Form@re
Open Journal per la formazione in rete

articoli

Correlations among natural language processing indicators and critical thinking sub-dimensions in HiEd students

Correlazioni tra indicatori del linguaggio naturale e sotto-dimensioni del pensiero critico in studenti universitari

Antonella Poce[a], Francesca Amenduni[b], Maria Rosaria Re[c], Carlo De Medio[d], Alessandra Norgini[e,1]

[a] *University of Roma Tre, antonella.poce@uniroma3.it*
[b] *University of Roma Tre, francesca.amenduni@uniroma3.it*
[c] *University of Roma Tre, mariarosaria.re@uniroma3.it*
[d] *University of Roma Tre, carlo.demedio@uniroma3.it*
[e] *University of Roma Tre, alessandra.norgini@uniroma3.it*

Abstract

Most of the attempts to develop and validate tools for the automatic assessment of Critical Thinking (CT) related-skills applied Natural Language Processing techniques (NLP) to English written texts, with a few applications in other languages. Therefore, this research was aimed at understanding which NLP features correlates with six CT sub-dimensions in essays written in Italian language. 206 Master Degree students' pre-post essays were assessed both by human evaluators and by an algorithm which automatically calculates different kinds of NLP features. We found a positive internal reliability and a medium to high inter-coder agreement of the human evaluators. Three NLP indicators significantly correlate with CT total score: *Corpus Length*, *Syntax Complexity*, and an adapted measure of *Term Frequency-Inverse Document Frequency*.

Keywords: critical thinking; assessment; natural language processing; higher education.

Sintesi

La maggior parte dei tentativi di sviluppare e validare strumenti per la valutazione automatica delle competenze di Pensiero Critico (CT) ha applicato tecniche di elaborazione del linguaggio naturale (NLP) a testi scritti in inglese, con poche applicazioni in altre lingue. Pertanto, questa ricerca mira a comprendere quali indicatori NLP, estratti da saggi scritti in italiano, correlino con sei sotto-dimensioni del CT. 206 saggi pre-post di studenti di laurea magistrale sono stati valutati sia da esperti umani che da un algoritmo che calcola automaticamente alcuni indicatori NLP. È stata riscontrata una buona attendibilità interna e un accordo inter-giudice medio-alto. Tre indicatori NLP correlano in modo significativo al punteggio di CT totale: lunghezza del testo, complessità della sintassi e una versione adattata del *Term Frequency-Inverse Document Frequency*.

Parole chiave: pensiero critico; valutazione; elaborazione del linguaggio naturale; istruzione universitaria.

[1] A. Poce coordinated the research presented in this paper. Research group is composed by the authors of the contribution that was edited in the following order: A. Poce par. 1, 1.1., 1.2, 2.1, 4. F. Amenduni par. 3, 3.1, 3.2. M. R. Re, 2.2, 2.3, 2.4. C. De Medio 3.3., A. Norgini 2.5.

1. Introduction

Nowadays, a debate regarding the role that higher education is supposed to cover in the broader society is present at an international level. The debate refers to a dialectical conflict between two different stances: should university prepare students to fulfil the job market needs? Or is the university supposed to transmit the knowledge without considering the economic pressure and professional skill training? To which extent is it possible to reconcile these contrasting perspectives? An education system that focuses on developing higher-order skills, especially Critical Thinking (CT), could be a way to overcome this conflict. Enhancing students' CT is not the only necessary skill to enter and fulfil the job market needs (OECD, 2012; Wagenaar, 2018). Also, it provides students with tools to be autonomous thinkers and active citizens (Davies & Barnett, 2015). Having said that, CT operationalization and definition still represents an open challenge, and therefore, there are many different perspectives regarding the best way to assess and capture it.

Assessment tests for CT could be classified in different ways. Hyytinen, Nissinen, Ursin, Toom, & Lindblom-Ylänne (2015) differentiated self-report from performance-based measurements. Moreover, the performance-based measurements can be classified into Multiple-Choice (MC) tests / questionnaires and Constructed Response Tasks (CRT). Another way to classify CT assessment is to distinguish between assessment tools focused on CT as a process or as an outcome (Garrison, Anderson, & Archer, 2001).

Some authors point out that the MC measures use cannot be proper for the higher-order skills assessment, such as CT; according to some authors, MC items can be answered without reading the respective text passage. MC tests may be answered merely by low-level processing, such as factual recognition and selection (Nicol, 2007). A further concern regarding MC items is that they make test-takers select between pre-determined answers rather than allowing individualised responses as in CRT (Rauch & Hartig, 2010). Another weakness concerns students:

A student may be able to recognise the correct answer that they would have never been able to generate on their own. In that sense, MC items can present an exaggerated picture of a students' understanding or competence, which might lead teachers to invalid inferences. (Popham, 2003).

Moreover, MC tests can never assess students' skills to synthesise or generate their answers (Popham, 2003). Lastly, all the tests based on MC are chargeable, which limits their accessibility and their use in educational contexts. To address the limitations of MC tests, researchers have developed alternative assessment methods, which involve the adoption of open-ended tasks.

1.1. Open-ended measures in Critical Thinking assessment

Open-ended measures are characterised by the requirements given to the examinees to create their answers to questions. In these measures, students usually need to analyse, evaluate and synthesise complex information and provide a reasoned explanation. It is possible to create more authentic contexts and assess students' ability to generate rather than select responses by using open-ended measures. Research has long established that the ability to recognise is different from the ability to generate (Shepard, 2000). These tasks are sometimes referred to as *authentic assessment* because they elicit the same

thinking processes that individuals use when they solve complex problems in their everyday lives (Andrews & Wulfeck, 2014). Indeed, in real-life situations where CT skills need to be exercised, no choices are provided. Instead, people are expected to come up with their own choices and determine which one is preferable based on the question at hand. Thus, according to some authors, open-ended measures could provide a better proxy of real-world scenarios than MC items.

Ennis (1993) was one of the first authors who highlighted the need to adopt open-ended measures for CT assessment. According to Ennis (1993), open-ended measures are necessary because MC tests are not comprehensive and miss much important CT elements: "The MC tests can, to varying degrees, be used for […] diagnosis, feedback, motivation, impact of teaching, and research. But discriminating judgment is necessary. For example, if a test is to be used for diagnostic purposes, it can legitimately only reveal strengths and weaknesses in aspects for which it tests. The less comprehensive the test, the less comprehensive the diagnosis. For a comprehensive assessment, unless appropriate multiple-choice tests are developed, open-ended assessment techniques are probably needed. Until the published repertoire of open-ended critical thinking tests increases considerably, and unless one uses the published essay test, […] it is necessary to make your own." (p. 184).

Although it has been almost 30 years since Ennis quote, the limitations of MC tests for CT assessment have not been completely overcome yet. Contrasting results have been found regarding the comparability of MC and CRT measures for CT assessment. In a report of 2009, Klein et al. reported high correlation levels between different MC tests and CRT tests for CT (which varies from 0.79 to 0.93). Hyytinen et al. (2015), who found opposite results, compared the two measures used in the OECD's AHELO (Assessment of Learning Outcomes in Higher Education) project for assessing CT skills: the CLA (Collegiate Learning Assessment) and an MC questionnaire from the Australian Council for Educational Research (ACER). The results showed that the correspondence between the CLA and the MC questionnaire was fully comparable in 45.5% of the students' test performance. Ten percent of the students had opposite test results. These students were further divided into two dissonant groups: (i) students with high MC scores but low CLA scores, and (ii) students with low MC scores but high CLA scores. By analysing the CLA responses qualitatively, the authors found out that students' responses in the first group were comprised of isolated and reproduced facts. In contrast, in the second group, the students' written responses indicated the in-depth material understanding. Based on these features they labelled these groups as (i) Superficial Processing (e.g. students reproduced or slightly modified portions of text sources, without explaining the content of the materials in their own words), and (ii) Thorough Processing (e.g. students evaluated the quality of the information and considered its premises, as well as the implications of different conclusions). The authors found out that the reason why the *Thorough Processing* group obtain a low score in the MC questionnaire was not due to the wrong answers, but due to the high number of unanswered questions. The authors concluded that MC tests do not measure students' skills to produce arguments and to give reasoned explanations, which are the essential elements of CT. Although the scoring of the CRT might be challenging, the students' written answers reveal the level of processing and understanding. Figure 1 presents the most known standardised tests to assess CT, which employ open-ended measures. As shown in Figure 1, CLA and HCTA presented MC items too; thus, they can be considered the multi-response format assessments. According to different authors (Ku, 2009; Liu, Frankel & Roohr, 2014), a measurement that elicits both open-ended and MC response formats should be pursued in CT assessment.

