Automatic Assessment of University Teachers' Critical Thinking Levels

https://doi.org/10.3991/ijac.v12i3.11259

Antonella Poce, Francesca Amenduni, Maria Rosaria Re,
Carlo De Medio
University of Roma Tre, Rome, Italy
antonella.poce@uniroma3.it

Abstract—The present work describes the structure of a pilot study which was addressed to test a tool developed to automatically assess critical thinking - CT levels through language analysis techniques. Starting from a Wikipedia database and lexical analysis procedures based on n-grams, a new approach aimed at the automatic assessment of the open-ended questions, where CT can be detected, is proposed. Automatic assessment is focused on four CT macro-indicators: basic language skills, relevance, importance and novelty. The pilot study was carried out through different workshops adapted from Crithinkedu EU Erasmus + Project model aimed at training university teachers in the field of CT. The workshops were designed to support the development of CT teaching practices at higher education levels and enhance University Teachers' CT as well. The two-hour workshops were conducted in two higher educational institutions, the first in the U.S.A (CCRWT Berkeley College NYC, 26 university teachers) and the second in Italy (Inclusive memory project - University Roma Tre, 22 university teachers). After the two workshops, data were collected through an online questionnaire developed and adapted in the framework of the Erasmus + Crithinkedu project. The questionnaire includes both open-ended and multiple-choice questions. The results present CT levels shown by university teachers and which kind of pedagogical practices they intend to promote after such an experience within their courses. In addition, a comparison between the values inferred by the algorithm and those calculated by domain human experts is offered. Finally, follow-up activity is shown taking into consideration other sets of macro-indicators: argumentation and critical evaluation.

Keywords—Critical thinking, automatic assessment, higher education

1 Introduction

In recent years, new ways to define and assess critical thinking assessment have been developed. Big amounts of behavioral data connected to learning processes are stored automatically in digital platforms (e.g. social media, LMS). The analysis of data collected from virtual learning environments has attracted much attention from different fields of study; therefore, a new research field is born, known as learning analytics.

In addition, in the field of critical thinking assessment, many researchers agree that multiple assessment formats are needed for critical thinking assessment. However, the use of open-ended questions raises some problems concerning the high cost of human scoring. Automated scoring could be a viable solution to these concerns, with automated scoring tools designed to assess both short answers and essay questions (Liu, Frankel, & Roohr, 2014). In the field of critical thinking assessment, Gordon, Prakken and Walton (2007) proposed a functional model for the evaluation of arguments in dialogical and argumentative contexts. Wegerif and colleagues (2010) described a computational model to identify places within e-discussion in which students adopt critical and creative thinking. Developing a computational model to identify critical thinking in students' written comments provides many advantages such as assisting the researcher in finding key aspects in big amounts of data and helping a teacher or a moderator identify when students are thinking more critically.

Despite the advantages, it is important to examine the accuracy of automated scores to make sure they achieve an acceptable level of agreement with valid human scores. Liu and colleagues (2014) asserted that, in many cases, automated scoring can be used as a substitute for the second human rater and can be compared with the score from the first human rater. If discrepancies beyond what is typically allowed between two human raters occur between the human and machine scores, additional human scoring will be introduced for adjudication.

In this work, we present a pilot study carried out to test a tool developed to automatically assess critical thinking levels through language analysis techniques. Starting from a Wikipedia database and the use of lexical analysis based on n-grams, we propose a new approach aimed at the automatic assessment of the open-ended questions, where critical thinking levels can be detected. The prototype devised so far is based on code framework developed in previous research (Poce, Corcione & Iovine, 2012; Poce, 2015) mainly inspired by the Newman, Webb and Cochrane model (1995). The above framework is composed of six macro-indicators: basic language skills, justification, relevance, importance, critical evaluation and novelty. The first macro-indicator, namely basic language skills, is useful to assess the correct use of the language. The justification macro-indicator evaluates students' ability to elaborate on their thesis and support their arguments throughout a discourse. Relevance is a macro-indicator that analyzes students' texts consistency, such as the correct use of outlines and students' capability to accurately use given *stimuli*. The importance macro-indicator evaluates the knowledge students use in their discourse. Finally, critical evaluation and novelty refer to personal and critical elaboration of sources, data and background knowledge with the use of new ideas and solutions associated with the initial hypothesis and students' personal thesis. At the moment, the prototype has been designed to assess four areas out of six: basic language skills, relevance, importance and novelty. To test the employability of the tool, we carried out a pilot study through a workshop adapted from Crithinkedu EU Erasmus + Project training course for university teachers. The workshop was designed to support both the development of critical thinking teaching practices at higher education levels and enhance university teachers' critical thinking.

