10.12871/CLICIT2014112

Towards a Decision Support System for Text Interpretation

Alessia Bellusci, Andrea Bellandi, Giulia Benotto,
Amedeo Cappelli, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. This article illustrates the first steps towards the implementation of a Decision Support System aimed to recreate a research environment for scholars and provide them with computational tools to assist in the processing and interpretation of texts. While outlining the general characteristics of the system, the paper presents a minimal set of user requirements and provides a possible use case on Dante's Inferno.

Italiano. *Questo articolo illustra i primi passi verso la realizzazione di un Sistema di Supporto alle Decisioni volto a ricreare un ambiente di ricerca per gli studiosi e assisterli, anche mediante strumenti computazionali, nell'elaborazione e nell'interpretazione di testi. Oltre a delineare le caratteristiche generali del sistema, l'articolo presenta una serie minima di requisiti utente e fornisce un possibile caso d'uso sull'Inferno di Dante.*

1 Introduction

A text represents a multifaceted object, resulting from the intersection of different expressive layers (graphemic, phonetic, syntactic, lexico-semantic, ontological, etc.). A text is always created by a writer with a specific attempt to outline a certain subject in a particular way. Even when it is not a literary creation, a given text follows its writer's specific intention and is written in a distinct form. The text creator's intention is not always self-evident and, even when it is, a written piece might convey very different meanings proportionally to the various readers analysing it. Texts can be seen, in fact, as communication media between writers and readers. Regardless of the epistemological theory about where meaning emerges in the reader-text relationship (Objectivism, Constructivism, Subjectivism), a text needs a reader as much as a writer to be expressive (Chandler, 1995). The reader goes beyond the explicit information given in the text, by making certain inferences and evaluations, according to his/her background, experience, knowledge and purpose. Therefore, interpretation depends on both the nature of the given text and the reader/interpreter; it can be understood as the goal, the process and the outcome of the analytic activity conducted by a certain reader on a given text under specific circumstances. Interpretation corresponds to the different – virtually infinite – mental frameworks and cognitive mechanisms activated in a certain reader/interpreter when examining a given text. The nature of the interpretation of a given text can be philological, historical, psychological, etc.; a psychological interpretation can be Freudian, Jungian, etc... Furthermore, the different categories of literary criticism and the various interpretative approaches might be very much blurred and intertwined, i.e. an historical interpretation might involve philological, anthropological, political and religious analyses.

While scholars are generally aware of their mental process of selection and categorization when reading/interpreting a text and, thus, can re-adjust their interpretative approach while they operate, an automatic system has often proved unfit for qualitative analysis due to the complexity of text meaning and text interpretation (Harnad, 1990). Nevertheless, a few semi-automatic systems for qualitative interpretation have been proposed in the last decades. The most outstanding of them is ATLAS.ti, a commercial system for qualitative analysis of unstructured data, which has been applied in the early nineties to text interpretation (Muhr, 1991). ATLAS.ti, however, appears too general to respond to the articulated needs

of a scholar studying a text, lacking of advanced text analysis tools and automatic knowledge extraction features. The University of Southampton and Birkbeck University are currently working on a commercial project, SAMTLA[1], aimed to create a language-agnostic research environment for studying textual corpora with the aid of computational technologies. In the past, concerning the interpretation of literary texts, the introduction of text annotation approaches and the adoption of high-level markup languages allowed to go beyond the typical use of concordances (DeVuyst, 1990; Sutherland, 1990; Sperberg-Mc Queen and Burnard, 1994). In this context, several works have been proposed for the study of Dante's Commedia. One of the first works involved the definition of a meta representation of the text of the Inferno and the construction of an ontology formalizing a portion of Dante's Commedia's world (Cappelli et al., 2002). Data mining procedures able to conceptually query the aforementioned resources have also been implemented (Baglioni et al., 2004). Among the other works on Dante we cite The World of Dante (Parker, 2001), Digital Dante of the Columbia University (LeLoup and Ponterio, 2006) and the Princeton Dante Project (Hollander, 2013). A "multidimensional" social network of characters, places and events of Dante's Inferno have been constructed to make evident the innermost structure of the text (Cappelli et al., 2011) by leveraging on the expressive power of graph representations of data (Newman, 2003; Newman et al., 2006; Easley and Kleinberg, 2010; Meirelles, 2013). A touch table approach to Dante's Inferno, based on the same social network representation, has been also implemented (Bordin et al., 2013). More recently, a semantic network of Dante's works has been developed alongside a RDF representation of the knowledge embedded in them (Tavoni et al., 2014). Other works involving text interpretation and graph representations have been carried out on other literary texts, such as Alice in Wonderland (Agarwal et al., 2012) and Promessi Sposi (Bolioli et al., 2013).

