Workshop
Toward Computational Models of Literary Analysis

May 22nd, 2006, Genoa (Italy),

jointly held with **LANGUAGE RESOURCES AND EVALUATION, LREC 2006**

It has been often noticed that computer based literary critics is still relying on studies of concordances as traditionally intended since the 13th century. All the intermediate digital representations (storage, indexes, data structures or records) are not capitalized although they can play the role of a new literary "monster" (like the Cheiron centaur) as a new meaningful, artistic and hermeneutic macro unit. It is indeed true that the digital representation, its metadata and its digital derivatives (e.g. indexes, parse trees, semantic references to external dictionaries) are new and more complex forms of "concordances" and should be used by the literary scholar in cooperation with the original content. New processes of narrative analysis should thus take all of this into account by exploiting the fruitful interactions among the parts of the monster within suitable software architectures (that are thus more complex than digital archives/catalogs).

In the Natural Language Processing research community, a wide range of computational methods have been successfully applied to information and document management, spanning from text categorization and information extraction, to ontology leaning, text mining and automatic semantic markup. Although these techniques are mostly applied to technical texts in application-driven contexts, their application range could be expanded to encompass a larger typology of texts, thereby gaining new powerful insights for the analysis of literary text content and paving the way for new experiments and forms of text hermeneutics. The development of language resources in this area is also rather limited and more interdisciplinary research is needed to open the field to realistic and effective applications.

The long term research enterprise in this field should aim to design novel paradigms for literary studies that are:

- more information-centered, as they work at a higher level of abstraction
- interactive with the scholar, as the software is proactive with respect to the literary work
- multifunctional and integrated as they support incremental refinement of internal knowledge of the opera along with more interaction with the expert takes place.

This workshop aims to gather studies, achievements and experiences from scholars belonging to different schools (literary studies, linguistics, computing technologies, artificial intelligence, human-computer interaction) in order to survey, compare and assess currently independent research enterprises whose focus is narrative and literary text analysis. The aim is to discuss at which extent the textual evidences currently observable through digital technologies can support the computational treatment of narrative and literary phenomena. Results in these area have an invaluable impact on the technological side (as a novel challenge for computational models of language and narrative) as well as on the cultural side (as new perspectives for human-computer interaction and modern literary analysis). Moreover, the enormous potentials offered to cultural heritage preservation and dissemination are evident.

The Workshop Organisers

Roberto Basili (University of Roma Tor Vergata, Italy) (co-chair)
Alessandro Lenci (University of Pisa, Italy) (co-chair)

The Workshop Programme

14:25-14:30 WELCOME AND INTRODUCTORY NOTES

14:30-15:10 INVITED TALK: *"When Will Computers Understand Shakespeare?"*
Jerry Hobbs (ISI, California, USA)

PAPER SESSION 1

15:10-15:30 *Semantic Domains and Linguistic Theory*
Alfio Ghiozzo (ITC-TCC, Trento, Italy)

15:30-15:50 *Computational linguistics meets philosophy: a Latent Semantic Analysis of Giordano Bruno's texts*
Simonetta Bassi, Felice Dell'Orletta, Daniele Esposito, Alessandro Lenci (University of Pisa, Italy)

15:50-16:10 *A geometrical approach to literary text analysis*
Roberto Basili, Paolo Marocco (University of Roma, Tor Vergata, Italy)

16:10-16:30 *The encoding model of Puccini's Correspondence Project*
Elena Pierazzo (University of Pisa, Italy)

16:30-17:00 COFFE BREAK

PAPER SESSION 2

17:00-17:20 *Stylogenetics: Clustering-based stylistic analysis of literary corpora*
Kim Luyckx, Walter Daelemans, Edward Vanhoutte (University of Antwerp, Netherlands)

17:20-17:40 *Sentiment Classification Techniques for Tracking Literary Reputation*,
Maite Taboada, Mary Ann Gillies, Paul McFetridge, (Simon Fraser University, Canada)

17:40-18:00 *Narrative Models: Narratology Meets Artificial Intelligence*
Pablo Gervás, Federico Peinado (Universidad Complutense de Madrid, Spain), Birte Lönneker-Rodman (University of Hamburg, Germany), Jan Christoph Meister (Ludwig-Maximilians University, Germany)

