Understanding User Behavior in Textual Analysis: A Thinking Aloud Approach for Digital Humanities Research Contexts

Patricia Martin-Rodilla
Institute of Heritage Sciences,
Spanish National Research Council
Avda. Vigo s/n Campus Sur
15705 Santiago de Compostela, Spain
patricia.martin-rodilla@incipit.csic.es

Cesar Gonzalez-Perez
Institute of Heritage Sciences,
Spanish National Research Council
Avda. Vigo s/n Campus Sur
15705 Santiago de Compostela, Spain
cesar.gonzalez-perez@incipit.csic.es

ABSTRACT

The exponential increase in availability of scientific papers, institutional reports or research monographies in digital contexts (i.e. in digital repositories, archives or social scientific networks) has led to the advancement of manual, semi-automatic or automatic-based methods to analyze these texts in the digital environment. These techniques cover a heterogeneous range, from manual expert analysis using computer methods (usually by annotation systems), to the application of natural language processing algorithms or discourse analysis techniques, which are able to identify cognitive relationships between text elements, e.g. causal structures or contrasts argumentations. This advancement is more evident in humanities research contexts, where most of the knowledge generated are expressed in textual formats. However, how the use of these techniques is affecting the analysis conducted by researchers in humanities' texts? Is it possible to measure the quality of the textual analysis? What kind of cognitive structures are identified in the text using these methods?

This paper presents an empirical study conducted with humanities researchers, with the goal of obtaining a better understanding about how texts in digital contexts are analyzed by these professionals using semiautomatic discourse analysis techniques. The paper also proposes a method, based on Thinking Aloud protocols, in order to design experiments and evaluate software cognitive aspects, such as digital textual analysis, with humanities professionals. Finally, the paper discusses about how empirical studies and the Thinking Aloud method constitute a solid basis to better understand the relationship between expert textual analysis in humanities and it conducting using software methods.

Categories and Subject Descriptors

• **General and reference~Empirical studies** • **Applied computing~Arts and humanities** • **Information systems~Language models** • *Computing methodologies~Discourse, dialogue and pragmatics* • **Applied computing~Document analysis.**

Keywords

Experimentation in digital humanities; digital textual analysis; cognitive aspects; information systems.

1. INTRODUCTION

Humanities research involves the analysis, comprehension, production and sharing of a huge amount of textual sources. In the last decades, these sources are generally available in scientific repositories, institutional archives and social networks specially designed for scientists, such as Research-Gate [1] or Academia Edu [2], among others. Thus, humanities professionals are currently using digital-based methods and techniques to analyze texts from digital sources, including annotation systems, or natural language processing techniques, in order to extract information and generate new knowledge based on these textual sources [3].

In this context, previous works [4] show how discourse analysis techniques offer a suitable and agile set of resources to analyze textual sources in humanities and extract cognitive aspects from them. We understand by cognitive process a relation, presented in the textual discourse, between text elements. Some examples are causal relations, contrasts or exemplifications inside texts. These relations are identified by Hobbs [5] as a total of ten types — Occasion, Elaboration, Explanation, Evaluation, Generalization, Exemplification, Contrast, Violated Expectation, Background and Parallelism— (for further information see Hobbs [5]). In previous works we formalized each of them in a formalization model, as shown in [4, 6]. Discourse analysis techniques allow humanities professionals to analyze in a semiautomatic manner textual sources. However, there is a lack of empirical studies about how professionals in humanities use this kind of techniques and methods, and how the textual analysis is influenced by these techniques.

In order to analyze in an empirical way how professionals use discourse analysis in analyzing humanities textual sources, we perform an empirical study based on Thinking Aloud Protocol [7] (hereinafter TAP). The final goal of the study is identifying how cognitive processes are most easily identified by these professionals, if there are coincidences or differences between personal textual analyses and how aspects can be identified in order to improve the textual analysis process by using software tools. In addition, the paper presents our TAP-based method created to design experiments and evaluate software cognitive aspects, such as digital textual analysis, with humanities professionals. This method allows us to offer a design framework for user studies in order to continue working in the better understanding of the user behavior in digital textual analysis.

The paper is structured as follows: Section II describes existing work on software-based techniques and methods for textual analysis commonly used in humanities research contexts. Also, introduces TAP characteristics and the reasons why the decision was made to use in order to validate the textual analysis in humanities contexts. Section III introduces the empirical study design performed and how this design allows us to define an entire TAP-based method for future empirical designs in evaluating cognitive aspects in software with humanities professionals. Section IV presents the empirical study results. Section V presents the discussion and conclusions about strengths and weaknesses of the method proposed, and future possibilities.

2. PREVIOUS WORK

This section presents an overview of existing works on textual analysis via software used in humanities research contexts. Due to our special interest in discourse analysis techniques, we also present discourse analysis techniques and how existing works validate discourse analysis applied to humanities texts. As a result of this review, we present TAP as a basis for designing our empirical study.