As discussed by semiologists, linguists and literary scholars (Eco, 1979; Todorov, 1973; Segre, 1985; Roque, 2012) the interpretation of a text may require a complex structuring and interrelation of the information belonging to its different expressive layers.

[1] http://samtla.dcs.bbk.ac.uk/

The Decision Support System (DSS) we here introduce aims to assist scholars in their research projects, by providing them with semi-automatic tools specifically developed to support the interpretation of texts at different and combined layers. We chose to start from the analysis of literary texts to be able to face the most challenging aspects related to text interpretation. This work is the third of a series describing the progressive development of the general approach: for the others refer to (Bellandi et al., 2013; Bellandi et al., 2014). In what follows, we describe the general characteristics of the DSS we plan to develop accompanied by a minimal set of user requirements (2.), we present a possible scenario, in which the system can be applied (3.), and we provide some conclusive notes (4.).

2 Towards a Decision Support System for Text Interpretation

In this section, we present our vision of a DSS (Shim et al., 2002) specifically aimed to recreate a *research environment* for scholars and provide them with computational tools developed to assist data elaboration and content interpretation of texts. Theoretically, each automatic act operated by a computational system on a given text can be seen as an interpretative act. Yet, in our view, users shall remain the main decision-makers within their interpretative process, while the system and the integrated tools we aim to create shall function only as instruments enabling users to achieve their research goals in a clearer and easier manner. In the computational metaphor, our DSS would represent the writing desk and library of the historian or the laboratory and microscope of the biologist.

Within the system, users shall be able to carry out a *research project* based on one or more *textual sources* from the beginning through its end, whether the project is the analysis of medical records, the interpretation of a literary work, the production of a critical edition of a given text, or the historical analysis of textual material. Similarly, our system shall assist the creation of text interpretations either for personal purposes (student exercise, amateur research) or for scientific productions (article, monograph, critical edition). Although conceived for the use of a single scholar, the system shall enable users also to selectively share their results in a collaborative space. With the aid of our DSS, users shall be able to consult, search and analyze

a text dynamically and according to their specific interest. The system shall enable to conduct the study of a given text on several and different *layers*, each of which is already implicit in the text and explicated by the interpretative activity of the reader/scholar through specific tools and visual solutions provided by the system.

2.1 Minimal User Requirements

In order to define a minimal set of user requirements we first introduce the following key terms: *textual source*, *layer*, *element*, *relation* and *network*. As ***textual source*** we intend every object presenting at least one grapheme, which has been either digitized or scanned as image and uploaded into the system (i.e., page from a digitized literary book, image of an inscribed pottery, image of a folium from a manuscript, transcription of a manuscript). The term *source* can refer to (i.) a textual corpus (i.e., Dante's writings), (ii.) a specific section/unit/book of the given corpus (i.e., Inferno), and (iii.) a passage from a specific book of a given corpus (i.e., XVI Canto of Inferno). A ***layer*** is a specific set of features embedded in a given *textual source*, which can be explicated by users through analysis and annotation tools. Each *source* exhibits, at least, a graphemic layer (grapheme/s on a given writing surface) and may include an unlimited number of *layers*, according to the user's research interest. Some basic *layers* (i.e., graphemic, phonetic, terminological, ontological) are already provided by the DSS, while others (arbitrary layers) can be defined by users (e.g., dialogical layer, anthropological layer). An ***element*** is an atomic unit forming a *layer*, i.e. a grapheme of the graphemic layer, a phoneme of the phonetic layer, a term of the terminological layer, or a concept of the ontological layer; an *element* can be visualized as a node of a network in the interface of the DSS. A ***relation*** is a link between two or more *elements*, intra and inter-layer; a relation can be visualized as an arc of a *network* in the interface of the DSS. Finally, a ***network*** is a set of *elements* and the *relations* among them visualized as a graph.