Test	Format of the Open-Ended Measures	Developers	Competences assessed
Ennis Weir Critical Thinking Essay Test (EWCTET)	Given an argumentative passage, the examinees have to evaluate the logic of the passage and defend their own argument.	Ennis & Weir, 1985	Recognising formal and informal fallacies; individuating alternative solutions; assessing quality of the arguments and producing own arguments
International Critical Thinking Essay Test (ICTET)	Given a literary text (e.g. the Art of Loving, Erich Fromm) examinees are required to (1) paraphrase; (2) explicate; (3) analyse; (4) evaluate; (5) role-playing the author.	Paul & Elder, 2006	Reflecting, self-monitoring, summarising, exemplifying, synthetizing, connecting with daily life experiences, explicating the thesis, analysing the logic, applying standards in writings
Collegiate Learning Assessment (CLA)	Given realistic problems, which include more or less relevant reading materials (e.g. letters, summaries of research reports, articles, graphs), examinees are asked to organise, analyse, synthesise and evaluate these multiple sources of information to arrive at a solution or explanation of a problem.	Council for Aid to Education, 2000	Analysis and problem-solving; writing effectiveness; writing mechanics
Halpern Critical Thinking Assessment Using Everyday situations (HCTAES)	Given 20 everyday scenarios, respondents are first asked an opened-ended question (e.g. "Based on this information, would you support this idea? Explain why") which is followed by a forced choice question.	Halpern, 2016	Verbal reasoning skills; argument and analysis skills; skills in thinking and hypothesis testing; using likelihood and uncertainty; decision-making and problem-solving

Figure 1. Validated Tests to Assess CT General Skills and Dispositions Based on Open-Ended Measures.

By looking at general features of these tests, it is possible to retrieve some common characteristics of the open-ended item format:

- they use ill-structured problems. Moss and Koziol (1991) explain that test questions should require students to go beyond the available information in the task to draw inferences or make evaluations. Besides, problems should have more than one plausible or defensible solution, and sufficient information and evidence should be present within the task to enable students to support multiple views;
- they provide contradictory materials or sources and focus on controversial topics. Fischer, Spiker, and Riedel (2009) argue that CT is a stimulus-bound phenomenon meaning that certain external task features may impact whether CT is elicited in a given assessment context. They demonstrated that certain tasks types are more likely to elicit CT than others. The level of consistency, or lack of contradictions, within stimulus materials did have the primary effect; it is more likely to prompt CT while using inconsistent or contradictory materials than consistent and coherent stimulus materials.

All the standardised assessment method presented in Figure 1 assess CT as an outcome. Among the first authors who emphasised the importance of studying CT processes rather than outcomes were Garrison et al. (2001). They stated: "Critical thinking is both a process and an outcome. As an outcome, it is best understood from an individual perspective—that is, the acquisition of deep and meaningful understanding as well as content-specific critical inquiry abilities, skills, and dispositions.[…] The difficulty of assessing critical thinking as a product is that it is a complex and (only indirectly) accessible cognitive process. However, and most relevant here, from a process perspective, it is assumed that acquiring critical thinking skills would be greatly assisted by an understanding of the process. Moreover, it is assumed that facilitating the process of higher-order learning could be assisted through the use of a tool to assess critical discourse and reflection" (Garrison et al., 2001, p. 8).

Chou, Wu, and Tsai (2019) found that the Garrison and colleagues' model was the most adopted qualitative method for studying CT in e-learning settings, between January 2006 and November 2017, followed Newman, Johnson, Webb, and Cochrane (1997) and Newman, Webb, and Cochrane (1995) coding framework. Both Garrison and colleagues' and Newman and colleagues' models adopted qualitative content analysis for retrieving manifestations of CT in students' written texts.

1.2. Road to Critical Thinking automatic assessment

Although the acknowledged importance of using open-ended measures in CT assessment, they are less widespread than closed measures because they present different disadvantages. The most important is the difficulty of scoring (Attali, 2014). The open-answer assessment is characterised as subjective and open to scoring bias because examinees' responses are traditionally scored by using human evaluation. The CRT scoring is also considered time consuming and expensive; a large amount of time and effort is needed to train scorers and to score the responses. According to Liu, Frankel, and Roohr (2014) automatic assessment of open-ended answers could be a viable solution to these concerns. Automatic assessment of learning outcomes is a *hot topic* in educational research for at least two reasons: firstly, the availability of learning data is growing exponentially due to the spreading of online education. Secondly, researchers in the field of *Big Data*, *Machine Learning*, and *Artificial Intelligence* can provide educators with sophisticated tools for processing an immense amount of linguistic and behavioural data. One of the first attempt to automatically score open-ended answers for CT assessment was reported by the CLA developers. Two tasks named the *break-an-argument* and *make-an-argument* are scored automatically through Natural Language Processing (NLP) programs. NLP is an analysis of a human language by using computers aimed at automated discourse analysis. The term *natural* was coined to refer to human language in contrast to computer languages. NLP techniques can provide information about multiple levels of text: from the simplest level constituted by the analysis of single words used in a discourse, to the more complex levels which are the semantics as well as the discourse structure (McNamara, Allen, Crossley, Dascalu, & Perret, 2017). Klein (2007) reported study results in which the NLP reliability of the CLA was tested. Students' answers to one of the analytical writing tasks were assessed both by human expert assessors and by two NLP algorithms developed by Educational Testing Service (ETS): e-rater (https://www.ets.org/erater/about) and c-rater (https://www.ets.org/accelerate/ai-portfolio/c-rater). In Klein's study, human assessors were provided with an assessment guide, which contained 40 separate items (graded 0 or 1) and a 5-point overall communication score. For the latter score, the assessors were asked to consider whether

the answer was well organised; whether it communicated clearly; whether arguments and conclusions were supported with a specific reference to the documents provided; and whether the answer used appropriate vocabulary, language, and sentence structure. Readers were instructed to ignore spelling mistakes. ETS built the e-rater algorithm for the communication score based on grades assigned by human evaluators; it contains modules for identifying the following features relevant to the scoring guide criteria: syntax, discourse, topical content, and lexical complexity. ETS's c-rater, designed for the short-answers assessment, was used to create scores for items 1 through 40. The correlation between hand and machine assigned mean scores, on the make-an-argument and break-an-argument tasks, was 0.78 (Klein, 2007). This correlation result is close to the 0.80 to 0.85 correlation between two human assessors on these prompts, suggesting a good level of NLP technique reliability. Beside the Klein's study, no other studies that tested the reliability of the CLA NLP system were identified. Additionally, Klein did not describe, in the paper, the 40 items assessed through the c-rater system. Consequently, it is difficult to understand on which aspects human assessors and NLP system *agree*.

In the last 20 years, a growing number of studies have been investigation how to exploit NLP systems to perform automatic assessment of CT sub-skills in CRT.

Among different NLP features, *n-grams* are growingly used in the field of automatic text analysis; they can be defined as groups of characters or words. The letter *n* refers to the number of grams included in the group. For instance, by using the term *bi-grams*, we can refer to groups of two words or syllables. N-grams are used among the linguistic features in ETS' c-rater-ML to automatically calculate the short-answers score (Heilman & Madnani, 2015). C-rater-ML specifically calculates words unigrams, words bigrams, and character n-grams (sequences of 2 - 5 characters).

