A MODELLING LANGUAGE FOR DISCOURSE ANALYSIS IN HUMANITIES: DEFINITION, DESIGN, VALIDATION AND FIRST EXPERIENCES

UN LENGUAJE DE MODELADO PARA EL ANÁLISIS DEL DISCURSO EN HUMANIDADES: DEFINICIÓN, DISEÑO, VALIDACIÓN Y PRIMERAS EXPERIENCIAS

Patricia Martín Rodilla
Instituto de Ciencias del Patrimonio. Consejo Superior de Investigaciones Científicas
patricia.martin-rodilla@incipit.csic.es

César González-Pérez
Instituto de las Ciencias del Patrimonio. Consejo Superior de Investigaciones Científicas
cesar.gonzalez-perez@incipit.csic.es

Abstract

Due to the fact that Humanities generally produce knowledge in textual formats (e.g. narrative conclusions or reports), a properly management needs methods for conceptualizing and extracting information from textual sources. Discourse analysis techniques allow extracting information in terms of the connection between discourse structure and elements of the reality referred in the text, as well as the inferential dimension. This semantic information is not available following other extraction methods from texts. In order to formalize the discourse analysis application for textual sources in Humanities, a modelling language has been defined and initially validated with Humanities specialists, showing the discourse structure and the semantic and inferential aspects extracted.

Keywords: Discourse Analysis. Modelling Language. Text Analysis. Information Modelling. Information Extraction.

Resumen

Las disciplinas humanísticas a menudo generan conocimiento en formatos textuales, como narrativas, informes, monografías, etc., por lo que una adecuada gestión del corpus en Humanidades necesita métodos de conceptualización y extracción de información desde fuentes textuales. Las técnicas de análisis del discurso permiten, frente a otros métodos estudiados, extraer información acerca de la conexión entre estructuras del discurso y las entidades de la realidad a las que refiere el texto, además de la dimensión inferencial subyacente. Con el objetivo de formalizar la aplicación de análisis del discurso en fuentes textuales en Humanidades, se ha definido y validado inicialmente con especialistas en Humanidades un lenguaje de modelado que permite capturar la estructura del discurso y extraer aspectos semánticos e inferenciales presentes en el texto.

Palabras clave: Análisis del discurso. Lenguaje de modelado. Análisis textual. Modelado de información. Extracción de información.

1. INTRODUCTION AND BACKGROUND

Far from other disciplines in which products are produced in more structured formats (datasets, analytical results, etc.), Humanities generally produce products in textual formats, such as narrative conclusions, reports, memories, etc. This situation usually involves the fact that Humanities present more needs in terms of method, techniques and tools that allow the conceptualization and extraction of information from its *corpus*, in order to manage them correctly.

Software engineering approaches in conceptualizing, capturing and extracting information from textual sources go back to decades. Firstly, there were approaches focused on providing techniques and tools to apply semi-automatic and automatic methods to extract information from texts. Most of them were related to information retrieval (Baeza-Yates and Ribeiro-Neto, 1999), and they were based on heuristic and probabilistic techniques that allowed extracting information in a quantitative level. For instance, these techniques extracted frequency results about the presence of specific elements in the text or similar indicators. We can find also more semantic approaches inside information retrieval disciplines, analysing textual sources based on topic maps (ISO/IEC, 2006) or thesauri solutions. Also, noteworthy

are the works in lexematization techniques (Torres-Moreno, 2010: 38-53), structure identification of discursive analysis or sentimental analysis —the identification of positive or negative connotations in a text— (Pang *et al*., 2002: 79-86; Borth *et al*., 2013). These approaches allow extracting semantic relationships between elements, such as hierarchical relationships. However, due to the degree of automation applied, it is not possible to achieve a satisfactory level of semantic extraction for the application to more narrative contexts.