We have grouped the minimal requirements we identified for the development of our DSS in four main categories. To the first group, **(A.) Upload and Source Management**, belong the following requirements: **(1.)** creation of a new *research project*; **(2.)** management of a variety of different *research projects* for each user; **(3.)** upload of the relevant *sources* for a specific *project*; **(4.)** running of OCR on the scanned source, when dealing with images of manuscripts or material objects; **(5.)** sharing of selected sources with selected users; **(6.)** execution of catalographic searches. To the second group, **(B.) Layers**, belong: **(1.)** use of predefined basic *layers* **(2.)** definition of arbitrary *layers*; **(3.)** use of (manual and automatic) tools for the elicitation of the elements of a specific *layer*; **(4.)** addition of notes (footnotes, endnotes, general notes, philological, linguistic, ...) and comments of different types to a specific *element*. To the third category, **(C) Research and Comparison**: **(1.)** execution of searches on the selected textual sources within one or more *layers*; **(2.)** execution of searches with boolean and regular expressions; **(3.)** execution of manual and semi-automatic comparisons between two or more *sources*, also on different *layers*, by presenting them together on the screen; **(4.)** highlighting of the differences between two or more *sources* selected for the comparison; **(5.)** highlighting of features shared by two or more *sources* selected for the comparison; **(6.)** visualization of the results of each specific search and comparison in structured lists. Finally, for the fourth category, **(D) Construction of Networks**, we identified the following requirements: **(1.)** manual or, when possible, automatic construction of a *network*, realized by defining *relations* among *elements* belonging to the same *layer* or different *layers*; **(2.)** editing of an automatically generated *network*.

3 A Possible Use Case on Dante's Inferno

Here, we present a possible use case on Dante's Inferno, a highly complex and rich writing, which gathers a great amount of information, thus requiring very different scholarly skills to be fully understood and analysed. Particularly, our use case studies the dialogues of Guelfi and Ghibellini, two rival Florentine political factions. Although in our vision the DSS would enable users to annotate chunks of text as dialogues and to define the *text ontology* (Bellandi et al., 2013) including the characters of the *al di là*, we chose to exploit an existing XML-encoded advanced representation of Inferno (Cappelli et al., 2011).

An analysis of this type can be articulated in a series of steps, each one bringing to the construction of a portion of the *network* (requirement