The n-grams and the word count approach allow analysing the explicit content of the text. However, when evaluating the text *relevance* related to a set of concepts, information regarding the latent meaning behind the words is crucial. Latent Semantic Analysis (LSA) is a technique that provides means to extract semantic meaning from texts and compare text samples for semantic similarities (McNamara et al., 2017). Besides meaning, many other language features can be used to train algorithms in measuring the quality of a given text: parts of speech, syntax, cohesion, and syntax complexity. This information is computed through machine learning techniques to predict learning outcomes. Among these NLP features, some have been recently used to predict CT related-skills, such as argumentation (Zhu, Liu, & Lee, 2020), reflective writing (Ullman, 2019), discourse coherence quality (Burstein et al., 2013), and the use of evidence (Rahimi et al., 2017). Most of these studies applied NLP to English written texts, and there are only a few attempts to generalise these techniques to other languages.

As shown previously, a tradition of content-analysis-based human interpretation and coding for CT assessment in essays, open-ended answers, and CMC is present. In coherence with this tradition, Computerised content analysis methods could be exploited to assess CT in students written answers. Kovanovic, Joksimovic, Gaevic, Hatala, and Siemens (2017) reviewed the application of content analytics related methods and discovered that one of the earliest application domains was the student essays analysis, also known as Automated Essay Scoring (AES). Based on their analysis, the authors found that the most widely applied technique for automated essay scoring is Latent Semantic Analysis (LSA), used to measure the semantic similarity between two text bodies through the analysis of their word co-occurrence. Regarding AES, LSA can be used to calculate the resemblance of an essay to a predefined set of other essays and the

internal document similarity, often considered as a document coherence. Based on those similarities, a numeric measure of the essay quality can be calculated. Another commonly adopted method for AES is the graph-based visualisation, also based on a text's word co-occurrences. Besides approaches based on word co-occurrences, linguistic and rhetorical essays' analysis has been used to assess the quality of *argumentation* (Simsek, Buckingham Shum, Sandor, De Liddo, & Ferguson, 2013). Similar content analytics are used for other types of student-written texts, for example, short answers and online social interactions (e.g., chat, forums). In short-answers cases and AES, a set of *golden-answers* can be used to facilitate the work of automatic scoring systems.

Other methods to automatically assess CT in online discussions were based on Garrison, Anderson, and Archer's model (2001). McKlin, Harmon, Evans, and Jones (2001) developed a neural network classification system to automate discussion message coding based on the four phases described in Garrison's model: triggering, exploration, integration, and resolutions.

More recently, different studies (Kovanović, Joksimović, Gašević, & Hatala, 2014; Waters, 2015) examined the use of different text-mining techniques for coding messages based on the four stages of the Garrison's model. Kovanović et al. (2014) developed an algorithm that detected Garrison's CT related processes with the accuracy of 58.38% and Cohen's kappa of 0.41. The authors developed their algorithm by computing different linguistic features (i.e. n-grams, part-of-speech n-grams, linguistic dependency triplets, the number of mentioned concepts, and discussion position metrics).

2. Methods

2.1. Goals of the research

As shown, many attempts have been carried out in order to develop and validate tools for the automatic assessment of CT related-skills. Most of these studies applied NLP techniques to English written texts and there are a few attempts to generalize these techniques to other languages. According to the above-mentioned premises, the main goal of this research is understanding which NLP features are best associated with six CT sub-dimensions, as assessed by human evaluators in essays written in Italian: use of language, argumentation, relevance, importance, critical evaluation and novelty (Poce, 2017). We will also try to answer the following Research Questions (RQ):

1. what is the reliability level of human evaluators' assessment?
2. how students CT levels change in a university course designed to support students' CT levels?
3. what is the level of internal coherence of NLP features and how they correlate with CT sub-dimensions?

2.2. Learning activities aimed at stimulating critical thinking skills

An experimentation was carried out within a Master Degree University module in *Experimental Education and School Assessment* at the Department of Educational Sciences (Roma Tre University). The University module lasted 9 months and 202 students (F = 193; M = 7; Prefer not to say = 2; Average age: 23.3) were involved in different kinds of activities designed to foster students CT throughout two semesters. In

the first semester students attended a seminar regarding theoretical assumptions of Open Education. After that, they were required to individually search for and assess ten Open Educational Resources (OERs) on topics related to 21st skills and Museum Education for primary school children. Students looked for educational resources in OERs repositories and used a rubric for the OERs assessment developed in the context of the European Erasmus Plus Project *Open Virtual Mobility* (Poce, Amenduni, Re, & De Medio, 2019). For each OER identified, students had to assess the following six indicators: (i) quality of the explanation; (ii) Support to the lesson (iii) Quality of the assessment (iv) Quality of the instruction (v) Technological quality (vi) Promotion of Higher Cognitive Skills. For each indicator, students provided a score from 0 to 3 or they declared that an indicator was *Not Assessable*. For example, when a selected OER did not include quizzes or assessment, students inserted N/A for the indicator *Quality of the assessment*. Students were also invited to insert the link of the OER, a short abstract, the link of the OER repository used and they could also add a facultative comment. This activity was aimed both at stimulating CT *evaluative* skills and preparing students for the second semester not-mandatory assignments where students were given the possibility to design collaboratively their own OERS, following the design principles of the Project-Based Learning (PBL) methodology (Sasson, Yehuda, & Malkinson, 2018).

In previous research (Kurubacak, 2007), the process of designing OERs proved to be successful when a PBL methodology was employed to improve students CT levels. Students worked in groups in order co-construct their own OERs, by using different kind of technologies. Out of 202 students, 40 students voluntarily participated in the OER design activity by working in 8 groups. OERs, produced by the students, were assessed by the teacher through the same rubric students used in the first semester to assess the OERs retrieved from the repositories. While the OERs individual assessment assignment was mandatory and carried out fully online, the collaborative PBL activity was optional. Students who chose to participate worked in a blended modality, alternating Face to Face meeting at the university with online work. CT level were assessed through a pre-post-test methodology, described in details in the following paragraph.

2.3. CT measure

The study used a corpus of pre-post essays written in Italian language by 202 students. Students were asked to read an extract of the *Dialogue concerning two chief world systems* (Galilei, 1632) entitled *Origin of the nerves according to Aristoteles and according to the doctors* (pp. 107-108, see Appendix 1). Students completed the same test at the beginning (October 2018) and the end of the course (June 2019). Students were asked to write an essay by including in their arguments the answers to the following six questions:

1. what are the two opposite positions regarding the origin of the nerves described in the text?
2. what are the differences between the methods supported by Simplicio and Sagredo?
3. what does the *principle of authority* consist of? When is it explicitly referred to in the text and when is it implicit?
4. why do you think the episode was settled in the Republic of Venice?
5. in your opinion, has the principle of authority affected scientific discoveries throughout history? If so, how?

6. choose one or more elements in the passage that, in your opinion, have played a role in the development of scientific knowledge in the modern and contemporary world. Explain the reasons for your choice.

The choice of the Galilei's stimulus was driven by different reasons, related both with the specific characteristics of the text and the contents. Firstly, the Galilei's text can be classified as literary text. According to some authors (Paul & Elder, 2006; Poce, 2017), when students read literary texts they strive to accurately represent in their own thinking what are they are reading. Reading literary texts requires active engagement, by creating an inner dialog with the text (questioning, summarizing and connecting ideas). Galilei's literary text is characterised by the use of a figurative and allusive language. Since many implicit references are presented in the Galilei's text, students had to go beyond the available information in the task to draw inferences or make evaluations (Moss & Koziol, 1991). A further characteristic of the Galilei's text is the presence of a dialogue on a controversial topic. As shown by Fischer et al. (2009) it is more likely to prompt CT while using inconsistent or contradictory materials than consistent and coherent stimulus materials. Last but not least, the Galilei's text concerns relevant topics for the course subject in *Experimental Education and School Assessmen*t such as the role of empirical research and research methods.

2.4. Exam grades

The exam consisted on a MC questionnaire composed by 80 questions aimed at assessing students' knowledge of the course's subject. Results will be presented as percentage of correct answers provided to the MC questionnaire.