2.1 Software Methods for Textual Analysis

Although the purpose of this paper is far from a systematic review of the existing software methods for textual analysis (focusing more in presenting the empirical study and the derived method for designing future empirical validations of the same nature), we introduce here the main categories of existing works in order to offer a research context of the proposal and also a contextualization of discourse analysis techniques, which are the final method evaluated during the empirical study presented.

We classify the existing work attending to the software automatization level required for the techniques or methods from the humanities professionals. Firstly, we can find textual analysis methods based on manual techniques. In these approaches, our humanities' user performed a manual analysis using software tools (with different levels of assistance and usability). Most of these approaches are based on textual editors that allow the human to label different elements of the text (attending semantic, syntactic or functional criteria). Also, these systems allow the systematic annotation [8] by the expert for future software treatment of the text. Secondly, we can find semi-automatic approaches, such as software systems that offers suggestions for annotation and analysis, or discourse analysis techniques [5, 9], which offer and automatic segmentation of the text, and a manual characterization of the cognitive processes presented in the discourse by the humanities expert. This characterization is done following the ten types of cognitive processes identified by Hobbs as we previously detailed. Finally, we can find more automatic approaches, based on the automatic application of natural language processing algorithms to the humanities' textual sources, such as Named Entity Recognition (NER) [3] or similar algorithms offered by programming suites [10, 11].

We focused our empirical study on semi-automatic approaches, and specifically on discourse analysis techniques, due to our final goal is to know more about the user's behavior on digital textual analysis, and how cognitive processes could be identified by them. Automatic approaches do not allow us to measure the user behavior in terms of textual analysis, due to the software tool performed the textual analysis without human collaboration (we could only measure in an empirical way aspects about the

usability of the software tool employed, but not about the textual analysis process itself).

2.2 Thinking Aloud Protocol

In order to performed an empirical study of the discourse analysis techniques applied to textual analysis in humanities research contexts, and due to the nature of the product to be validated (the textual analysis performed using discourse analysis techniques in a semi-automatic software environment), it was necessary to select techniques which would allow us to obtain as much information as possible regarding how the cognitive processes were carried out during this validation work. This necessity led us to consider TAP techniques as a possible basis for the validation of the proposed characterization.

In this section, we shall describe existing works in empirical studies for cognitive aspects in textual analysis, and the reasons why the decision was made to use Thinking Aloud Protocol [7] in order to validate the textual analysis in humanities contexts. Later, we shall go on to describe the design of the validation model created based on TAP, with the aim of validating cognitive processes in digital textual analysis. This design represents a scientific contribution in itself, given that, at the time the review of the studies listed here was carried out, no bibliographic sources were found documenting empirical TAP studies with this aim, although some were found in other contexts of Software Engineering, as we detailed in next paragraphs. The proposed design was applied in order to validate cognitive processes identified in discourse analysis processes when humanities researchers analyze textual sources.

On textual analysis in digital contexts, most of the existing validations follow a Software Engineering approach, validating the use of the software tool or the method for text analysis itself (annotation, automatic algorithms, etc.). There are many methods of very different types for performing validations in Software Engineering, ranging from scenarios or case studies [12], to statistical experimentation [13] or prototyping and concept tests [14]. Within the types of validations with an empirical component, approaches based on observation, documentation and formalization of the resolution of problems or tasks are the most common [13, 15].

The term TAP (Thinking Aloud Protocols) is applied to a set of techniques which originated in the field of experimental cognitive psychology, in which experts in the field being studied are asked to voice the words in their minds or, in other words, to manifest each idea that they think of when carrying out certain selected tasks which they must perform. Although this technique had been used in the past, it was in the 1980s that its use became generalized, especially in connection to studies such as that of Ericsson and Simon [16], who developed the use of these techniques for studying high level cognitive processes, especially those relating to memory, from a scientific and methodologically rigorous perspective. Years later, studies such as those by Olson [17], indicated that the use of TAP is one of the most effective ways to evaluate the highest level of cognitive processes (for example, those involving working memory) and that the technique could also be used to study the cognitive differences between individuals when carrying out the same task. Ericsson and Simon [16] concluded that, although the vision of cognitive processes provided by a TAP session may prove incomplete, the results represent a reliable source as far as the cognitive processes being

carried out are concerned. Due to these ideas, TAP protocols have become common as a technique employed in research contexts studying areas which are traditionally verbal or narrative, such as psychological [17] and linguistic [18, 19] studies. They are also commonly used in problem-solving contexts in different fields, such as Physics and Mathematics or Biomedicine [20, 21]. These two approaches (more textual and/or narrative or more focused on problem-solving) employ their own application methodologies of the protocol.