Secondly, there are existing approaches focused on modelling a specific domain, in order to achieve the desirable semantic conceptualization and extraction from textual sources. For instance, existing applications in biomedicine (Jensen *et al*., 2006: 119-129) combine conceptual modelling techniques, annotation and natural language processing methods. These approaches usually present good results in the context of a particular domain, using case studies or well-defined *corpus* in the context of a particular project and designing *ad hoc* information extraction methods. However, an ad hoc design involves a domain-dependency of the solution created. Thus, it is not possible to achieve a high degree of generalization for the proposed solutions.

In this context, more linguistic and semantic approaches have recently become popular; because they allow enriching the information extraction methods from textual sources with an acceptable degree of domain independence. In particular, discourse analysis (Hobbs, 1985) is a set of techniques from Linguistics used to discover semantic relations between elements in the texts based on the discursive structure of them. In other words, applying discourse analysis we can identify what discourse elements are present in a text (sentences, clauses…) and link them to the entities of the reality referred (about what entities is talking about a specific text). In addition, we can identify what inferential relations are connecting those two parts (causal relations, exemplifications, etc.). Discourse analysis techniques incorporate a validation phase with the author of the text analyzed, to keep the original semantic content. What does discourse analysis techniques offers us in order to extract information from humanities texts over other approaches? The connection between discourse structure and elements of reality referred in the text, as well as the inferences made by the author, constitute semantic information which is not available following other extracting methods from textual sources.

For this reason, current studies (Polanyi, 1988: 601-638; Mc Kevitt *et al*., 1999: 947-989) are working on the application of discourse analysis techniques for extracting information from textual sources. Hence, we based our work on the approach made by Hobbs (1985) and subsequent work based on it, and we defined a modelling language that allows applying discourse analysis for extracting information from textual sources in Humanities. This language was previously presented at (Martín-Rodilla and González-Pérez, 2014). In the next section,

we aim to introduce the modelling language. In later sections, we present for the first time the language modelling validation context and the results obtained.

2. THE MODELLING LANGUAGE

To provide to Humanities researchers the necessary method to extract structural information from texts —not only in a quantitative or structural level, but also in a highly semantic and inferential level— based on successful experiences on teaching conceptual modelling to Humanities specialists, we carried out a two-year research about the application of discourse analysis to textual sources in Humanities. We created a conceptual language that allows creating models from textual sources (Martín-Rodilla and González-Pérez, 2014), capturing the discourse structure and extracting semantic and inferential information from them. Using this language, Humanities specialists can model a textual *corpus* in terms of its elements (sentences, clauses) and link these elements to the entities they represent through coherence relations. The discourse analysis process used follows the four-step approach made by Hobbs (1985), adapted for a conceptual modelling tool —the language defined— as we have explained in previous works (Martín-Rodilla and González-Pérez, 2014).

Let's take a few examples of discourse fragments in Humanities to show this. All of them are extracted from an archaeological and historical study in Cyprus (Le Brun and Peltenburg, 2004: 194-196). The publication is available online: "In the mid-1970s, there were only farm tracks leading to this area, but even then, some plots lay uncultivated" (Peltenburg, 2009).

The discourse fragment is divided into two sentences (S1 and S2), following Hobbs discourse analysis techniques: "In the mid-1970s, there were only farm tracks leading to this area and but even then, some plots lay uncultivated" (Peltenburg, 2009).

The modelling language designed allows us to create a class model, identifying the entities that the discourse is talking about (e.g. *farming area*, *plot*), the features of these entities (e.g. the state of the plots is uncultivated) and the relationships between them (the plots only can be accessed by those farm tracks). In addition, the language allows modelling the *contrast coherence relation* —in Hobbs' terms— existing between the two sentences, extracting information about what reasoning processes are performed by author of the text. Thus, consequently, we have all the information about entities, features, relationships and discourse relations present in the model. With the identification of this contrast relation, we can continue modelling the discourse. We can now get to know, for example, if this contrast leads to raise other activities in the area that were not agricultural activity. The following figure shows the model built for the discourse fragment selected:

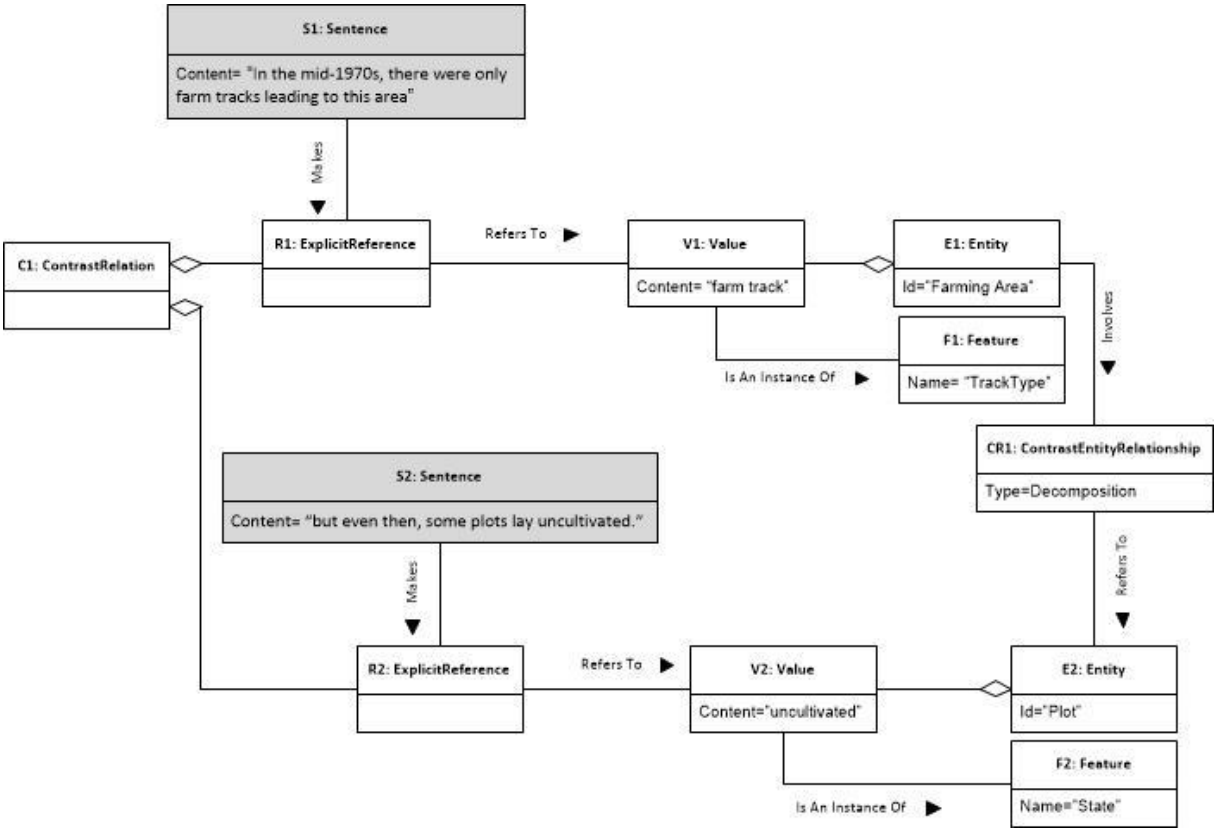

Figure 1. Object model corresponding to the first discourse fragment analysed.

Another example from the archaeological and historical study selected, which presents more complexity in structural and semantic terms, is detailed below: “few examples survive with both rim and base intact. Instead, the bases most commonly survive possibly because rims and sides tended to be finer and taller than those of other types and hence more fragile” (Le Brun and Peltenburg, 2004: 194-196).

The discourse fragment is divided into two parts, separated by the link *Instead*, following Hobbs discourse analysis techniques. The first sentence is already an atomic element. The second part needs to be further divided into smaller sentences: “Instead, the bases most commonly survive and possibly because rims and sides tended to be finer and taller than those of other types and hence more fragile”.