2.5. Data analysis

In this analysis, 103 students' pre-post tests were included: thus, the corpus is composed by 206 essays. All the essays were assessed by human evaluators and through an algorithm which calculates different kinds of NLP features simultaneously. One human expert assessed all the essays based on a rubric composed by six macro-indicators on a scale from 1 to 5: use of the language, argumentation, relevance, importance, critical evaluation and novelty (based on Poce, 2017).

The two remaining human evaluators assessed 80 essays (40 from the pre-test and 40 from the post-test) to perform inter-rater reliability analysis. At the same time, different NLP features were automatically measured: (i) corpus length, (ii) mean sentence length, (iii) readability (Vacca, 1972) and (iv) syntax complexity (Yang, Lu, & Weigle, 2015), since the best essays are more syntactically and semantically complex than others; (v) hapax (Poce, 2012) and (vi) lexical extension, because the more synonymous and unique words there are in a text, the better the writer (Crossley, Weston, McLain Sullivain, & McNamara, 2011), (vii) verbatim copying (Chang & Ku, 2015); (viii) TD-IDF (Salton & McGill, 1983), to evaluate how relevant a word is to a document and to a corpus on the basis of the number of times that word appears in that document and in that corpus, in order to check its relevance. Based on recent research results, TFxIDF is thought to be used to support the assessment of the sub-indicator *Novelty* (Wang, Dong, & Ma, 2019).

This because higher is the index, lower is the number of unique concepts introduced in the text compared to all the other students' text. Figure 2 describes the assumed correspondence among the six CT sub-skills identified by Poce (2017) and the selected NLP descriptors and features. Although the algorithm integrates most of NLP features

presented in the Figure 2, the measurement of some indicators relies on external tools or is not yet fully implemented in the algorithm.

CT Indicators	NLP descriptors	NLP features	State of development
Use of language	Grammar and syntax mistakes	https://scuolaelettrica.it/correttore/correttorea.php https://www.prepostseo.com/grammar-checker	Used as external tools
	Lexicon	Corpus Length; Mean Sentence Length (MSL); Hapax: $V1^2/N$ x 100; Lexical extension;	Implemented in the algorithm
Justification/ Argumentation	Readability	Flesch reading; F(Reading ease) = 206 – (0,65 x ASW) – ASL[3]	Implemented in the algorithm
	Syntax complexity	Tint (The Italian NLP Tool) was used to count the number of syntactic patterns, typical of persuasive and argumentative texts (e.g. adverb + adjective + conjunction + adjective) included in an essay;	
Relevance	Topics' relevance to a document and to a corpus	TF-IDF (Term Frequency-Inverse Document Frequency) = (sum of all P (t) of: R (p)) / PTotals[4] Higher is TF-IDF, higher will be level of relevance of the essay	Not yet implemented in the algorithm
Importance	Coherence, semantic similarity	LSA (Latent Semantic Analysis). Co-occurrence statistics on the content words preceding and following the target word; then weighting of the occurrences and reduction of dimensionality-	Not yet implemented in the algorithm
Critical evaluation	Degree of personal elaboration	*Verbatim copying*: number of instances of verbatim copying/four main concepts × the number of students	Implemented in the algorithm
Novelty	Divergent Thinking	TF-IDF. Lower is TF-IDF, higher will be level of relevance of the essay	Implemented in the algorithm

Figure 2. Correspondence among six CT sub-skills and the selected NLP descriptors and features.

Different statistical tests have been adopted. Descriptive statistics (average, frequencies, SD) were used to describe the sample features and the main variables under investigation. Welch's unequal variance t-test was used when we wanted to test the hypothesis that two populations have equal means. Welch's unequal variance t-test is an adaptation of Student's t-test, and is more reliable when the two samples have unequal variances and/or unequal sample sizes (Ruxton, 2006). Quadratic-Weighted Kappa (QWK) and Pearson product-moment correlation index was adopted to assess the degree of agreement

[2] V1 is the number of words that only appears once in a work

[3] ASL: Average Sentence Length; ASW: Average Syllables per word

[4] PTotals = all words; T set of texts t; P (t) the set of words p in the text; R(p) the number of repetitions of the word p in all texts of T except t

between the expert evaluators. The QWK index is an inter-rater reliability measure, that quantifies the degree of agreement among evaluators. The QWK index is a number between 0 and 1, in which 0 indicates the absence of agreement and 1 the perfect agreement (Fleiss & Cohen, 1973). The correlation index of Pearson is another index that allows for the evaluation of the degree of agreement consistency between two evaluators. High levels of inter-rater agreement show that other evaluators, using the same rubric, would reach similar evaluation results, thus proving the evaluation tool is reliable. Kendall's tau-b (τb) correlation coefficient (Kendall's tau-b, for short) was used to calculate correlation between NLP features and CT indicators as assessed by human evaluators.

3. Results

Figure 3 shows descriptive features of the group of participants. It is composed by 103 (F = 96; M = 7) Master Degree students enrolled in the course of *Experimental education and School Assessment*. 26 out of 103 attended all the course activities in blended modality whilst the remaining 77 students attended only the online activities.

Variables	Values	Frequency
Gender	Male	7
	Female	96
Attendance	100% online	77
	Blended	26
Exam grades %	Less than 45%	17
	Between 46% and 55%	22
	Between 56% and 65%	17
	Between 66% and 75%	23
	Between 75% and 80%	13
	Higher than 80%	2
	Missing	11
Total		**103**

Figure 3. Descriptive statistics of the group of participants.

Approximately 50% of the students passed the exam with a score higher than 60% at their first try. The lowest score at the exam was 35% of correct answers and the highest was 86.25% (Average = 60.53; SD = 13.72). In the pre-test, students spent in average 58 minutes (SD = 23.36) to complete the CT essay whilst in the post-test students spent in average 30,5 minutes (SD = 29.73). Regression analysis suggest that time to complete the essay test do not contribute to explain the variability in CT scores, neither in pre-test or post-test. Welch's unequal variance t-test found no significant difference in scores between men (M = 19.85, SD = 4.99) and women (M = 17.23, SD = 4.39) on CT total score (p = .067). Thus, neither gender and time to complete the assessment could explain the variability in CT total score.

3.1. Critical Thinking Human Assessment Reliability

Cronbach's alpha was used to test the internal reliability of items in the CT test. Cronbach's alpha value is 0.894. Utilising Ponterotto and Ruckdeschel's (2007) reliability matrix, an alpha of 0.85 or above is deemed to be excellent. Thus, Alpha

indicates a high level of internal consistency for our scale with this specific group of participants. Figure 4 shows the correlation between each CT sub-indicators and CT total.

	Correlation between item and total score	Internal reliability with item removed
Use of language	.667	.883
Argumentation	.754	.869
Relevance	.551	.898
Importance	.813	.860
Critical Evaluation	.807	.861
Novelty	.706	.877

Figure 4. Cronbach's Alpha values for CT sub-indicators assessed by human evaluators.

Three graders marked responses on the CT test scores developed by Poce (2017) in order to test inter expert reliability. Figure 5 presents the results. Use of language and Argumentation obtained the higher level of agreement between evaluators whilst Critical Evaluation the lowest. The overall inter-rater reliability is medium to high, which suggest there is still room for improvement in terms of inter expert reliability.

	Correlation	QWK
Use of Language	0.815**	0.803**
Argumentation	0.768**	0.742**
Relevance	0.635**	0.488**
Importance	0.599**	0.503**
Critical Evaluation	0.534**	0.430**
Novelty	0.633**	0.549**

Figure 5. Intercoder agreement between experts.

3.2. Comparison of Critical Thinking pre-post test scores

The distribution of the CT total score is close to normal distribution both in pre and post-test (see Figure 6 for a population pyramid of CT total scores in the pre-tests and in the post-tests).