The studies found which use TAP techniques in humanities fields are worthy of special mention. The majority of them employ the protocol as the base methodology for ethnographic work or for interviews relating to tourism [22] but they do not characterize the cognitive processes which are performed during the sessions. Therefore, their application of the TAP protocol is of no use to us as a basis for our objective. Studies of Humanities students regarding their temporal reasoning [23, 24] or the extraction of specific sub-processes, such as the transcription of ancient texts [25], are the only reference found regarding the application of TAP for an objective similar to our own, showing satisfactory results in both sub-disciplines.

As for how and why TAP protocols have been used in Software Engineering, there are numerous studies which extract cognitive processes during the carrying out of modelling tasks [26], decision-taking in programming [27] and prototyping and validation tasks [28-31]. Furthermore, TAP protocols have been used for tasks more closely related to the identification of cognitive processes during textual analysis, such as in on-line search tasks [32], ontological work [33] and the creation of knowledge [34]. However, there is no defined and agreed upon methodology for the application of TAP in the validation of cognitive aspects in software contexts.

In general, "the literature of think-aloud research shows its strong theoretical foundation and confirms its value as a way of exploring individuals' thought processes" [35]. In addition, it gives satisfactory results in its application in Software Engineering contexts, thus confirming its value as a way of exploring cognitive processes [35]. The absence of a specific methodology for the application of TAP protocols to validate cognitive processes in software-based textual analysis contexts, as well as the need to resort to a hybrid context combining the experience of TAP application in textual and/or narrative contexts and in problem solving, makes it necessary to design a specific method for applying the TAP protocol in order to validate cognitive processes in software-based textual analysis contexts.

3. EMPIRICAL STUDY: DISCOURSE ANALYSIS IN HUMANITIES RESEARCH CONTEXTS

In summary, our goal is evaluating in an empirical way how humanities professionals use discourse analysis techniques (in a software semiautomatic process) to analyze digital textual sources. With this goal in mind, we design a TAP-based method in order to design this kind of empirical studies. Subsequently, we applied the method proposed to perform our empirical study with real humanities researchers.

3.1 . A TAP-based Method for Evaluating Software Cognitive Aspects in Humanities

As other authors have pointed out [13], empirical validations in the field of Software Engineering can prove complex, due to the large number of aspects to be taken into account. In addition, the interdisciplinary nature of the research which concerns us here increases this complexity to a certain degree. Therefore, and in order to guarantee the rigor of all the studies and empirical validations carried out with humanities professionals, a reference framework for experimentation in Software Engineering by Wohlin [13] has been selected which allows us to design with precision all the aspects to be taken into account in the validation and guides us throughout the entire process. According to Wohlin, the process of experimentation in Software Engineering contexts can be defined in five phases:

Scoping: In this phase, the scope of the empirical study is marked out and the problem and general aim of the research in which the study is set is clearly defined, along with the specific objectives of the study. These specific objectives are defined by way of a template with five sections: object(s) of study, purpose, quality focus, perspective and context.

Planning: this phase proceeds to the complete design of the study to be carried out and the possible threats to validity. This phase includes the definition of the initial hypothesis of the experiment, the experimental context and the selection of the variables and participants. This phase condenses the most important part of the process of experimentation as it establishes the steps to be followed in the subsequent stages, laying the foundations for the possible conclusions which can be reached by analyzing this experimentation.

Operation: this corresponds to the carrying out of the empirical study itself, in which the data is gathered according to the previously created design. It consists of three stages: preparation, execution and data validation.

Analysis and Interpretation: in this phase, the data obtained is analyzed in three stages: descriptive statistics analysis, data set reduction and hypothesis testing.

Presentation and Package: in this final phase, the way in which the results are to be presented is decided and the necessary reports are prepared. In this case, this paper corresponds to the presentation of the results of our empirical study carried out.

We proposed a method in order to adapt this Wohlin framework to the application of TAP for validating software cognitive aspects in humanities research contexts. A short analysis of the most important aspects of TAP is included below.

Tasks: One of the most important aspects to be taken into account is that of selecting the tasks which the participants in the TAP sessions are to carry out. Requiring tasks with a high cognitive level could lead to interference in their verbalization [36]. For this reason, many authors recommend the selection of tasks of a simple or intermediate level of difficulty, with a certain verbal load and organized into ascending levels of difficulty [36].

Tasks involving a software assistance system show a high level of heterogeneity as far as their number and level of difficulty are concerned. Therefore, in this point, it is considered necessary to determine at least one specific task for the validation of each cognitive process present in the textual sources which is to be

analyzed. This allows the degree of possibilities to be reduced, along with the confusion associated with them, in the defined tasks and the cognitive processes which are to be validated. In addition, the tasks defined should arise spontaneously from the direct observation of everyday work in the field being validated. That is to say, they should be tasks which the participants carry out (independently of the method and with different tools) as part of their habitual work in the humanities discipline. In our case, they should be tasks relating to the analysis of texts carried out by specialists in humanities. The tasks selected should be of a low level of difficulty.