2 The Context of the Research, Research Questions and Objectives

In the context of the Crithinkedu EU Erasmus + Project 'Critical Thinking Across the European Higher Education Curricula', funded by the European Commission under the Erasmus+ Program, a specific *training course*, aimed to improve the quality of CT teaching and learning in universities across the curricula, was designed. The main idea underpinned by the training course is that HE teachers do not only need to be trained about methods to improve critical thinking in their students, but also to develop a critical thinking attitude themselves within their own professional practices. Indeed, critical thinking development in higher education is often considered a priority not only because it improves deep-comprehension ability and allows teachers to be effective in the workplace, but also because critical thinking is a necessary mindset to be active citizens of the wider social environment (Davies & Barnett, 2015). For this reason, university teachers need to receive the proper training to incorporate critical thinking instructions into their curricula. From previous research, it is clear that improvement in students' CT skills and dispositions cannot be a matter of *implicit expectations* (Marin & Halpern, 2011; Dominguez, 2018). Educators should make CT objectives explicit and include them in training and faculty development. In addition, a gap between university and workplaces' expectations (Dominguez, 2018) was observed and defined in terms of who is a "critical thinker" and what he/she should be able to do.

All the above taken into consideration, the research group identified the following research questions:

- How are CT objectives made clear in the HE curricula?
- How do university teachers interpret CT skills and dispositions?
- Which levels of critical thinking are shown by the university teachers' sample analyzed and which kind of pedagogical practices do they intend to promote after the workshop?
- Is it possible to automatize CT assessment?
- If yes, can CT automatic assessment support the human one?
- Can a tool based on a language-analysis procedure be useful to assess the macro-indicators present in the Newman, Webb and Cochrane adapted model?
- How much do the values inferred by the algorithm predict the values calculated by domain expert?

In the first part of this paper the structure of the Crithinkedu adapted workshop from the *training course* model (Dominguez, 2018) for university teachers is described. The workshop model is aimed to

- Support the development of critical thinking teaching practices at higher education level
- Enhance university teachers' critical thinking. As mentioned, the two-hour workshop model was conducted in two higher educational institutions, the first carried

out at Berkeley College NYC (U.S.A)[1] and the second at the University of Roma Tre[2] (Italy).

3 Methodology

3.1 The workshop structure

The two workshops carried out both in the United States and in Italy followed a general structure inspired by the Crithinkedu *training course,* although there were some differences due to the specific context. The course carried out in the United States took place in the setting of the 6th Annual Conference "Defining Critical in the 21st Century[3]". The conference was devoted to critical thinking in higher education and university professors were invited because of their interest to improve their teaching and professional practices.

On the other hand, the course carried out in Italy took place in the framework of a project named "Inclusive Memory"[4]. Local university professors from different fields were involved in developing critical thinking knowledge, skills and dispositions in order to produce inclusive museum object-based learning paths to be used in their own courses. Our goal was to see if CT knowledge they have to acquire for the sake of the project would also be used in their university teaching practices.

Both the workshops lasted two hours and they were implemented bearing in mind the following objectives:

- Participants should be introduced to more general/transversal elements of CT
- Participants should be able to discuss and apply CT in their discipline/field
- Participants should be encouraged to redesign their courses aiming at the strengthening/embedding the 'teaching CT' aspects
- Participants should have the opportunity to discuss field/discipline specific instances of teaching CT

At the beginning of each activity, the goals of the workshop were explained and negotiated with the participants. Then, the most used definitions of critical thinking based on the Facione (1990) and Jiménez-Aleixandre and Puig (2012) conceptualizations were shown to the participants. After that, they were invited to reflect upon learning strategies that can be used to improve critical thinking (e.g. jigsaw methods, conceptual maps, problem and project-based learning) and on methods to assess critical thinking, according to the skills and dispositions they intended to improve. After the theoretical presentation, participants were invited to work in small groups and they