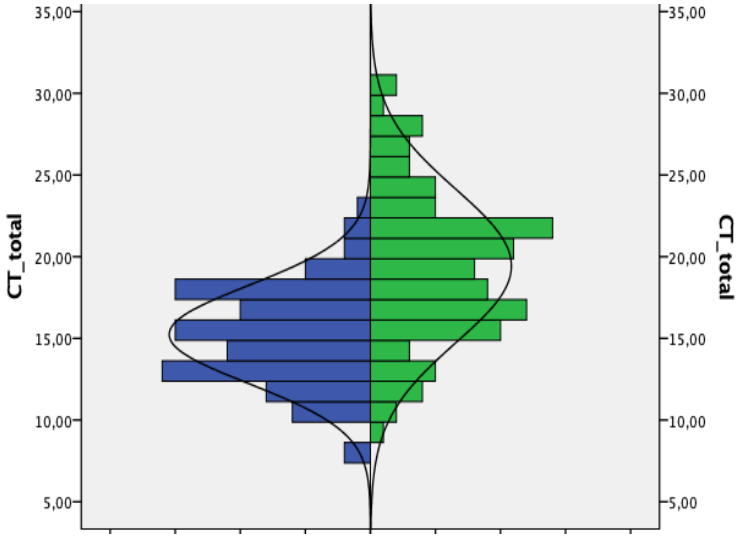

Figure 6. A population pyramid of CT total scores in the pre-tests and in the post-tests.

The mean score on the CT test was respectively 15.24 (SD = 2.99) in the pre-test and 19.43 (SD = 4.69) in the post-test. We used Welch's unequal variance t-test to compare the difference between pre-post-test CT total score. Welch's unequal variance t-test found a statistically significant difference between pre and post CT total score ($p < .000$).

We investigated the difference between CT sub-indicators, as assessed by human experts. Figure 7 shows the comparison of the averages obtained for each sub-indicator. Welch's unequal variance t-test found a statistically significant difference between pre and post CT total score ($p < .000$) for all the CT sub-indicators. For almost all the CT sub-indicators, the average was lower than 3 (the median score) in the pre-test, with the exception of the sub-indicator *relevance*. In the post-test, the average was always higher than 3, except for the *novelty* indicator. This suggest that the group of students in average shift from insufficient to sufficient scores. We tried to understand if difference in CT scores could be explained by the attendance of the blended course vs 100% online course.

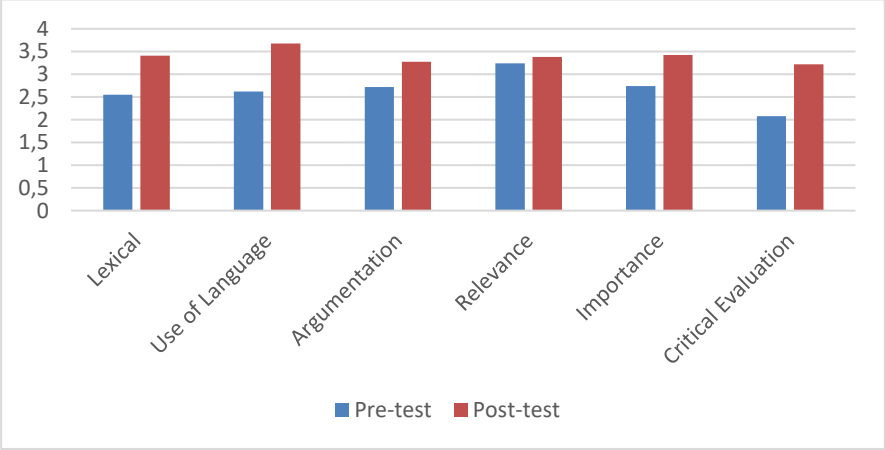

Figure 7. Comparison of the average scores between pre and post-tests as assessed by human evaluators * < 0.05; ** < 0.01; *** <0.001.

No statistical difference has been identified in CT level post-test (as assessed by human evaluators) between students who attended the course activities in blended modality compared with students who completed only the online activities.

Figure 8 shows the difference between pre-post of blended course attendance vs online 100% students in CT-total average. Both the groups started from similar CT total scores average. Both the groups improved in the post-tests but students who attended only the online activities (blue line) improved a little bit more (Average = 19.92) compared with students who attended the blended activities (green line, Average = 18.00). However, differences in the post-test were not statistically significant between the groups. Thus, the kind of attendance could not explain the difference between pre-post-test.

On the other hand, participants who attended the course had a higher score in the *Exam-grade* (Average = 64.12) compared with students who didn't (Average = 59.23), although the difference between the average is not statistically significant.

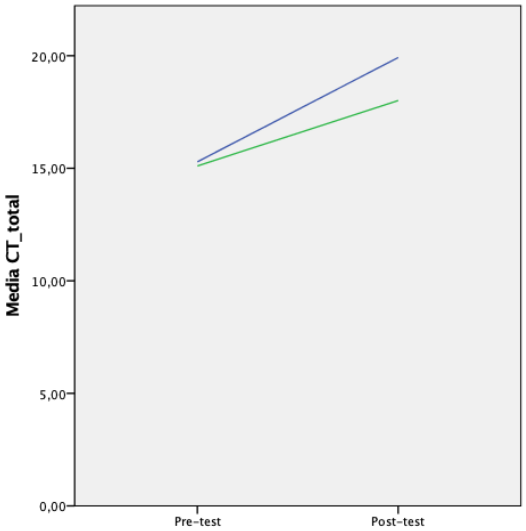

Figure 8. Difference between pre-post of blended course (green line) vs online 100% (blue line) students' attendance in CT-total average.

Figure 9 shows the difference between pre-post of students who obtained an exam grade not sufficient (green line) and a sufficient exam grade (yellow line) in CT-total average. In the pre-test, students who obtained in the final exam an insufficient grade had slightly lower average (Average = 14.69) compared to students who obtained a sufficient grade at the end of the exam (Average = 15.81). However, the difference in CT pre-test was not statistically significant between these groups. Both groups improved in their CT level in the post-tests and the difference between the two groups' averages was reduced, as showed in the Figure 9.

Kendall's tau-b (τb) correlation coefficient was used to explore correlation between CT score pre-test and exam grade (Figure 10). A low and significant correlation has been identified with two out of six CT sub-indicators: *use of language* and *argumentation*.

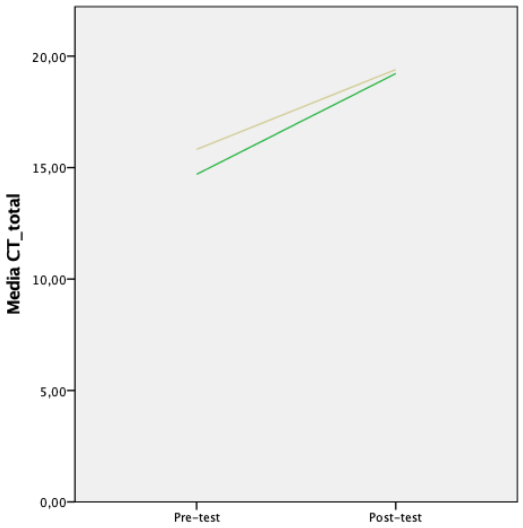

Figure 9. Difference between pre-post of students who obtained a not sufficient (green line) a (yellow line) sufficient exam grade in CT-total average.

		Use of language	Argumentation	Relevance	Importance	Critical Evaluation	Novelty
Exam grade	Tau	.170*	.198*	.049	-.014	.108	.066
	Sign	.032	.014	.547	.866	.187	.418
	N	88	88	88	88	88	88

Figure 10. Correlation between CT sub-indicators and exam grade.

3.3. NLP features' properties

Figure 11 presents descriptive statistics related with the NLP indicators extracted from the students' essays. The essay length was, in average, composed by 245 words and the average length of each sentence in the essay was approximately of 37 words. The Hapax index was in average 37. This indicates that, in average, each essay was composed by 37% of words used only one time. Lexical extension was 69. This indicates that in an essay, the range of words used is of 69%. Reading ease was in average of 27.32 which indicates that a text is written in a complex form, typically adopted by people with higher levels of education. The complexity of syntax is on average 24 which means that students used in average 24 complex argumentative syntax forms in their essays. Students, in average, copied verbatim 6 times in their essay the words of the test' questions. The TfxIDF was 39.38 in average. This means that each text contains in average 39% of words that are not used in other students' texts.