The Dynamics of the Sessions: There are a large number of approaches relating to the way data is to be extracted during TAP sessions. The majority of them advocate recording the sessions though minimizing the presence of recording equipment and locating it, along with the person conducting the sessions, beside the participant or at a certain distance, not in front of them, in order to minimize the degree of intimidation of the participant [37]. Other authors complement the sessions with questionnaires [36] or include control groups which carry out the same tasks outside of the TAP session [16].

Another important aspect regarding dynamics is the need, or lack thereof, for training the participants and/or giving them prior explanations. Ideally, the participants should not require previous training [35], although on occasions they may be offered a prior orientation session in order to reduce tension levels at the start of the session (the so-called "cold start effect") [38].

As far as the number of participants is concerned, TAP sessions generally involve a reduced number of participants, basically due to two reasons: (1) The objectives of the sessions are normally qualitative, so it is not necessary to carry out a huge number of sessions in order to obtain results and (2) The great and time consuming workload which the design of the sessions implies.

In our case, we opted to record the sessions, attempting to minimize the degree of intimidation to the participant as we believe that they are a valuable source of information for future study in the discipline. Furthermore, some simple instructions were given to the participants in the sessions regarding the discourse analysis techniques, the cognitive processes that they can identified through the textual analysis and the tasks to be carried out, although they were not trained in TAP techniques beforehand. The decision was also made to carry out a reduced number of sessions.

The Analysis of TAP Results and Complementary Data: Finally, it is necessary to take into account aspects related to the analysis of the results when designing TAP sessions. One of the most relevant aspects is the choice of the subject of the tasks to be performed in the sessions as this will, later, affect the results. In this context, studies [35] demonstrate the effectiveness of the choice of a specific case study to be carried out during the TAP sessions as a way of, later, being able to interpret the results from a general and applied point of view.

Furthermore, the subsequent method of analysis and enrichment of the TAP results varies notably. Charters carried out a review in which the majority of the TAP sessions analyzed combine qualitative focuses with a certain quantitative classification of the results. Although there are also purely qualitative studies [18], the presence of quantitative indicators is a constant in the analysis of results [39], obtaining them by way of answers to interview

questions. However, a certain lack of structure should be maintained in some parts of the sessions as this permits valuable data to be obtained which is impossible in an entirely structured format [40].

In our case, we opted for a quantitative and variable-based analysis of the data obtained during the TAP sessions, although the recording of the full session allowed us to create and gather data in less structured conditions during the session itself, which we believe may be of interest for future studies in this area.

Taking into account all of these aspects, a TAP-based model for evaluating cognitive aspects in software contexts, shown in Figure 1, has been designed.
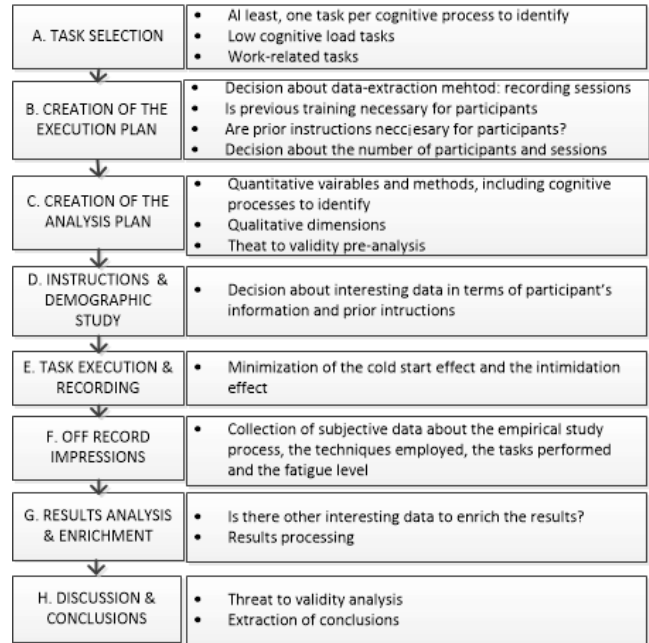

Figure 1. TAP-based method proposed for evaluation cognitive aspects in Digital Humanities Research Contexts.

Table 1 presents the proposed method and how each part of the TAP process is corresponding to specific Wohlin framework phases during the design of empirical experiments.

Table 1. TAP-based method proposed steps and their correspondence inside Wohlin framework.

Wohlin Phase	TAP-based method
Scoping	Previous to apply TAP
Planning	A, B, C
Operation	D, E, F
Analysis & Interpretation	G, H
Presentation & Package	Independent of applying TAP

This design would allow for a method of reference, which would aid researchers in the validation of the cognitive processes they have identified as relevant in their field of application, allowing them then to obtain a better understanding of the user behavior in textual analysis applying software-based methods and tools.