[1] CCRWT October 19th 2019 https://ccrwt.weebly.com/2018-ccrwt.html
[2] Inclusive memory project, November 6th 2019
[3] https://ccrwt.weebly.com/uploads/2/2/7/1/22712194/5683_ccrwt_program_onlinedoc_final_pdf.pdf
[4] The project is aimed to support inclusiveness of minorities and disadvantaged groups through the fruition of cultural heritage in museums and through the development of the 4Cs (Collaboration, Creativity, Communication and Critical Thinking).

were divided according to their field background (STEM, humanities, social sciences, foreign language and literature, engineering, in the case of the USA group, or to their role in the "Inclusive Memory" project, in the case of Italian group). All of them were invited to:

- **Define their CT learning goals**. Based on Facione and Jiménez-Aleixandre & Puig definitions, they had to choose which CT skills or dispositions they aimed to focus on.
- **Define their activities.** They had to decide methods that could support the chosen CT skill or disposition to be developed.
- **Define their assessment method.** They eventually had to decide assessment methods consistent with their CT learning goals.

At the end of the workshop, groups were invited to present and compare their ideas in plenary sessions and to comment on the choices made by other groups.

3.2 Data collection and analysis

To answer to the research questions described above, data were collected after the two workshops through an online questionnaire developed and adapted in the *Erasmus + Crithinkedu* project. The questionnaire includes both open-ended and multiple-choice questions. We received 22 answers from the Italian group and 26 answers from the American group.

The two questionnaires covered the same areas of interest, even if each one was adapted to the context. Both tools presented closed questions regarding the following topics:

- Personal contacts and information;
- Departments and discipline field (STEM, humanities, social sciences);
- Kind of skill and disposition they were going to develop within their classes;

At the end of both questionnaires the following open-questions were inserted:

- Mention max. 3 activities that you would adopt in your teaching to promote critical thinking. Please also mention why you decided to include those activities in your course.
- In what way do you think the planned activities would influence participants' critical thinking?
- In what way could participants' critical thinking development contribute to achieve other learning objectives?

In order to detect critical thinking levels shown by university teachers on the pilot activity and how they intend to change their pedagogical practices, we analyzed the open questions mentioned above by comparing human assessment with the one carried out by a prototype for the automatic assessment of critical thinking devised by the research group on purpose.

3.3 The automatic tool for critical thinking assessment

Our prototype is composed of four main modules that allow one to perform all the operations necessary to obtain the experimental results.

Authentication manager: The module allows online registration via email and provides a secure login form to access the services offered.

Input module: This module manages the insertion of the questions and answers to be evaluated. For each question, in addition to the title and the text of the question, users are also asked to include words representing the *concepts* and the *successors*. *Concepts* could be defined as the topics that should be covered in a correct and exhaustive answer. *Successors* represent, instead, deepening or related topics of the given concepts. *Concepts* and *successors* will be used by the automatic response analysis module to evaluate the four indicators of critical thinking. It is possible to insert more questions or answers at the same time using the import function from Google forms and uploading the generated XML file. The module interacts with Hibernate, a framework for the automatic management of entities in the local database where all the questions and answers are saved.

Manual evaluator: Through this module, field experts can manually evaluate the indicators for the answers entered. It is possible to select any question on the system and the system will propose in series all the answers not yet evaluated. The user can then decide whether to evaluate or delegate to another teacher. For each question it is possible to associate only one anonymous evaluation; these evaluations will be compared with the automatic evaluations to verify the validity of the proposed approach.

Automatic evaluator: This module is at the heart of the system. It will use two external modules to perform the automatic evaluation of the four indicators presented.

Basic linguistic skills: To evaluate language skills, the system makes use of the collaboration of an external system, JLanguageTool. This tool, developed as an online web service rest, allows you to send texts and receive information on grammar errors in just a few milliseconds. It also allows you to receive a version of the text with the most probable corrections. This correct version is fundamental for more advanced analysis because an incorrect text introduces noise that lowers the performance of the whole system. The value of the indicator is given by normalizing the number of errors considering the number of words that make up the text of the answer.

Relevance: The relevance is assessed by exploiting a Wikipedia analysis: initially, the text of the question and of the answer are sent to an online tagging service through Wikipedia pages, TAGME (https://tagme.d4science.org/tagme/). The service returns a set of pages associated with a given text, in our case the text of the question or answer. To see how many topics related to the question have been described, the system performs the intersection of the titles of all the pages linked to the entities in the outgoing link related to the text of the question with those reported by the TAGME service for answers. The hypothesis that we want to show is that outgoing links from pages representing the concepts of the question points to concepts that must be covered by the answers.