According to the assumptions presented in Figure 2, five NLP indicators can be useful to support the assessment of the CT sub-indicator *Use of Language*: 1. Hapax 2. Lexical Extension, 3. Corpus Length 4. MSL; 5. Verbatim copying. All these indicators are expected to be related with the Lexicon use. The correlation between hapax and lexical extension is indeed strong: $\tau b = 0{,}853$ sign. 0.000 (Figure 12).

	Min	Max	Average	SD
Corpus Length	91.00	456.00	245.3795	68.92533
MSL	16.85	73.13	36.9976	9.90324
Hapax	31.36	73.74	52.0249	7.40308
Lexical extension	51.75	85.86	69.7351	5.71921
Reading ease	-376.74	56.38	27.3201	31.27196
Syntax complexity	6.00	61.00	24.2205	10.02870
Verbatim copying	.00	16.00	6.3	.00720
TFxIDF	21.17	51.50	39.3863	5.57181

Figure 11. Descriptive statistics related with the NLP indicators.

Corpus Length negatively correlate with both hapax and lexical extension. This means that higher is the number of words used in an essay, lower is the probability that people use unique words in the text (Hapax) and a higher range of words (Lexical Extension). Verbatim copying moderately and negatively correlates with lexical extension. This means that students who copy verbatim the words used in test' questions use a lower range of words. According to the assumptions presented in Figure 2, NLP features associated with the sub-indicator *Argumentation* are 1. Reading ease 2. Syntax complexity. A moderate negative correlation has been identified between these two

indicators. This means that more difficult is a text to read, higher is the complexity of its syntax. Moreover, syntax complexities strongly correlate with Corpus Length τb = 0.543 sign .000. This means that higher is the number of words used in a text, higher is the number of complex structures adopted in that text. TFxIDF was thought to be used to support the assessment of the sub-indicator *Novelty*. TFxIDF negatively correlate with Lexical Extension and Hapax and positively correlates with the number of words used, syntax complexity and repetition.

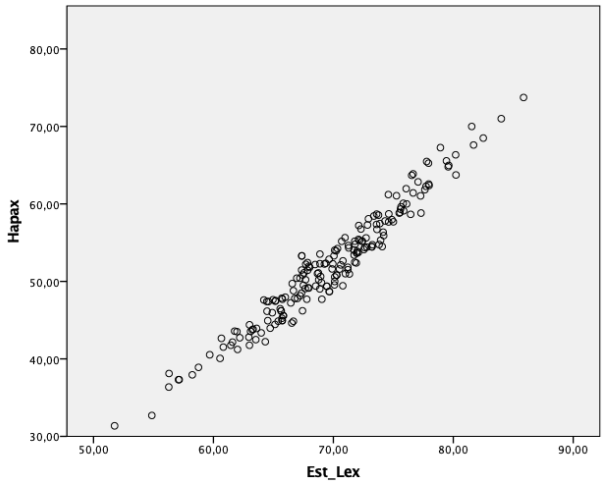

Figure 12. Scatterplot for the correlation between Hapax and Lexical Extension.

		MSL	Hapax	Lexical extension	Reading Ease	Syntax	Verbatim copying	TFxIDF
Corpus Length	τb	.101*	-.376**	-.421**	-.147**	.543**	.040	.612**
	Sig	.036	.000	.000	.002	.000	.434	.000
MSL	τb	1.00	-.060	-.069	-.694**	.083	.049	.052
	Sig	.	.210	.153	.000	.087	.336	.468
Hapax	τb		1	.853**	.021	-.216**	-.088	-.315**
	Sig		.	.000	.667	.000	.082	.000
Lexical extension	τb		.853**	1	.044	-.249**	-.115*	-.321**
	Sig		.000	.	.359	.000	.022	.000
Reading ease	τb		.021	.044	1	-.245**	-.021	-.026
	Sig		.667	.359	.	.000	.673	.723
Syntax complexity	τb		-.216*	-.249**	-.245**	1	-.019	.445**
	Sig	.087	.000	.000	.000	.	.709	.000
Verbatim copying	τb	.049	-.088	-.115*	-.021	-.019	1	.121
	Sig	.336	.082	.022	.673	.709	.	.091

Figure 13. NLP features internal coherence.

Three NLP indicators significantly correlate with CT total score. The Corpus Length, the complexity of the syntax, and the TFxIDF (Figure 14).

		Corpus Length	Syntax complexity	TFxIDF
CT total score	Tb	.198**	.230**	.228**
	Sign	.000	.000	.001
	N	195	195	195

Figure 14. Correlation between NLP features and CT total score.

Figure 15 presents the correlation between the 6 CT sub-indicators, as assessed by human experts, and five NLP indicators. The CT sub-indicator Use of Language is moderately and negatively associated with the average sentence length (Figure 16). Average sentence length included between 15 and 45 correspond to score higher than 3 in the evaluation of the sub-indicator Use of Language. On the other hand, when the average sentence length is higher than 45, the assessment of the sub-indicator Use of Language is not sufficient. According to our expectations, the use of complex argumentative syntax forms correlates with the sub-indicator Argumentation, although the correlation is moderate. This association was graphically explored in the Figure 17, where it is possible to see that a higher numbers of complex syntax forms correspond to higher level in Argumentation scores as assessed by human experts. Syntax complexity is also significantly associated with all the others CT sub-indicators. As expected, one of the strongest positive correlation found was between TfxIDF and the sub-indicator Relevance. Higher is the TfxIDF, higher is the coherence with words and concepts used in one essay and the other ones. In other words, higher is the TfxIDF, higher is the relevance of the topics covered in an essay in relation to the others (Figure 18). Contrary to our expectations, TfxIDF moderately and positively correlates with novelty.

		Corpus Length	MSL	Est Lex	Syntax complexity	TFxIDF
UOL_human	τb	.084	-.146**	.056	.144**	.156*
	Sign	.100	.004	.272	.006	.029
Arg_human	τb	.155**	-.071	.014	.175**	.181*
	Sign	.003	.177	.797	.001	.011
Rel_human	τb	.274**	.004	-.126	.235**	.208**
	Sign	.000	.942	.020*	.000	.003
Imp_human	τb	.175**	-.080	.003	.201**	.203**
	Sign	.001	.131	.960	.000	.004
CE_human	τb	.118*	-.099	-.009	.152**	.199**
	Sign	.026	.062	.871	.005	.005
Nov_human	τb	.203**	-.044	-.054	.174**	.141*
	Sign	.000	.412	.312	.001	.050

Figure 15. Correlation between the 6 CT sub-indicators, as assessed by human experts, and five NLP indicators.

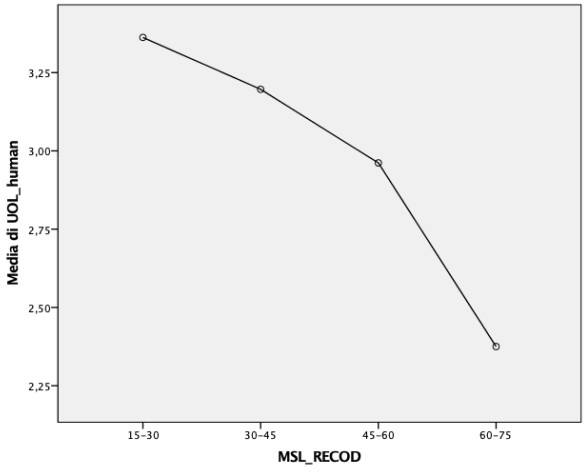

Figure 16. Graphic representation of the correlation between MSL and Use of Language.

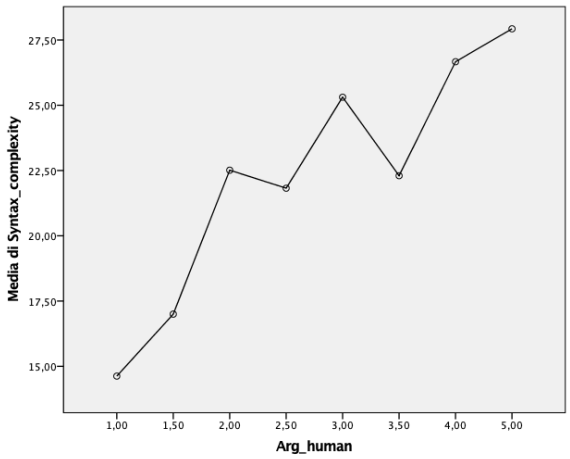

Figure 17. Graphic representation of the correlation between Syntax Complexity and Argumentation.

