulla scrittura scolastica e sulla didattica della scrittura*. Il Mulino.

Dominique Brunato and Felice dell'Orletta. 2015. ISACCO: a corpus for investigating spoken and written language development in Italian school–age children. In *Proceedings of CLIC-it*.

Edoardo Buroni. 2009. Politicamente corretto? Aspetti grammaticali nei quotidiani politici della "Seconda Repubblica" tra norma, uso medio e finalità pragmatiche. *Studi di Grammatica Italiana*, 2007:107–163.

Luca Cignetti, Silvia Demartini, and Simone Fornara. 2016. *Come TIscrivo? La scrittura a scuola tra teoria e didattica*. Aracne.

Paolo D'Achille. 2003. *L'italiano contemporaneo*. Il mulino Bologna.

Maurizio Dardano. 1986. *Il linguaggio dei giornali italiani*, volume 18. Laterza.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*, pages 302–308.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Judy M. Parr. 2010. A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research*, vol. 2(issue 2):129–150. Query date: 2018-06-25.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding Projector: Interactive visualization and interpretation of embeddings. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102.