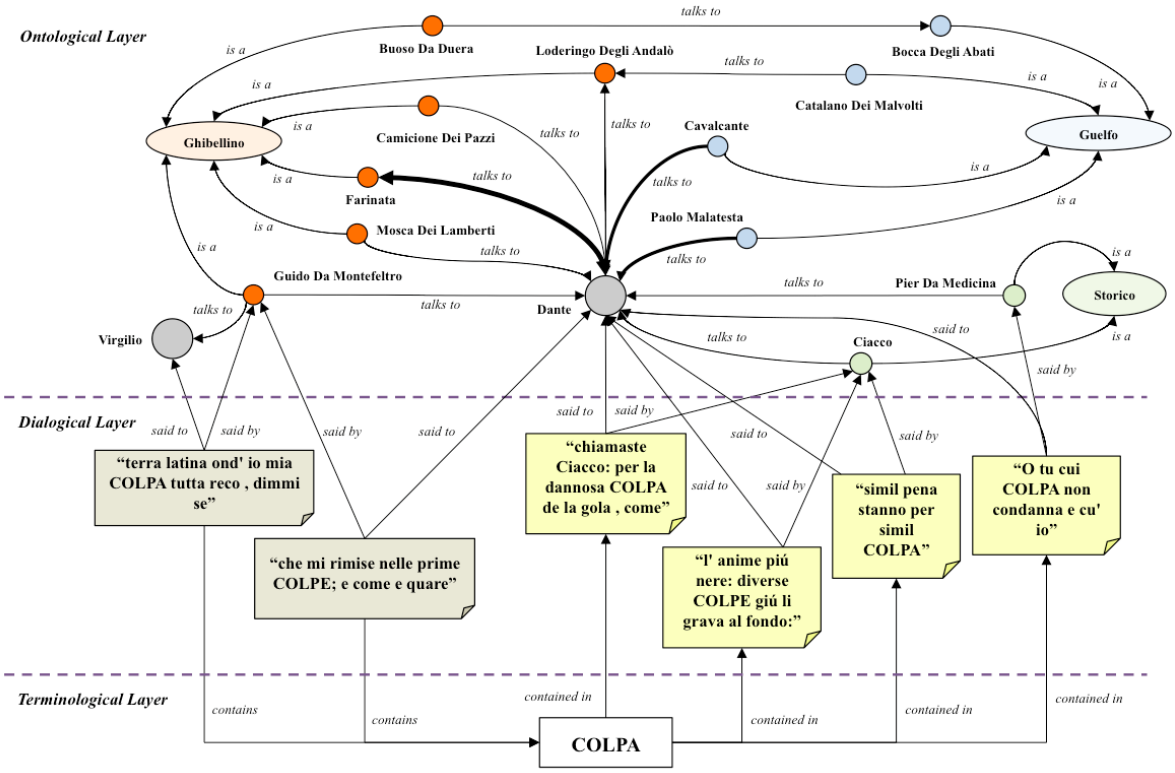

Figura 1: Example of network

D.1), of Figure 1. The first step involves the *ontological layer* (requirement B.1): the user would build the upper part of the network by introducing the relation *talks to* (the thickness of the relative arc representing the number of dialogical interactions) among the elements *Guelfo*, *Ghibellino*, *Dante*, and *Virgilio*. The obtained network shows that the only interactions between the two factions are those of *Buoso Da Duera* who talks to *Bocca degli Abati*, and *Catalano Dei Malvolti* who talks to *Loderingo Degli Andalò*. Furthermore, *Guido Da Montefeltro* is the only *Ghibellino* who talks to both *Dante* and *Virgilio*. The user could then be interested in analysing his dialogues (the two added on the left part of the network as elements of the dialogical layer), by using a terminology extractor, bringing to the elicitation of the elements (terms) constituting the terminological layer (requirement B.3). The user could select the term *colpa* ("guilt" in English) since being present in both dialogues and add it to the network. In the final part of this example the user could verify if the term *colpa* appears in other dialogues. To do this the user would search the pattern "colp[ae]" (representing the singular and plural forms of the lemma *colpa*) inside the elements of the dialogical layer (requirement C.2). As a result, the network would be populated with four more dialogues, showing that only *Ciacco* and *Pier Da Medicina* talk to *Dante* using the term *colpa*. These two characters are not politically characterized, being classified, in the ontology, as "Storico" (historical character).

4 Conclusions

In this work, we presented our vision of a Decision Support System for the analysis and interpretation of texts. In addition to outlining the general characteristics of the system, we illustrated a case study on Dante's Inferno showing how the study of a text can involve elements belonging to three different layers (ontological, dialogical and terminological) thus allowing to take into account, in an innovative way, both textual and contextual elements.

The next steps will consist in the extension of the user requirements and the design of the main components of the system. We plan to start with the basic features allowing a user to create a project and upload documents and then provide the minimal text processing tools necessary for the definition and management of (at least) the graphemic layer.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 88-96, Montréal, Canada. June 8.

Miriam Baglioni, Mirco Nanni, and Emiliano Giovannetti. 2004. Mining literary texts by using domain ontologies. In *Proceedings of the Workshop on Knowledge Discovery and Ontologies (KDO-2004).* Pisa, Italy. September 20-24.