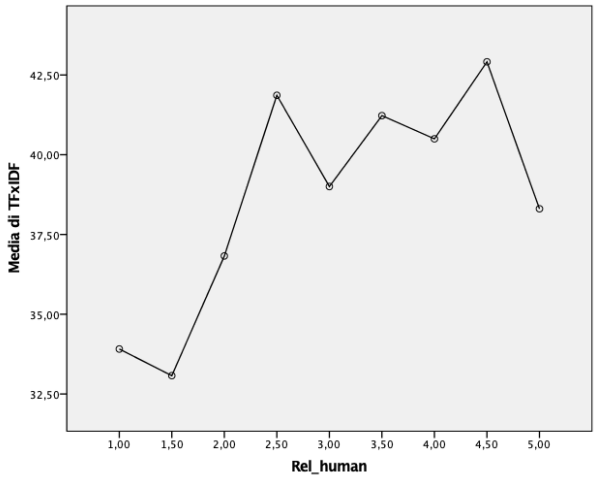

Figure 18. Graphic representation of the correlation between TFxIDF and Relevance.

The only indicator that correlates with the exam grade is *syntax complexity* ($\tau b = 0.129$ sign < .01), which could denote rooms for improvement in terms of criterion validity of NLP towards academic performance.

4. Discussion and final remarks

The absence of a shared definition of CT has led to the development of multiple methods and tools for the evaluation of this set of skills, dispositions and behaviours. On one side, a high number of tests are available in the standardised testing market (Rear, 2019). On the other side, a recent literature review showed how non-standardised instruments created *ad hoc* by the teacher and by the researcher are frequently used too (Tiruneh, Verburgh, & Elen, 2014). In this work, we tried to take an intermediate position between the need to assess CT validly and ecologically from one side and the priority to guarantee measurement validity and reliability on the other side. In our perspective, it is possible to

observe CT manifestations or, instead, failures in its application, in complex communicative acts, mediated by the use of language. For this reason, it is believed that the evaluation of CT within CRT guarantees the highest levels of external and ecological validity. Having said that, we acknowledge that CRT scoring can be time consuming and expensive both for teachers and researchers. For this reason, our research team has been working in the direction of *automated scoring* to support and facilitate the scoring of students' written responses. International research shows that the automatic scoring of students written answers has achieved good level of reliability, in specific cases and with English language. However, there are only a few attempts to generalize these techniques to other languages. The present research has started to fill this gap by investigating which NLP features are best associated with six CT sub-dimensions, as assessed by human evaluators, in essays written in Italian: use of language, argumentation, relevance, importance, critical evaluation and novelty (Poce, 2017). The experimentation was carried out with 103 students enrolled in a Master Degree course in *Experimental Education and School Assessment* at Roma Tre University. The course includes different activities aimed at stimulating students CT: OERs individual assessment and OERs collaborative design. The first activity was mandatory and all the students completed it online. The second activity was optional and it was completed by a total of 40 students. Students had to work in groups and alternate Face to Face meeting with online work. In this work, we tried to answer to three RQs. The first RQ concerns the reliability level of human evaluators' assessment. We found an excellent internal reliability and a medium to high inter-coder agreement of the human evaluators. *Use of language* and *Argumentation* obtained the higher level of agreement between evaluators.

Those two indicators also correlate with students' final exam grade and have good internal coherence with CT total scores. Thus, *Use of language* and *Argumentation* can be considered two reliable and valid indicators. On the other hand, *Critical Evaluation* obtained the lowest level of agreement between evaluators. The overall inter-rater reliability is medium to high, which suggest there is still room for improvement in terms of inter expert reliability.

We also wanted to explore how students CT levels change in the university course designed to support students' CT levels. Students CT level improved significantly in the post-test. We compared the CT students' performance of 100% online and blended attendance. Both the groups improved in the post-tests but students who attended only the online activities improved a little bit more (Average = 19.92) compared with students who attended the course in a blended modality (Average = 18.00). However, differences in the post-test were not statistically significant between the groups. Thus, the kind of attendance could not explain the difference between pre-post-test.

In the pre-test, students who obtained in the final exam an insufficient grade had slightly lower average (Average = 14.69) compared to students who obtained a sufficient grade at the end of the exam (Average = 15.81). However, the difference in CT pre-test was not statistically significant between these groups. Both groups improved in their CT level in the post-tests and the difference between the two groups average was reduced. This suggests that CT course entry level could be used to predict students' final exam grade. Moreover, it is possible that the course design had a stronger effect on students' CT level with a lower level of academic preparation. Further research would be necessary to test those hypotheses. In our last research question, we wanted to explore the level of internal coherence of NLP features and how they correlate with CT sub-dimensions.

According to the expectations, we found correlations between 5 NLP features assumed to be associated with the CT sub-indicator *Use of Language*: 1. Hapax 2. Lexical Extension, 3. Corpus Length 4. MSL; 5. Verbatim copying. Moreover, a moderate negative correlation has been identified between 1. Reading ease and 2. Syntax complexity, both assumed to be associate with the CT sub-indicator *Argumentation*. This means that more difficult is a text to read, higher is the complexity of its syntax. Three NLP indicators significantly correlate with CT total score. The Corpus Length, the complexity of the syntax, and the TFxIDF. As expected, the Medium Sentence Length negatively correlate with the CT sub-indicator *Use of Language*. Complexity of the syntax positively correlate with *Argumentation.* In addition, TFxIDF positively correlates with *Relevance*. On the other hand, some of our expectations were not confirmed.

Although Hapax and Lexical Extension correlates, they did not show any significant correlation with the expected CT sub-indicator *Use of Language*. This result can be explained by an issue retrieved within many of the essays assessed. We found that students often used a not coherent language within their essays, by alternating refined with everyday expressions. In this condition, whilst human evaluators provide low scores to the CT sub-indicator *Use of Language*, the algorithm can find good level of Hapax (number of unique words in the text) and Lexicon Extension.

The second expectations not confirmed concerns the correlation between TFxIDF and the CT sub-indicator *Novelty*. We expected a negative correlation between these indicators (based on Wang, Dong, & Ma, 2019) but we found a moderate and positive correlation. In previous studies, TFxIDF was used to assess divergent forms of novelty. However, the *Novelty* required in CT should be convergent. Indeed, divergent thought from a single starting point generates varied ideas, whereas convergent thought starting from multiple points seeks one most true or useful conclusion (Brophy, 2001).

In our case, it is likely that algorithm and human evaluators looked for different forms of *Novelty*: the algorithm retrieved divergent new ideas, whilst human evaluators search for convergent new ideas.

This study was exploratory in nature and we acknowledge its limitations: in future studies, we would need to expand the sample size, by including the remaining 200 essays collected. It is necessary to collect a large amount of data so that it is possible to conduct some *training* in *machine learning* mechanisms for the implementation of the NLP prototype performance (Grimmer & Stewart, 2013). At that moment, due to the limited number of cases, it had not been possible to create a *training set* for the application of a *supervised learning model*. We have also explored a limited number of NLP indicators because of the difficulties to find out Open Tools for Italian Language to be incorporated in our automatic system. In future studies, we are going to use larger corpora and to test new NLP features. These improvements will allow us to carry out more sophisticated statistical analysis such as structural equation modelling and Latent Factor Analysis (MacArthur, Jennings, & Philippakos, 2019). In future studies, we will test the approach with other kinds of open-ended answers produced by HiEd students, both in the field of humanities and STEM disciplines.