3.2 Empirical Study Design

Following the proposed method, we have designed and performed an empirical study with humanities researchers in order to analyze discourse analysis techniques applying when humanities researchers analyze digital texts. Next, we explain in detail the empirical study design following our proposed method.

3.2.1 Scoping

The objective of this empirical study, in Wohlin's terms, can be defined as:

Analyze the textual analysis performed by humanities professionals for the purpose of evaluation regarding the degree of cognitive processes identified using discourse analysis techniques from the perspective of researchers in humanities fields in the context of public and private research institutions.

3.2.2 Planning the Study

Regarding this Wohlin's phase, we apply phases A, B and C of our TAP-based method (see Table 1).

Context: The empirical study was carried out with real tools and problems (20 fragments of discourse extracted from management and research reports from humanities repositories, which are available on-line). The study is set in a specific but broad context (a sample of specialists in humanities drawn from both public and private institutions on a national scale). It could be considered that the context of this empirical study is on-line as it was carried out in a professional context in which it would be used (public and private management and research institutions). It should be remembered, at this point, that the study was carried out using a methodology based on TAP (Thinking Aloud Protocol), specifically designed for the purpose of this empirical study and described in the previous section.

The formulation of the hypothesis: a hypothesis was verified, which was informally defined as:

"The textual analysis employed discourse analysis does not allow the cognitive processes identification in humanities research contexts. This is reflected in the fact that the percentage of coincidence between the categorizations of the selected fragments of discourse among the different experts in the field is low. Therefore, the discourse analysis employed may not be sufficiently representative, suitable, understandable or generalizable in the field."

In this case, we shall formalize the suitability of the textual analysis performed in the form of cognitive processes identified by humanities' researchers. Thus, our one response variable is: The average percentage of coincidence between the assignations of cognitive processes which the specialists in the field made to specific fragments of discourse. Taking this informal hypothesis as a basis, we can now define H0 as a null hypothesis and H1 as an alternative hypothesis:

H0: The average percentage of coincidence (CAP) among the assignations of cognitive processes to the selected fragments of discourse among different specialists in the field in each group of fragments is less than 50%. H0: CAP < 50%

Alternative hypothesis H1: CAP >= 50%

Selection of variables: The fragments to be evaluated and the associated cognitive processes were selected as the independent variables. In addition, experience with TAP protocols, controlled via a questionnaire prior to participation, was taken into account as an independent variable.

The dependent variable would be the suitability of the textual analysis performed, formalized via the average percentage of coincidences of assignations of cognitive processes performed by the specialists in order to characterize the fragments of textual discourse.

Selection of subjects: The sample of participants corresponds to the Simple random sampling model. A sample of six humanities specialists was selected from three different institutions (both public and private) (Incipit CSIC, Xunta de Galicia (the regional government of Galicia) and a private company specialized in cultural management) chosen randomly from the 20 specialists who showed an interest in collaborating in this process.

Design principles: As far as randomization is concerned, the objects were not assigned randomly to the subjects. That is to say, all the participants evaluated the 20 selected fragments (named from A to T) extracted from four different reports. The participants evaluated the fragments of discourse and categorized them randomly. However, we believe that the order of evaluation is not relevant.

As for the need to block variables, we consider that the decision to evaluate several fragments of texts from different sources blocked, to a certain degree, the impact of the sources and the text itself in the resulting models. The influence of the participants' textual analysis skills using discourse analysis techniques was also taken into account, with an attempt to measure this aspect being made via a prior questionnaire.

Last of all, it would have been desirable to apply the balancing principle but, due to the difficulty in finding a high volume of subjects, they all evaluated the same fragments, therefore, balancing did not occur.

Instrumentation: The choice of the objects can be considered as random, due to the fact that, although some of the texts belonged to the participants' own institutions, others came from international institutional repositories unrelated to their institutions. In the case of the reports from the institutions consulted, none of the participants were the authors of the texts.

As far lines of action in the execution of the study are concerned, the participants were provided with a document containing the instructions for participation, along with a questionnaire and information regarding the characterization of cognitive processes which should be used for their answers. No prior training was necessary for the participants to reply to the questionnaire.

Evaluation of aspects of validity: According to the Wohlin's definition, based on Cook & Campbell [41] it was considered necessary to highlight the following threats to the validity of the study beforehand:

As far as its internal validity is concerned, we consider that the low number of participants in the study could compromise the results if the objective of the study were to establish causal conclusions. However, as was explained at the beginning of the chapter, the empirical study does not have the objective of establishing this type of relations, thus minimizing this threat.

As far as its external validity is concerned, we consider that the probability of the results being repeated in other environments is

high, due to the randomization in the objects used and the fact that other similar studies and experiments have been carried out in humanities contexts [42].