Novelty: The same analysis carried out for the evaluation of the relevance is performed to evaluate these indicators using the set of concepts defined during the creation of the demand for possible inferred developments.

Importance: The importance is evaluated by exploiting an analysis of the concepts defined during the creation of the application. The text is processed by a POS Tagger (https://nlp.stanford.edu/software/tagger.shtml), which analyzes the text of the response and extracts all the nouns. This set of nouns is applied to an algorithm that generates n-grams of length from one to three and is compared with the concepts defined for the question. The number of the intersections between the n-grams and the concepts will give the relevance of the answer to the topic treated. The analysis of concepts through Wikipedia is also applied to the previous indicators.

We decided to take advantage of Wikipedia for these analyses because most of the teachers, about 87 percent, use Wikipedia in their didactic activities and the reliability of Wikipedia (primarily of the English-language edition) has also been assessed: an early study in the Nature journal said that, in 2005, Wikipedia's scientific articles came close to the level of accuracy of the Encyclopedia Britannica (*Giles, 2005*).

For this first evaluation of the prototype, we analyzed only the American group because Wikipedia.it contains only 1 million pages against the 5.5 million of the English version and this leads to a considerable decline in performance in finding the concepts associated with questions and answers. In the future, we hope to extend the approach to every different language.

The first interaction that users have with the system after entering the URL to reach the platform (currently locally on a Roma3 server) is with the login form. If the user reaches the platform for the first time, he/she is asked to perform an email registration, with confirmation from the system administrator.

The submission of the login form redirects the user to the main page of the system. Here the user will find all the questions inserted in the system and for each question he/she can perform a manual evaluation of the answers based on the four criteria: basic linguistic skills, relevance, novelty and importance.

When a user chooses the manual evaluation, the text of the question and the answer will be visualized. Through four checkboxes it will be possible to manually insert the values of each critical thinking indicator.

On the other hand, the system can perform the automatic evaluation of the answers and create the entry in the database for future evaluation. By clicking on the "insert a question button," the user will visualize an insertion form where to write the text of the answer, two sets of concepts that should be treated in the answer and represent possible developments or conclusions.

4 Findings and Discussion of Results

In the case of the American group, most of the teachers are based in the field of humanities and social sciences (see Figure 3).

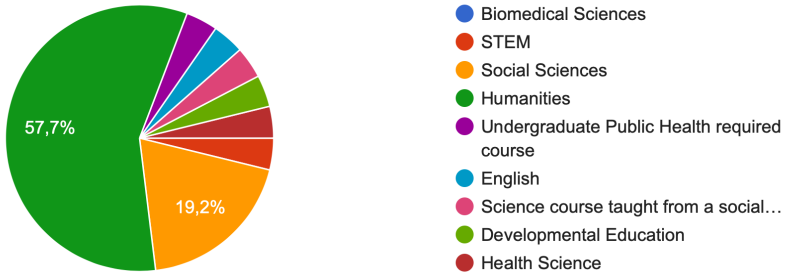

Fig. 1. Teaching discipline sectors of the American group

In the Italian group, most of the teachers come from the department of Educational Sciences (45.5%), Foreigner literature (22.7%), Engineering (9.1%), Economics (9.1%) and Business School (9.1%).

Manual evaluation was carried out by two domain experts and the averages of the values collected were taken into consideration as a reference for comparisons. For each sub-skill, each question collected is marked from a minimum of 1 to a maximum of 5. The two groups, as shown in Figures 5 and 6, obtained similar scores in terms of sub-skills related to critical thinking.

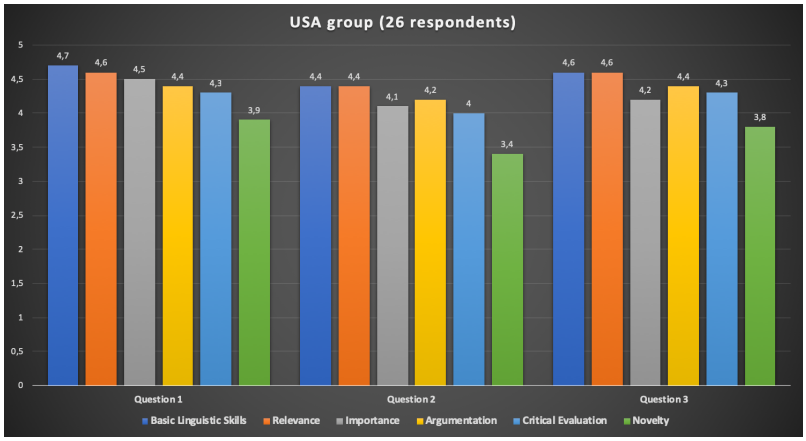

Fig. 2. Scores obtained by the USA group on the six sub-skills of critical thinking from human evaluators