The modelling language designed allows us to create an object model, identifying the entities that the discourse is talking about (parts of material evidences), their features and relationships. In addition, the language allows modelling the *explanation coherence relation* —in Hobbs’ terms— existing between these two last sentences —the second one explains the cause for the first one—. The modelling process continues with the creation of the complete model, involving the first atomic sentence *Few examples survive with both rim and base intact*. We can identify at this point the *contrast relation* —in Hobbs’ terms— between this first atomic sentence and the sentence S2a in Figure 2:

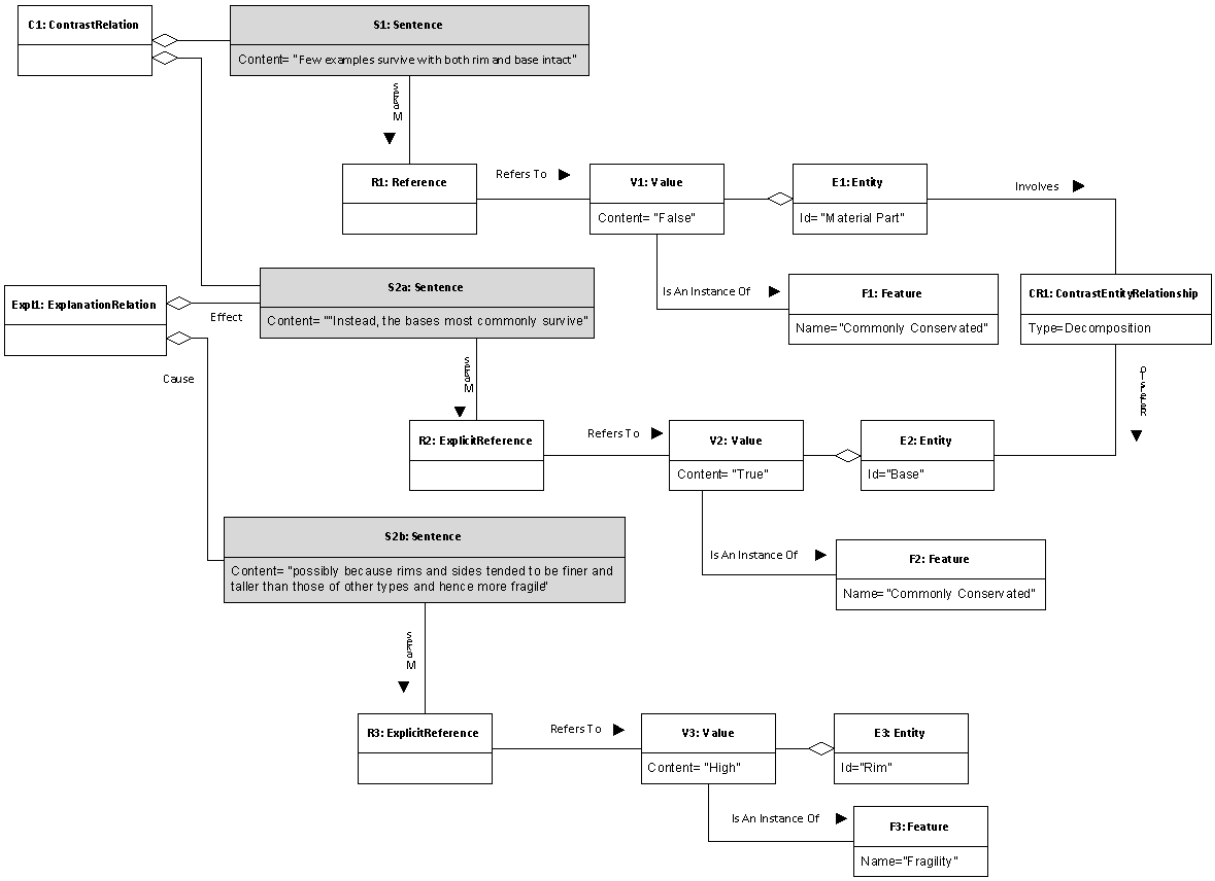

Figure 2. Object model corresponding to the second discourse fragment analysed.