18:00-18:15 *Cognitive Emotion Theories, Emotional Agent, and Narratology*,
Gesine Lenore Schiewer (University of Bern, Switzerland)

18:15-18:30 *Locating proverbs with finite-state transducers in literary texts*
Olympia Tsaknaki (IGM, Université de Marne-la-Vallée, France)

18:30-19:00 PANEL SESSION

19:00 CLOSING REMARKS

Workshop Organiser(s)

Roberto Basili, (University of Roma Tor Vergata, Italy)
basili@info.uniroma2.it

Alessandro Lenci (University of Pisa, Italy) (co-chair)
alessandro.lenci@ilc.cnr.it

Workshop Programme Committee

Roberto Basili	(*University of Roma Tor Vergata, Italy*)
Simonetta Bassi	(*University of Pisa & SIGNUM, SNS, Italy*)
Marc Cavazza	(*University of Teeside, UK*)
Richard Coyne	(*University of Edinburgh, UK*)
Pierantonio Frare	(*University of Milan, Italy*)
Andrea Gareffi	(*University of Roma, Tor Vergata, Italy*)
Graeme Hirst	(*University of Toronto, Canada*)
Jerry Hobbs	(*ISI, University of Southern California, USA*)
Hugh Craig	(*University of Newcastle, Australia*)
Alessandro Lenci	(*University of Pisa, Italy*)
Marco Pennacchiotti	(*University of Roma, Tor Vergata, Italy*)
Mirko Tavoni	(*University of Pisa, Italy*)
Yorick Wilks	(*University of Sheffield, UK*)

Workshop Home Page:

http://ai-nlp.info.uniroma2.it/basili/LREC2006/TowardsCompModels.html

Table of Contents

Author Index

Semantic Domains and Linguistic Theory

Alfio Gliozzo
University of Rome Tor Vergata
Department of Computer Science, System and Production
00133 Roma (Italy)
gliozzo@itc.it

Abstract

This paper is about the relations between the concept of Semantic Domain and the "Theory of Semantic Fields", a structural model for lexical semantics proposed by Jost Trier at the beginning of the last century. The main limitation of the Trier's notion is that it does not provide an objective criterion to aggregate words around fields, making the overall model too vague, and then unuseful for computational purposes. The notion of Semantic Domain improves that of Semantic Field by providing such a criterion. In particular, the structuralist approach in semantics has been connected to the Wittgenstein's *meaning-is-use* assumption, providing an objective criterion to infer Semantic Domains from corpora relying on a lexical coherence assumption. The task based evaluation we did for our claims shows that the notion of Semantic Domains is effective because it allows to define an uniform methodology to deal with many different Natural Language Processing tasks. In the paper we discuss the epistemological issues concerning the possibility of adopting a task based methodology to support linguistic theory, showing the case study of Semantic Domains in Computational Linguistics as a paradigmatic example for our claims.

1 Introduction

The predominant view in lexical semantic is the Saussure's structural semantics (de Saussure, 1922), claiming that a word meaning is determined by the "horizontal" paradigmatic and the "vertical" syntagmatic relations between that word and others in the whole language (Lyons, 1977). Structural assumptions are also widely adopted in Computational Linguistic. For example, many machine readable dictionaries describe the word senses by means of semantic networks representing relations among terms (e.g. WORDNET (Miller, 1990)). The main limitation of the "radical" structuralist view is that it is almost impossible to describe the associations among all the possible terms in a natural language, because the huge number of concepts and semantic relations among them.

The Semantic Fields Theory (Trier, 1931) goes a step further in the structural approach to lexical semantics by introducing an additional aggregation level and by delimiting to which extend paradigmatic relations hold. The basic assumption of this theory is that the lexicon is structured into Semantic Fields: semantic relations among concepts belonging to the same field are very dense, while concepts belonging to different fields are typically unrelated. In fact, a word meaning is established only by the network of relations among the terms of its field. Another property of great interest is that there exists a strong correspondence among Semantic Fields of different languages, while such a strong correspondence cannot be established among the terms themselves.