In the first five skills (basic linguistic skills, relevance, importance, argumentation, critical evaluation), both the groups obtained a score higher than four, showing a good level of CT on the five areas. With regard to the last skill, novelty, both showed a similar score with a result lower than four.

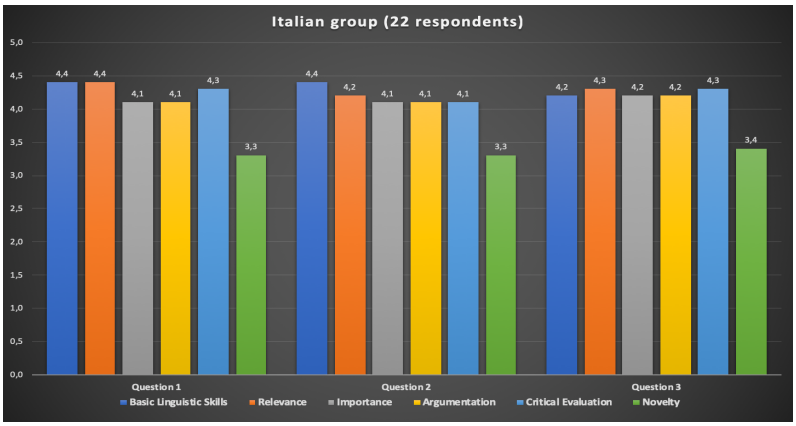

Fig. 3. Scores obtained by the Italian group on the six sub-skills of critical thinking from human evaluators

All in all, the two groups have achieved similar total scores. The maximum critical thinking score possible is 30 coming from the sum of all the sub-skills scores. Both groups have achieved a medium-high grade in the first question. This could be explained by the fact that teachers deeply reflected on the educational activities aimed at developing critical thinking.

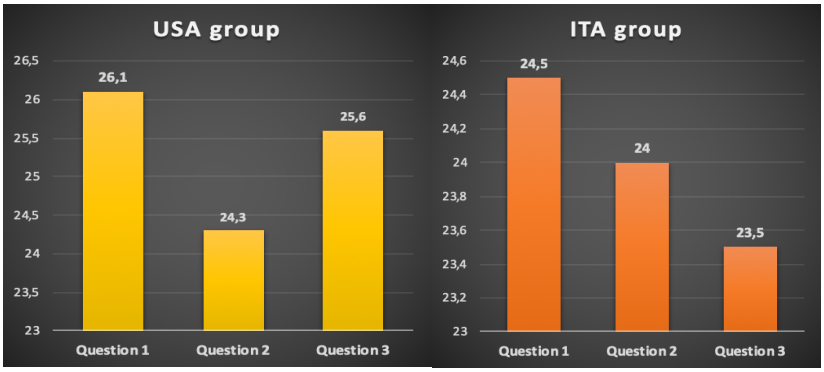

Fig. 4. Total scores obtained by the Italian and the USA groups on critical thinking level from human evaluators

In Figures 8 and 9, the results of the automatic analysis of the same questions assessed by the human experts before are reported.

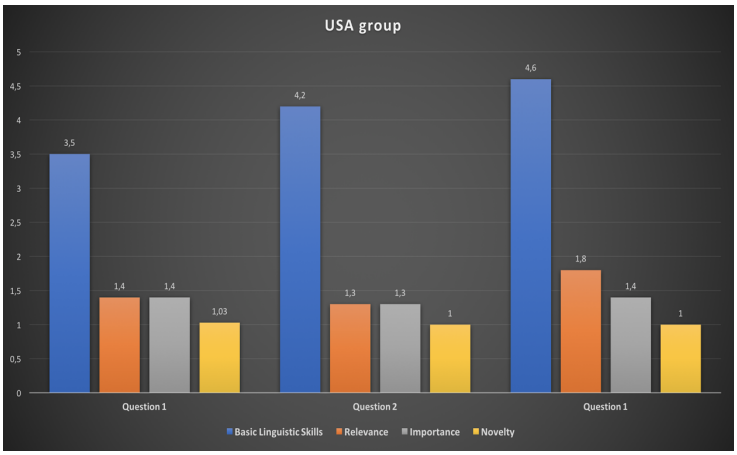

Fig. 5. Scores obtained by the USA group on the four CT sub-skills through automatic evaluation