The modelling language allows us to have inside our model all the information about entities, features, relationships and coherence relations present in the discourse. An example of use could be to improve the software searching methods about material evidences, so we can find out which parts are kept and which are not, and relate them to their height and fragility.

Those examples illustrate how our modelling language allows the specialists in Humanities to extract enriched information from textual sources and how this extracted information could be used: in the first example, it is used for continuing the knowledge generation process, in the second example, it is used for extracting desirable searching functionalities in databases or related information systems applications.

3. LANGUAGE VALIDATION AND RESULTS

Using the proposed language, we have modelled a selected *corpus* of texts from historical and archaeological contexts, to validate the approach.

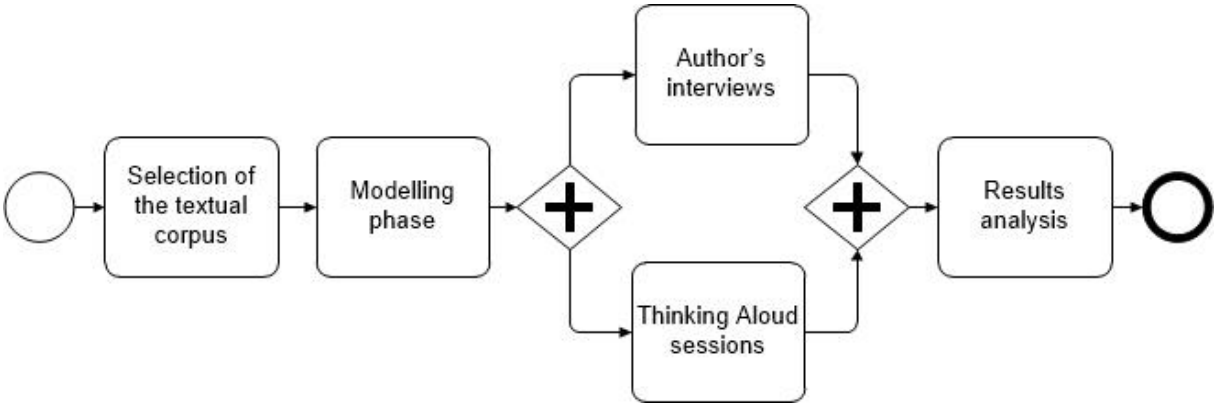

Figure 3. Business Process Diagram showing the validation process performed.

The validation process is shown in previous figure, expressed in BPMN (Business Process Modelling Notation) notation (White, 2004). The circles establish the initial and final events, squares correspond to task performed, and the plus symbol indicates a parallel gateway (tasks performed in parallel). As we can see, the validation process included an interview phase with the authors of the texts —following discourse analysis recommendations— and a group of sessions based on Think Aloud protocols —TAP— (Someren *et al*., 1994). TAP establish recorded sessions with real users —in our case, Humanities specialists— that can *think aloud* during the creation and/or validation of the models, identifying the cognitive processes that they perform in function of the tasks presented in each session. The purpose of the interviews and sessions TAP was to investigate:

- Differences and similarities between models created by the author of the text and models created by software engineers, regarding the same text source.
- Differences and similarities between models created by the author of the texts and models created by specialist in Humanities (belonging to the same domain or colleagues of the author), regarding the same text source.

What areas of conflict exist in both cases? The validation process has allowed us to extract inferential information from texts, such as detection of contrasts, causality or generalization and exemplifications relations. That identification would not be possible using existing approaches from software engineering. That identification would not be possible using existing approaches from software engineering. In addition, the validation results allowed us to figure out about the generalization possibilities of the models created, as well as what inferences presented a higher level of disagreement. All this information is shown in Figure 4 and Figure 5:

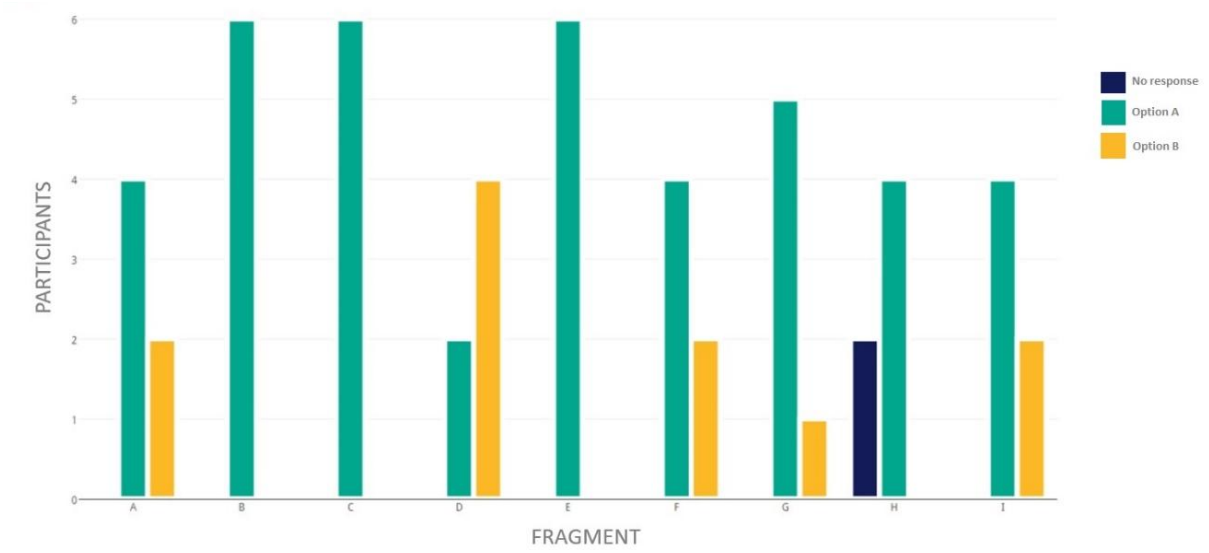

Figure 4. Bar diagram showing modelled responses from six Humanities specialists in A-I fragments analysis.

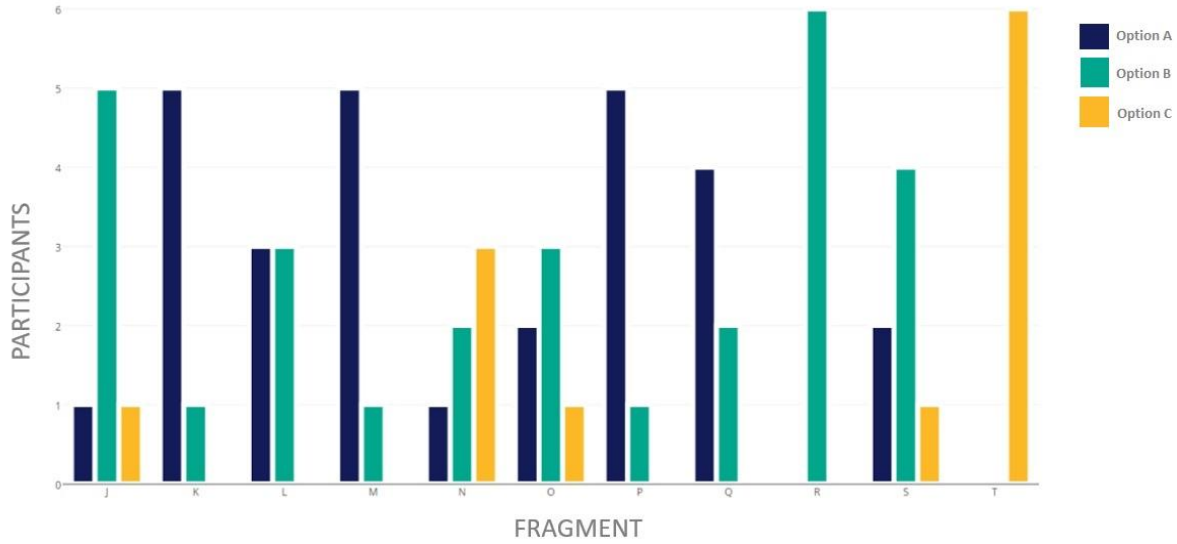

Figure 5. Bar diagram showing modelled responses from six Humanities specialists in J-T fragments analysis.