It has been observed that the main limitation of the Trier's notion is that it does not provide an objective criterion to aggregate words around fields, making the overall model too vague, and then unuseful for computational purposes. The notion of Semantic Domain improves that of Semantic Field by providing such a criterion. In particular, the structuralist approach in semantics has been connected to the *meaning-is-use* assumption introduced by Ludwig Wittgenstein in his celebrated "Philosophical Investigations" (Wittgenstein, 1965). A word meaning is its use into the concrete "form of life" where it is adopted, i.e. the *linguistic game*, in the Wittgenstein's terminology. Frequently co-occurring words in texts are then associated to the same linguistic game. It follows that fields can be identified from a corpus based analysis of the lexicon, exploiting the connections between linguistic games and Semantic Fields already depicted. The notion of Semantic Domain arises from this convergence, providing an objective criterion to identify semantically related words in texts, supported by a *lexical coherence* assumption, that we empirically corroborated in text in the earlier stages of our work.

The notion of Semantic Domain is intimately related to several phenomena in the language at both a lexical and a textual level. At a lexical level Semantic Domains can be used as a (shallow) model for lexical ambiguity and variability, while at a textual level semantic domains provide meaningful topic taxonomies that can be used to group texts into semantic clusters. In addition, the inherent multilingual nature of semantic domains allows an uniform representation of both the lexicon and the texts in most of the natural languages.

Exploiting Semantic Domains for Natural Language Processing (NLP) allowed us to improve sensibly the state-of-the-art in all those tasks in which they have been applied, providing and indirect evidence to support their linguistic properties. The major goal of this paper is to discuss the possibility of adopting such a task based methodology to support linguistic theory, showing the case study of Semantic Domains in computational linguistics as a successfully paradigmatic example of our methodology.

The paper is structured as follows. Section 2 is about the Semantic Fields Theory while Section 3 concerns the relations between this theory and the Wittgenstein's meaning-is-use assumption. Section 4 describes the concept of Semantic domains as the confluence of both perspectives, highlighting its technological impact in developing state-of-the-art systems for NLP, while Section 5 conclude the paper discussing the possibility of adopting the indirect task based evaluation to support linguistic theory.

2 The Theory of Semantic Fields

Semantic Domains are a matter of recent interest in Computational Linguistics (Magnini and Cavaglià, 2000; Magnini et al., 2002; Gliozzo et al., 2005a), even though their basic assumptions are inspired from a long standing research direction in structural linguistics started in the beginning of the last century and widely known as "The Theory of Semantic Fields" (Lyons, 1977). The notion of *Semantic Field* has proved its worth in a great volume of studies, and has been mainly put forward by Jost Trier (Trier, 1931), whose work is credited with having "opened a new phase in the history of semantics"(Ullmann, 1957).

In that work, it has been claimed that the lexicon is structured in clusters of very closely related concepts, lexicalized by sets of words. Word senses are determined and delimitated only by the meanings of other words in the same field. Such clusters of semantically related terms have been called Semantic Fields[1], and the theory explaining their properties is known as "The theory of Semantic Fields" (Vassilyev, 1974).

Semantic Fields are conceptual regions shared out amongst a number of words. Each field is viewed as a partial region of the whole expanse of ideas that is covered by the vocabulary of a language. Such areas are referred to by groups of semantically related words, i.e. the Semantic Fields. Internally to each field, a word meaning is determined by the network of relations established with other words.

There exists a strong correspondence among Semantic Fields of different languages, while such a strong correspondence cannot be established among the terms themselves. For example, the field of COLORS is structured differently in different lan-

[1]There is no agreement on the terminology adopted by different authors. Trier uses the German term *wortfeld* (literally "word field" or "lexical field" in Lyons' terminology) to denote what we call here semantic field.

guages, and sometimes it is very difficult, if not impossible, to translate name of colors, even whether the chromatic spectrum perceived by people in different countries (i.e. the conceptual field) is the same. Some languages adopt many words to denote the chromatic range to which the English term `white` refers, distinguishing among different degrees of "whiteness" that have not a direct translation in English. Anyway, the chromatic range covered by the COLORS fields of different languages is evidently the same. The meaning of each term is defined in virtue of its oppositions with other terms of the same field. Different languages have different distinctions, but the field of COLORS itself is a constant among all the languages.