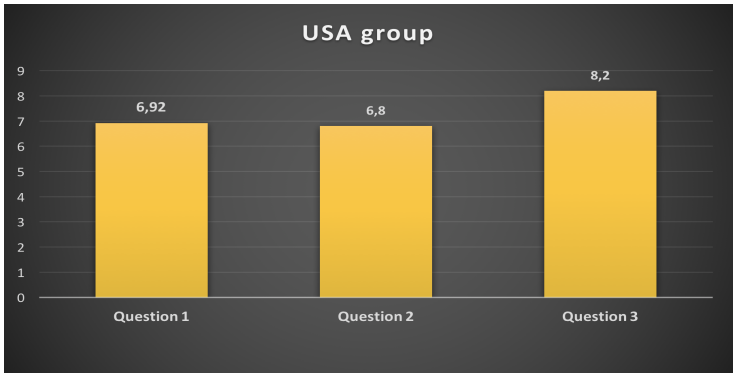

Fig. 6. Total scores obtained by the USA group on critical thinking level through automatic assessment

In the automatic classification shown in Figure 9, only four critical thinking indicators are considered and therefore the maximum score for each question is 20. To analyze the performance of the prototype, the metric used to compare manual and automatic evaluation is accuracy. The *accuracy* is the ration between *number of correct predictions* and *number of total predictions.*

For this first test we made some simplifications:

- The terms inserted in the system to contextualize relevance and novelty are defined as the titles of the Wikipedia pages associated with the concepts in order not to introduce noise in the evaluation process.
- The manual scoring was divided into three classes of values (negative, neutral and positive) for each indicator of critical thinking.

For this first pilot an *ad hoc* dictionary of terms for the recognition of synonyms and related concepts has been built as a simple query expansion module applied to concept in order to maximize the number of retrieved entities for the relevance, importance and novelty indicators.

The prototype, currently, can only evaluate texts related to the domain of the three questions considered. Following the typical approach of the development of the classifiers, we try to identify the best features to describe the problem and then try to generalize these conclusions outside the analyzed domain. In the future we'll try to automatically create these dictionaries through the exploit of bases of knowledge such as DBpedia (https://wiki.dbpedia.org/) or Wordnet (https://wordnet.princeton.edu). In these conditions the prototype agreed with the domain expert in 30% of cases. Analyzing only a sub-sample of the dataset, the one with the best answers (more complete and longer in terms of words), the value grows to almost 34%.

The best evaluations were obtained from the basic linguistic skills and importance indicators with accuracy values of 67% and 39% respectively. The result is not satisfactory yet for an effective classification considering the application of the study to the domain (only three questions) and the restrictions made, but has allowed us to identify many points to extend the approach. An analysis of the negatively classified instance highlighted some evidence: the process of defining the associated concepts to the importance must be very specific, otherwise the system can't evaluate the indicator correctly because general concepts lead the system out of topic in the Wikipedia analysis. Moreover, it has been found that the more general the question is, the more the system performance worsens calculating the relativity of the answer, due to the number of concepts found in the Wikipedia pages explored and that are not related to the question.

Finally, to increase the accuracy of the classification it may be interesting to analyze a semantic database for a better contextualization of the questions and answers considered; specifically, it could be interesting to extract the set of associated concepts and travel the tree of the Wikipedia categories linked to the pages to go back to common nodes to better recognize the level of relevance.

5 Final Remarks

Taking into consideration the starting research questions, for the sake of the present contribution, some final remarks can be made. First of all, data collected here are limited to a pilot activity where a small number of participants was involved (48 in total) so any generalization is impossible. University teachers within the sample used have a fairly correct CT interpretation and knowledge and they show positive results in four out of the five CT macro-indicators. The attempt to automatize CT assessment through open-ended questions is at its start but proves to be a useful support to human evaluation. The use of language analysis procedures seems to be a possible direction according to the first results collected in the study herewith presented. The research group feels therefore encouraged to follow up the research described above, through further experimentation, working also on different macro-indicators from the Newman, Webb and Cochrane adapted model employed so far.