Both previous figures present results of agreements in models for 20 discourse fragments (fragments labelled from A to T) modelled by the author of the texts and models created by specialists in Humanities for the same discourse fragment. Each discourse fragment was modelled by 6 participants (including the author of the text). The degree of agreement, in terms of inferential information extracted from the textual sources, is higher than the 66% in all cases, except a lower degree of agreement in generalization relations. In these cases, we found a high degree of disagreement among the models made by the author of the text and the models made by researchers in the same discipline. Depending on whether the researcher considered particularly relevant a given example as a basis to generalize, he/she agreed or not.

Regarding the differences and similarities between models created by the author of the text and models created by software engineers, the following table shows some promising results for the same 20 discourse fragments (fragments labelled from A to T):

Fragment	A	B	C	D	E
C_{AP}	66,6%	100%	100%	66,6%	100%
Main Cognitive Process	Explanation	Evaluation	Elaboration	Parallel	Generalization
Cognitive Process (pre-choice)	Explanation	Evaluation	Elaboration	Parallel	Generalization
Fragment	**F**	**G**	**H**	**I**	**J**
C_{AP}	66,6%	83,3%	66,6%	66,6%	50%
Main Cognitive Process	Generalization	Violated	Contrast	Explanation	Parallel
Cognitive Process (pre-choice)	Generalization	Violated	Contrast	Explanation	Parallel
Fragment	**K**	**L**	**M**	**N**	**O**
C_{AP}	83,3%	50%	83,%3	50%	50%
Main Cognitive Process	Evaluation	Explanation	Evaluation	Parallel	Evaluation
Cognitive Process (pre-choice)	Evaluation	Explanation	Evaluation	Parallel	Evaluation
Fragment	**P**	**Q**	**R**	**S**	**T**
C_{AP}	83,3%	66,6%	100%	66,6%	100%
Main Cognitive Process	Occassion	Contrast	Contrast	Exemplification	Background
Cognitive Process (pre-choice)	Occassion	Contrast	Contrast	Exemplification	Background

Table 1. Percentages of coincidence between the cognitive process modelled by the software engineering and the cognitive process modelled by Humanities specialists.

The cognitive process (pre-choice) shown is the coherence relation modelled by the software engineer. The main cognitive process is the coherence relation modelled by the author of the text and the Humanities specialists. For each discourse fragment modelled, the table offers a CAP —Cognitive Agreement Percentage— value, that is, the percentage of agreement between the coherence relation chosen by the software engineering and the coherence relation modelled by Humanities experts (including the author of the text), for each discourse fragment. As we can see in the previous table, in most cases the percentages of agreement are higher than 66,6%.

4. CONCLUSIONS AND FUTURE WORK

The modelling language we created allowed us to extract information from Humanities texts, not only from the entities that are referenced in a text or the discourse structure, but also from inferences and underlying argumentation and how they are used and their connections

with the text elements. Furthermore, we presented here empirical results about the degree of agreement and possible generalization within the community related to a specific *corpus*.

These results can be used in subsequent steps in many ways. Particularly relevant for future work are: (1) the analysis of new *corpus* that will allow us to implement mechanisms to detect inconsistencies and other functionalities presented above and will encourage self-reflection inside the disciplines of the analysed *corpus*, getting to know more about how knowledge is generated using narrative formats in humanistic disciplines. This information about the knowledge generation process is crucial in the development of software systems for the Humanities; (2) the application of the extracted information to the expansion and improvement of existing annotation systems, including inferential information, will enrich the *corpus* analysis; and (3) the detection of relations between entities and underlying inferences as an initial step towards the study of the potential of knowledge discovery and data mining in Humanities texts. For example, the detection of hidden causalities in texts can open the application to existing methods of semi-automatic data-mining based on the causal mechanism, such as the association rules (Agrawal *et al*., 1993: 207-219). This connection has been already pointed out in previous studies (Martín-Rodilla, 2013), although it is still at an early stage of development.