As far as the conclusion validity is concerned, we consider that the random selection of textual reports and of the fragments to be characterized within these reports constitutes a guarantee of representatively of the texts. Furthermore, conducting the sessions in person, with only one participant at a time, minimized the threats regarding the quality of the data obtained.

As far as its construct validity is concerned, we must take into account the fact that the low number of participants did not allow us to make a statistical generalization of the results. Another possible threat could be the suitability of the selected measures as the average percentage may not permit a more in-depth analysis of the data obtained. However, carrying out the studies according to the TAP protocol allows us to obtain more data in the future. The differences in the participants' profiles and professional skills could also have an influence on the categorization of the selected fragments of discourse.

3.2.3 The Execution of the Study —Operation—
Regarding this Wohlin's phase, we apply phases D, E and F of our TAP-based method (see Table 1).

Preparation: The participants in the study did not know the fragments of text which were to be characterized. They were given information about the discourse analysis technique to be used in the textual analysis but not about the hypotheses being tested in the study. By offering to participate as volunteers in this study, they gave their consent to these conditions. The necessary materials were prepared beforehand: an initial questionnaire about their professional profile and previous experience in TAP (None, Low, Average, and Expert), the selected fragments of text and the cognitive processes to be identify by discourse analysis.

Execution: The participants were received one by one in individual sessions of between 45 minutes and one hour in length, thus avoiding problems relating to fatigue in TAP protocols. Each of them was given an explanation of the discourse analysis techniques and the cognitive processes that can be identified in the texts. They also were told that they would be recorded according to the TAP protocol described in Figure 1. Then, they were asked to characterize the fragments of discourse, describing aloud the reasons for their choices and/or doubts.

Validation of the data: No invalid data was detected. This can be attributed to the individuality and supervision of the process of execution of the study.

The results of the empirical study conducted are showed in next section.

4. EMPIRICAL RESULTS
This section shows the results obtained during the empirical study, corresponding to the Analysis and Interpretation phase in Wohlin terms and applying phases G and H of our TAP-based method (see Table 1). All results are detailed in function of the variable defined for the instrumentation of the initial hypothesis planned. According to percentage of coincidence (CAP), Figure 2 and Figure 3 show the participants' answers regarding the characterisation of the 20 fragments of discourse. Figure 2 shows the fragments in which the variability of the answers is only of two different values. In other words, the specialists in humanities

only differed in two possible cognitive processes when characterising the fragments of discourse. Figure 3 shows those fragments with a greater degree of variability in the answers.

As can be seen in both figures, the majority of the specialists coincided in the cognitive process which they associated with the textual fragment in question.
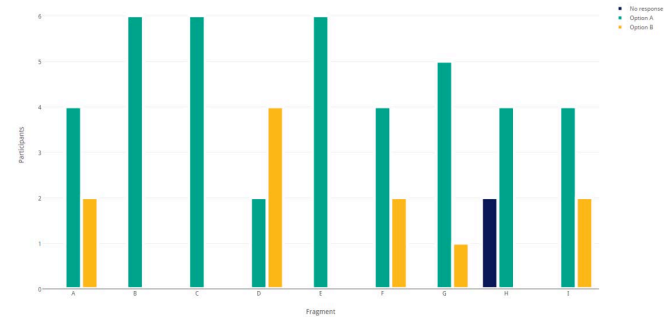

Figure 2. The participants' textual analysis for fragments A-I. In the majority of cases, there is a consensus as far as the cognitive process assigned to the fragment is concerned.
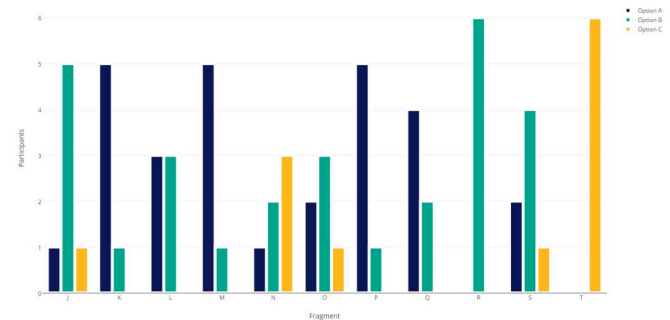

Figure 3. The participants' textual analysis for fragments J-T. In the majority of cases, there is a consensus as far as the cognitive process assigned to the fragment is concerned.

It can be concluded that, although a much higher volume of participants would be needed to obtain an acceptable degree of statistical generalisation, the coincidence percentage CAP in the characterisation of cognitive processes among specialists in the area is acceptable, passing the 50% (0.5) established as a hypothesis (see **Table 2**). Furthermore, this coincidence in all cases of the most selected cognitive process, by the author and by the specialists in the field, enables a line of action to be drawn. Thus, the textual analysis performed by different humanities professionals shows some degree of coincidence when they apply discourse analysis techniques in a software semiautomatic environment.