Appendix 1

Sagredo. One day I was at the home of a very famous doctor in Venice, where many persons came on account of their studies, and others occasionally came out of curiosity to

see some anatomical dissection performed by a man who was truly no less learned than he was a careful and expert anatomist. It happened on this day that he was investigating the source and origin of the nerves, about which there exists a notorious controversy between the Galenist and Peripatetic doctors. The anatomist showed that the great trunk of nerves, leaving the brain and passing through the nape, extended on down the spine and then branched out through the whole body, and that only a single strand as fine as a thread arrived at the heart. Turning to a gentleman whom he knew to be a Peripatetic philosopher, and on whose account he had been exhibiting and demonstrating everything with unusual care, he asked this man whether he was at last satisfied and convinced that the nerves originated in the brain and not in the heart. The philosopher, after considering for a while, answered: "You have made me see this matter so plainly and palpably that if Aristotle's text were not contrary to it, stating clearly that the nerves originate in the heart, I should be forced to admit it to be true."

Simplicio. Sir, I want you to know that this dispute as to the source of the nerves is by no means as settled and decided as perhaps some people like to think.

Sagredo. Doubtless it never will be, in the minds of such opponents. But what you say does not in the least diminish the absurdity of this Peripatetic's reply; who, as a counter to sensible experience, adduced no experiment or argument of Aristotle's, but just the authority of his bare ipse dixit.

Reference list

Andrews, D. H., & Wulfeck, W. H. (2014). Performance assessment: Something old, something new. In J. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 303-310). New York, NY: Springer.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795–808. http://dx.doi.org/10.1177/0013164414527450 (ver. 15.12.2020).

Brophy, D. R. (2001). Comparing the attributes, activities, and performance of divergent, convergent, and combination thinkers. *Creativity research journal*, *13*(3-4), 439–455.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 267-280). New York, NY: Routledge.

Chang, W. C., & Ku, Y. M. (2015). The effects of note-taking skills instruction on elementary students' reading. *The Journal of Educational Research*, *108*(4), 278–291.

Chou, T. L., Wu, J. J., & Tsai, C. C. (2019). Research trends and features of critical thinking studies in e-learning environments: A review. *Journal of Educational Computing Research*, *57*(4), 1038–1077.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311.

Davies, M., & Barnett, R. (Eds.). (2015). *The Palgrave handbook of critical thinking in higher education*. New York, NY: Springer.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into practice*, *32*(3), 179–186.

Ennis, R. H., & Weir, E. E. (1985). *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Pacific Grove: Midwest Publications.

ETS c-rater. https://www.ets.org/accelerate/ai-portfolio/c-rater (ver. 15.12.2020).

ETS. e-rater. https://www.ets.org/erater/about (ver. 15.12.2020).

Fischer, S. C., Spiker, V. A., & Riedel, S. L. (2009). Critical thinking training for army officers. Volume 2: A model of critical thinking (Technical Report). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7–23.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Halpern, D. (2016). *Manual Halpern Critical Thinking Assessment*. Mödling, Austria: Schuhfried GmbH.

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 81–85.

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, *44*, 1–8.

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In D. Nolan & T. Speed (Eds.), *Probability and Statistics: Essays in Honor of David A. Freedman* (Vol. 2) (pp.76-89). Beachwood, OH: Institute for Mathematical Statistics

Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C., ... & Steedle, J. C. (2009). Test Validity Study (TVS) Report. https://cp-files.s3.amazonaws.com/26/TVSReport_Final.pdf (ver. 15.12.2020).

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated content analysis of online discussion transcripts. In K. Yacef & H. Drachsler (Eds.), *Proceedings of the Workshops at the LAK 2014 Conference* (LAK-WS 2014), 24-28 March 2014, Indiana, USA. http://ceur-ws.org/Vol-1137/LA_machinelearning_ submission_1.pdf (ver. 15.12.2020).

Kovanovic, V., Joksimovic, S., Gaevic, D., Hatala, M., & Siemens, G. (2017). Content analytics: The definition, scope, and an overview of published research. In C. Lang, G. Siemens, A. F. Wise, & D. Gaevic (Eds.), *The handbook of learning*

analytics (pp. 77-92). Alberta, Canada: Society for Learning Analytics Research (SoLAR).

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, *4*(1), 70–76.

Kurubacak, G. (2007). Building knowledge networks through project-based online learning: A study of developing critical thinking skills via reusable learning objects. *Computers in human behavior, 23*(6), 2668–2695.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*(1), 1–23.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. *Reading and Writing*, *32*(6), 1553–1574.

McKlin, T., Harmon, S. W., Evans, W., & Jones, M. G. (2001). Cognitive presence in Web-based learning: A content analysis of students' online discussions. Annual *Proceedings of the National Convention of the Association for Educational Communications and Technology*, Atlanta, GA. https://eric.ed.gov/?id=ED470101 (ver. 15.12.2020).

McNamara, D., Allen, L., Crossley, S., Dascalu, M., & Perret, C. (2017). Natural Language Processing and Learning Analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gaevic, (Eds.), *The handbook of learning analytics* (pp. 93-104). Alberta, Canada: Society for Learning Analytics Research (SoLAR).

Moss, P. A., & Koziol Jr, S. M. (1991). Investigating the validity of a locally developed critical thinking test. *Educational Measurement: Issues and Practice, 10*(3), 17–22.

Newman, D. R., Johnson, C., Webb, B., & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co-operative learning. *Journal of the American Society for Information science*, *48*(6), 484–495.

Newman, D. R., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology, 3(*2), 56–77.

Nicol, D. (2007). E-assessment by design: Using multiple-choice test to good effect. *Journal for Further and Higher Education, 31*, 53–64.

OECD. Organisation for Economic Co-operation and Development (2012). Assessment of higher education learning outcomes. AHELO Feasibility Study Report. http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf (ver. 15.12.2020):

Paul, R., & Elder, L. (2006). *The miniature guide to critical thinking: Concepts & tools*. Dillon Beach, CA: Foundation for Critical Thinking.

Poce, A. (Ed.). (2012). *Contributi per la definizione di una tecnologia critica: un'esperienza di valutazione*. Milano: FrancoAngeli.

Poce, A. (2017). *Verba sequentur: pensiero e scrittura per uno sviluppo critico delle competenze nella scuola secondaria*. Milano: FrancoAngeli.

Poce, A., Amenduni, F., Re, M. R., & De Medio, C. (2019). Establishing a MOOC quality assurance framework. A case study. *Open Praxis, 11*(4), 451–460.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and motor skills, 105*(3), 997–1014.

Popham, W. J. (2003). *Test better, teach better. The instructional role of assessment*. Alexandria, VA: ASCD.

Prepostseo. Online Grammar Checker https://www.prepostseo.com/grammar-checker (ver. 15.12.2020).

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. International *Journal of Artificial Intelligence in Education, 27*(4), 694–728.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education, 44*(5), 664–675.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688–690.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill Book Co.

Sasson, I., Yehuda, I., & Malkinson, N. (2018). Fostering the skills of critical thinking and question-posing in a project-based learning environment. *Thinking Skills and Creativity, 29*, 203–212.

Scuola elettrica. Correttore grammaticale. https://scuolaelettrica.it/correttore/correttorea.php (ver. 15.12.2020).

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher, 29*(7), 4–14.

Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A., & Ferguson, R. (2013). *XIP Dashboard: Visual analytics from automated rhetorical parsing of scientific metadiscourse*. Paper presented at the 1st International Workshop on Discourse-Centric Learning Analytics, 8 April 2013, Leuven, Belgium. http://oro.open.ac.uk/37391/1/LAK13-DCLA-Simsek.pdf (ver.15.12.2020).

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies, 4*(1), 1–17.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257.

Vacca, F. (1972). *Per una critica quantitativa: romanzi a chilometri*. Il Messaggero.

Wagenaar, R. (2018). What do we know–What should we know? Measuring and comparing achievements of learning in European Higher Education: initiating the new CALOHEE approach. In O. Zlatkin-Troischansaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn, *Assessment of learning outcomes in higher education* (pp. 169-189). Cham: Springer.

Wang, K., Dong, B., & Ma, J. (2019). Towards computational assessment of idea novelty. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Honolulu, United States.

Waters, Z. (2015). *Using structural features to improve the automated detection of cognitive presence in online learning discussions* (B.Sc. Thesis). Queensland University of Technology.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668 (ver. 15.12.2020).