In future studies, we are going to expand the textual corpus because our prototype achieved slightly better performance with longer and more elaborated open-answers. We will conduct further validation studies with a larger sample and with different kinds of questions.

6 Acknowledgment

A. Poce coordinated the research presented in this paper. The research group is composed by the authors of the contribution that was edited in the following order: A. Poce (1. Introduction, 2. The context of the research 5. Final remarks), C. De Medio (3.3 The automatic tool for critical thinking assessment, 4. Findings and discussion), F. Amenduni (3.2 Data collection and data analysis), M. R. Re (3.1 The workshop structure).

7 References

[1] Davies, M., & Barnett, R. (Eds.). (2015). The Palgrave handbook of critical thinking in higher education. Springer. https://doi.org/10.1057/9781137378057

[2] Dominguez, C. (coord.) (2018). *The CRITHINKEDU European course on critical thinking education for university teachers: from conception to delivery*. Vila Real: UTAD. ISBN: 978-989-704-274-4

[3] Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).

[4] Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438, 900–901 https://doi.org/10.1038/438900a

[5] Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. Artificial Intelligence, 171(10-15), 875-896. https://doi.org/10.1016/j.artint.2007.04.010

[6] Jiménez-Aleixandre, M. P., & Puig, B. (2012). Argumentation, evidence evaluation and critical thinking. In Second international handbook of science education (pp. 1001-1015). Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9041-7_66

[7] 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. https://doi.org/10.1002/ets2.12009

[8] Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. Thinking Skills and Creativity, 6(1), 1-13. https://doi.org/10.1016/j.tsc.2010.08.002

[9] 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.

[10] Poce, A., Corcione, L., & Iovine, A. (2012). Content analysis and critical thinking. An assessment study. CADMO.

[11] Poce, A. (2015). Developing critical perspectives on technology in education: A tool for MOOC evaluation. European Journal of Open, Distance and E-learning, 18(1).

[12] Wegerif, R., McLaren, B. M., Chamrada, M., Scheuer, O., Mansour, N., Mikšátko, J., & Williams, M. (2010). Exploring creative thinking in graphically mediated synchronous dialogues. Computers & Education, 54(3), 613-621. https://doi.org/10.1016/j.compedu.2009.10.015

8 Authors

Antonella Poce is Associate Professor in Experimental Pedagogy at the University Roma Tre – Department of Education. Her research concerns innovative teaching practices in higher education at a national and international level. She is a member of the EDEN – European Distance and E-Learning Network (since 2009) and has been elected Chair of NAP SC in 2017 and she has been a EDEN Executive Committee member since then. She is a member of ICOM–CECA (Committee for Education and Cultural Action) (since 2006). She coordinated four departmental projects and the Erasmus+ projects. She chairs the two-year post graduate course, "Advanced Studies in Museum Education".

Carlo De Medio is a Ph.D. candidate in Computer Science at the University of Roma Tre. His research interests are in the field of adaptive learning and critical thinking evaluation tools. Contribution: analysis technological tools to promote social inclusion in museum contexts, production and evaluation of Inclusive Memory OERs and MOOC, transversal skill development assessment, project evaluation.

Francesca Amenduni is a Ph.D. student in Culture, Education and Communication at the University of Roma Tre in collaboration with University of Foggia. Her expertise is in the e-learning field both as practitioner and researcher. She has worked as an e-learning tutor and instructional designer since 2015. She carried research related to blended learning, and her current Ph.D. project regards semi-automated assessment of critical thinking in e-learning forums.

Maria Rosaria Re is a Ph.D. student in Culture, Education and Communication at the University Roma Tre in collaboration with University of Foggia. She used to be temporary researcher in the academic year 2015/2016, Department of Education – Università Roma Tre, carrying out research work in interactive teaching and learning online with specific reference to MOOC (Massive Online Open Courses) employment in museum education. She has been cooperating with Laboratory of Experimental Research and Centre for Museum Studies (University of Roma Tre) since 2013 and took part in national research projects and European projects.

This article is a revised version of a paper presented at the International Conference on E-Learning in the Workplace 2019 (ICELW 2019), held in June 2019, at Columbia University in New York, NY, USA. Article submitted 2019-07-12. Resubmitted 2019-08-17. Final acceptance 2019-08-22. Final version published as submitted by the authors.