These coincidences in humanities professional's textual analysis allow us to consider discourse analysis a suitable and agile basis to identify these relations in humanities textual analysis, and to continue working in the application of semi-automatic discourse analysis in humanities.

Table 2. Coincidence Percentage CAP obtained for each of the fragments A-T. The resulting cognitive processes names [5] are abbreviated, i.e. Occasion: OCC, Explanation: EXP, Generalization: GEN.

Textual fragment	A	B	C	D	E
C_{AP}	66,6%	100%	100%	66,6%	100%
Textual analysis result	EXP	EVAL	ELAB	PAR	GEN
Textual fragment	F	G	H	I	J
C_{AP}	66,6%	83,3%	66,6%	66,6%	50%
Textual analysis result	GEN	VIOL	CONT	EXP	PAR
Textual fragment	K	L	M	N	O
C_{AP}	83,3%	50%	83,3%	50%	50%
Textual analysis result	EVAL	EXP	EVAL	PAR	EVAL
Textual fragment	P	Q	R	S	T
C_{AP}	83,3%	66,6%	100%	66,6%	100%
Textual analysis result	OCC	CONT	CONT	EXEM	BACK

5. CONCLUSIONS AND OPEN ISSUES

This paper presents an empirical study about how humanities professionals analyze textual sources thanks to software semiautomatic techniques of discourse analysis. The discourse analysis techniques allow professionals to identify in the text cognitive processes, such as causal structures, exemplifications or contrast argumentations.

This empirical study enabled us to obtain an initial idea of the degree of agreement among the specialists regarding the identification of certain fragments of discourse depending on the underlying cognitive process, as discourse analysis techniques allow. The results of the Thinking Aloud empirical study showed a high degree of agreement in the community, with average percentages of coincidence of around 66%. Thanks to the use of the Thinking Aloud Protocol, we were able to detect other aspects of interest, which, although they have not been formalized in the study, provide important qualitative information with a view to using the discourse analysis techniques by humanities professionals. For example, disagreements may also occur due to the fact that the participants state that cognitive processes reflected in the text are not solidly supported or are expressed

with confusion, although the discourse analysis techniques still allows the underlying cognitive process to be identified.

In addition, the paper presents a TAP-based method to design empirical studies inside Wohlin framework to evaluate cognitive aspects in humanities research contexts.

As future work, it is necessary to carry out a set of empirical studies with humanities professionals using other semiautomatic or automatic software technique to analyze textual fragments, in order to know what techniques present best results of coincidences in textual analysis. Also, we want to continue testing the TAP-based method presented to design this kind of empirical studies, in order to detect problems in application and also possible improvements for the method.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] R. Gate. (2008-2016). *Reseach Gate*. Available: https://www.researchgate.net/

[2] Academia. (2016). *Academia Edu*. Available: https://www.academia.edu/

[3] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 2009, pp. 147-155.

[4] P. Martin-Rodilla and C. Gonzalez-Perez, "An ISO/IEC 24744-derived modelling language for discourse analysis," 2014.

[5] J. R. Hobbs, *On the coherence and structure of discourse*: CSLI, 1985.

[6] P. Martin-Rodilla, "Empirical Approaches to the Analysis of Archaeological Discourse," in *Across Space and Time. Papers from the 41st Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Amsterdam University Press, Amsterdam*, Perth, 25-28 March 2013, 2013.

[7] C. Lewis, *Using the" thinking-aloud" method in cognitive interface design*: IBM TJ Watson Research Center, 1982.

[8] M. Vargas-Vera, E. Motta, J. Domingue, S. B. Shum, and M. Lanzoni, "Knowledge Extraction by Using an Ontology Based Annotation Tool," presented at the Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation, Victoria, B.C., Canada, 2001.

[9] P. Muller and A. Reymonet, "Using inference for evaluating models of temporal discourse," in *Temporal Representation and Reasoning, 2005. TIME 2005. 12th International Symposium on*, 2005, pp. 11-19.

[10] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 423-430.

[11] Nltk.org. (2015). *Natural Language Toolkit — NLTK 3.0 documentation*. Available: http://www.nltk.org

[12]	K. Pohl, *Requirements engineering: fundamentals, principles, and techniques*: Springer Publishing Company, Incorporated, 2010.

[13]	C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*: Springer Science & Business Media, 2012.

[14]	S. E. Sim, S. Easterbrook, and R. C. Holt, "Using benchmarking to advance research: A challenge to software engineering," in *Proceedings of the 25th International Conference on Software Engineering*, 2003, pp. 74-83.

[15]	B. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *Software Engineering, IEEE Transactions on,* vol. 28, pp. 721-734, 2002.

[16]	K. A. Ericsson and H. A. Simon, "Verbal reports as data," *Psychological review,* vol. 87, p. 215, 1980.