BIBLIOGRAPHICAL REFERENCES

AGRAWAL, R., IMIELIŃSKI, T. and SWAMI, A. (1993). "Mining Association Rules between Sets of Items in Large Databases *Proceedings of the ACM SIGMOD Conference* (*Washington, May 1993*), 22.2, 207-216.

BAEZA-YATES, R. and RIBEIRO-NETO, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

BORTH, D., JI, R., CHEN, T., BREUEL, T. and CHANG, S.-F. (2013). "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs". *Proceedings of the 21st ACM International Conference on Multimedia*. Retrieved from http://www.ee.columbia.edu/ln/dvmm/vso/download/visual_sentiment_ontology_FINAL.pdf on 03/20/2017.

HOBBS, J.R. (1985). *On the Coherence and Structure of Discourse*. Stanford: Stanford University, Center for the Study of Language and Information.

ISO/IEC (2006). *Topics Maps. ISO/IEC* 13250/2006.

JENSEN, L.J., SARIC, J. and BORK, P. (2006). "Literature Mining for the Biologist: from Information Retrieval to Biological Discovery". *Nature Reviews Genetics*, 7.2, 119-129.

LE BRUN, A. and PELTENBURG, E. (2004). "The Colonization and Settlement of Cyprus. Investigations at Kissonerga-Mylouthkia, 1976-1996". *Paléorient*, 30.1, 194-196. Retrieved from http://www.jstor.org/41496692 on 03/20/2017.

MARTÍN-RODILLA, P. (2013). "Software-Assisted Knowledge Generation in the Archaeological Domain: A Conceptual Framework". *Proceedings of the Doctoral Consortium of the 25th International Conference on Advanced Information Systems Engineering* (*CAiSE*) (*Valencia, June 21, 2013*). Retrieved from http://ceur-ws.org/Vol-1001/paper8.pdf on 03/20/2017.

MARTÍN-RODILLA, P. and GONZÁLEZ-PÉREZ, C. (2014). *An ISO/IEC 24744-Derived Modelling Language for Discourse Analysis*. Oral presentation at: *Research Challenges in Information Science* (*RCIS*)*, 2014 IEEE Eighth International Conference*.

Mc KEVITT, P., PARTRIDGE, D. and WILKS, Y. (1999). "Why Machines Should Analyse Intention in Natural Language Dialogue". *International Journal of Human-Computer Studies*, 51.5, 947-989. Retrieved from https://pdfs.semanticscholar.org/347b/c8f647d846f2e94104d55f2b6b00e59759e1.pdf on 03/20/2017.

PANG, B., LEE, L. and VAITHYANATHAN, S. (2002). "Thumbs up?: Sentiment Classification Using Machine Learning Techniques". *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), *10* (*Philadelphia, July 2002*), 79-86. Association for Computational Linguistics. Retrieved from https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf on 03/20/2017.

PELTENBURG, E. (2009). *Kissonerga-Mylouthkia, Cyprus 1976-1996*. Retrieved from http://archaeologydataservice.ac.uk/archives/view/mylouthkia_ba_2009/ on 27/04/2017.

POLANYI, L. (1988). "A Formal Model of the Structure of Discourse." *Journal of Pragmatics*, 12.5, 601-638.

SOMEREN, M.V., BARNARD, Y.F. and SANDBERG, J.A. (1994). *The Think Aloud Method: A Practical Approach to Modelling Cognitive Processes.* San Diego: San Diego State University, Department of Educational Technology, Academic Press.

TORRES-MORENO, J.-M. (2010). "Reagrupamiento en familias y lexematización automática independientes del idioma." *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*,14.47, 38-53.

WHITE, S.A. (2004). *Introduction to* BPMN. Retrieved from http://www.omg.org/news/meetings/workshops/soa-bpm-mda-2006/00-T4_White.pdf on 27/04/2017.