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti, and Enrico Carniani. 2013. Content Elicitation: Towards a New Paradigm for the Analysis and Interpretation of Text. In *Proceedings of the IASTED International Conference on Informatics.* Innsbruck, Austria. February 17-19.

Andrea Bellandi, Alessia Bellusci, Amedeo Cappelli, and Emiliano Giovannetti. 2014. Graphic Visualization in Literary Text Interpretation. In *Proceedings of the 18th International Conference on Information Visualisation.* Paris, France. July 15-18.

Andrea Bolioli, Matteo Casu, Maurizio Lana, and Renato Roda. 2013. Exploring the Betrothed Lovers. *OASIcs-OpenAccess Series in Informatics*, 32:30–35.

Silvia Bordin, Massimo Zancanaro, and Antonell De Angeli. 2013. Touching Dante: A Proximity-based Paradigm for Tabletop Browsing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, 27:1-10. Trento, Italy. September 16-20.

Amedeo Cappelli, Maria Novella Catarsi, Patrizia Michelassi, Lorenzo Moretti, Miriam Baglioni, Franco Turini, and Mirko Tavoni. 2002. Knowledge Mining and Discovery for Searching in Literary Texts. In *Proceedings of LREC 2002.* Las Palmas, Canary Islands, Spain. 29-31 May.

Amedeo Cappelli, Michele Coscia, Fosca Giannotti, Dino Pedreschi, and Salvo Rinzivillo. 2011. The social network of Dante's Inferno. *Leonardo*, 44(3):246–247.

Daniel Chandler. 1995. *The Act of Writing: A Media Theory Approach* . Aberystwyth, pp. 4-8.

Jan De Vuyst. 1990. Knowledge representation for text interpretation. *Literary and linguistic computing* 5(4): 296–302.

David Easley, and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Umberto Eco. 1979. *Lector in Fabula*. Bompiani, Milano.

Stevan Harnad. 1990. Against Computational Hermeneutics. *Social Epistemology*, 4:167–172.

Robert Hollander. 2013. The Princeton Dante Project. *Humanist Studies and the Digital Age* 3(1):53-59. http://etcweb.princeton.edu/dante/index.html

Jean W. LeLoup, and Robert Ponterio. 2006. Dante: Digital and on the Web. *Language Learning & Technology* 10(1): 3–8. http://dante.ilt.columbia.edu

Isabel Meirelles. 2013. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers.

Thomas Muhr. 1991. ATLAS.ti - A Prototype for the Support of Text Interpretation. *Qualitative Sociology* 14:349–371. Human Science Press, New York.

Mark E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*. 45:167–256.

Mark E. J. Newman, Albert László Barabási, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton University Press.

Deborah Parker. 2001. The World of Dante: a hypermedia archive for the study of the inferno. *Literary and linguistic computing* 16(3): 287–297. http://www.worldofdante.org/about.html

Antonio Roque. 2012. Towards a computational approach to literary text analysis. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 97–104, Montréal, Canada. June 8.

Cesare Segre, 1985. Testo letterario, interpretazione, storia: linee concettuali e categorie critiche. In *Alberto Asor Rosa: Letteratura italiana* 21–140. Einaudi, Torino.

Jung P. Shim, Merril Warkentin, James F. Courtney, Daniel J. Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision support systems*, 33(2):111–126. Elsevier.

Michael Sperberg-McQueen, and Lou Burnard. 1994. *Guidelines for electronic text encoding and interchange*, 1. Chicago and Oxford: Text Encoding Initiative.

Kathryn Sutherland. 1990. A Guide Through the Labyrinth: Dickens's Little Dorrit as Hypertext. *Literary and Linguistic Computing*. 5(4):305–309.

Mirko Tavoni, Paola Andriani, Valentina Bartalesi, Elvira Locuratolo, Carlo Meghini, and Loredana Versienti. 2014. Towards a semantic network of Dante's works and their contextual knowledge. In *Proceedings of The Digital Humanities 2014 conference*. Lausanne, Switzerland. July 7-12.

Tzevetan Todorov. 1973. Postscriptum. *In R. Jakobson, Questions de poétique* 485–504. Editions du Seuil, Paris.