[17]	G. M. Olson, S. A. Duffy, and R. L. Mack, "Thinking-out-loud as a method for studying real-time comprehension processes," *New methods in reading comprehension research,* vol. 253, p. 286, 1984.

[18]	L. Cullum, "Encouraging the Reluctant Reader: Using a Think-Aloud Protocol To Discover Strategies for Reading Success," 1998.

[19]	I. Kovacic, "Thinking-aloud protocol-interview-text analysis," *BENJAMINS TRANSLATION LIBRARY,* vol. 37, pp. 97-110, 2000.

[20]	G. Schreiber and B. Wielinga, *KADS: A principled approach to knowledge-based system development* vol. 11: Academic Press, 1993.

[21]	M. W. Van Someren, Y. F. Barnard, and J. A. Sandberg, *The think aloud method: A practical guide to modelling cognitive processes* vol. 2: Academic Press London, 1994.

[22]	T. Clemmensen, M. Hertzum, K. Hornbæk, Q. Shi, and P. Yammiyavar, "Cultural cognition in the thinking-aloud method for usability evaluation," *ICIS 2008 Proceedings,* p. 189, 2008.

[23]	S. J. Foster, J. D. Hoge, and R. H. Rosch, "Thinking aloud about history: Children's and adolescents' responses to historical photographs," *Theory & Research in Social Education,* vol. 27, pp. 179-214, 1999.

[24]	M. B. Wolfe and S. R. Goldman, "Relations between adolescents' text processing and reasoning," *Cognition and Instruction,* vol. 23, pp. 467-502, 2005.

[25]	M. Terras, "Reading the readers: modelling complex humanities processes to build cognitive systems," *Literary and Linguistic Computing,* vol. 20, pp. 41-59, 2005.

[26]	J. Pinggera, S. Zugal, M. Furtner, P. Sachse, M. Martini, and B. Weber, "The Modeling Mind: Behavior Patterns in Process Modeling," in *Enterprise, Business-Process and Information Systems Modeling*, ed: Springer, 2014, pp. 1-16.

[27]	R. Brooks, "Towards a theory of the cognitive processes in computer programming," *International Journal of Man-Machine Studies,* vol. 9, pp. 737-751, 1977.

[28]	A. Holzinger, "Usability engineering methods for software developers," *Communications of the ACM,* vol. 48, pp. 71-74, 2005.

[29]	E. Krahmer and N. Ummelen, "Thinking about thinking aloud: a comparison of two verbal protocols for usability testing," *Professional Communication, IEEE Transactions on,* vol. 47, pp. 105-117, 2004.

[30]	J. Nielsen, *Usability engineering*: Elsevier, 1994.

[31]	L. Van Velsen, T. Van der Geest, and R. Klaassen, "Testing the usability of a personalized system: comparing the use of interviews, questionnaires and thinking-aloud," in *Professional Communication Conference, 2007. IPCC 2007. IEEE International*, 2007, pp. 1-8.

[32]	L. van Waes, "Thinking aloud as a method for testing the usability of Websites: the influence of task variation on the evaluation of hypertext," *Professional Communication, IEEE Transactions on,* vol. 43, pp. 279-291, 2000.

[33]	S. M. Falconer and M.-A. Storey, *A cognitive support framework for ontology mapping*: Springer, 2007.

[34]	P. Wouters, "What is the matter with e-Science?-thinking aloud about informatisation in knowledge creation," in *Pantaneto forum*, 2006.

[35]	E. Charters, "The use of think-aloud methods in qualitative research an introduction to think-aloud methods," *Brock Education Journal,* vol. 12, 2003.

[36]	A. Akyel and S. Kamisli, "Composing in First and Second Languages: Possible Effects of EFL Writing Instruction," 1996.

[37]	D. Nunan, *Research methods in language learning*: Cambridge University Press, 1992.

[38]	B. Gibson, "Talking the Test: Using Verbal Report Data in Looking at the Processing of Cloze Tasks," *Edinburgh Working Papers In Applied Linguistics,* vol. 8, pp. 54-62, 1997.

[39]	M. Pressley and P. Afflerbach, *Verbal protocols of reading: The nature of constructively responsive reading*: Routledge, 1995.

[40]	A. Fontana and J. H. Frey, "The interview: From structured questions to negotiated text," *Handbook of qualitative research,* vol. 2, pp. 645-672, 2000.

[41]	T. D. Cook, D. T. Campbell, and A. Day, *Quasi-experimentation: Design & analysis issues for field settings* vol. 351: Houghton Mifflin Boston, 1979.

[42]	P. Martin-Rodilla, "Knowledge-assisted Visualization in the Cultural Heritage Domain-Case Studies, Needs and Reflections," in *GRAPP/IVAPP*, 2013, pp. 546-549.