Another implication of the Semantic Fields Theory is that words belonging to different fields are basically unrelated. In fact, a word meaning is established only by the network of relations among the terms of its field. As far as paradigmatic relations are concerned, two words belonging to different fields are then un-related. This observation is crucial form a methodological point of view. The practical advantage of adopting the Semantic Field Theory in linguistics is that it allows a large scale structural analysis of the whole lexicon of a language, otherwise infeasible. In fact, restricting the attention to a particular field is a way to reduce the complexity of the overall task of finding relations among words in the whole lexicon, that is evidently quadratic in the number of words.

The main limitation of the Trier's theory is that it does not provide any objective criterion to identify and delimitate Semantic Fields in the language. The author himself admits "what symptoms, what characteristic features entitle the linguist to assume that in some place or other of the whole vocabulary there is a field? What are the linguistic considerations that guide the grasp with which he selects certain elements as belonging to a field, in order then to examine them as a field?" (Trier, 1934). The answer to this question is an issue opened by the Trier's work.

3 Semantic Fields and the meaning-is-use view

In the previous section we have pointed out that the main limitation of the Trier's theory is the gap of an objective criterion to characterize Semantic Fields. The solutions we have found in the literature (Weisgerber, 1939; Porzig, 1934; Coseriu, 1964) rely on very obscure notions, of scarse interest from a computational point of view. To overcome such a limitation, in this section we introduce the concept of Semantic Domain (see Section 4).

The notion of Semantic Domain improves that of Semantic Fields by connecting the structuralist approach in semantics to the *meaning-is-use* assumption introduced by Ludwig Wittgenstein in his celebrated "Philosophical Investigations" (Wittgenstein, 1965). A word meaning is its use into the concrete "form of life" where it is adopted, i.e. the *linguistic game*, in the Wittgenstein's terminology. Words are then meaningful only if they are expressed into concrete and situated linguistic games that provide the conditions for determining the meaning of natural language expressions. To illustrate this concept, Wittgenstein provided a clarifying example describing a very basic linguistic game: "...Let us imagine a language ...The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones; there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words *block*, *pillar*, *slab*, *beam*. A calls them out; – B brings the stone which he has learnt to bring at such-and-such a call. – Conceive of this as a complete primitive language." (Wittgenstein, 1965)

We observe that the notions of linguistic game and Semantic Field show many interesting connections. They approach the same problem from two different points of view, getting to a similar conclusion. According to Trier's view, words are meaningful when they belong to a specific Semantic Field, and their meaning is determined by the structure of the lexicon in the field. According to Wittgenstein's view, words are meaningful when there exists a linguistic game in which they can be formulated, and their meaning is exactly their use. In both cases, meaning arises from the wider contexts in which words are located.

Words appearing frequently into the same linguistic game are likely to be located into the same field. In the previous example the words *block*, *pillar*, *slab* and *beam* have been used in a common lin-

guistic game, while they clearly belong to the Semantic Field of BUILDING INDUSTRY. This example suggests that the notion of linguistic game provides a criterion to identify and to delimitate Semantic Fields. In particular, the recognition of the linguistic game in which words are typically formulated can be used as a criterion to identify classes of words composing lexical fields. The main problem of this assumption is that it is not clear how to distinguish linguistic games between each other. In fact, linguistic games are related by a complex network of similarities, but it is not possible to identify a set of discriminating features that allows us to univocally recognize them. "I can think of no better expression to characterize these similarities than 'family resemblances'; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way. - And I shall say: 'games' form a family" ((Wittgenstein, 1965), par. 67).

We observe that linguistic games are naturally reflected in texts, allowing us to detect them from a word distribution analysis on a large scale corpus. In fact, according to Wittgenstein's view, the content of any text is located into a specific linguistic game, otherwise the text itself would be meaningless. Texts can be perceived as open windows through which we can observe the connections among concepts in the real world. Frequently co-occurring words in texts are then associated to the same linguistic game.

It follows that the set of concepts belonging to a particular field can be identified from a corpus based analysis of the lexicon, exploiting the connections between linguistic games and Semantic Fields already depicted. For example, the two words *fork* and *glass* are evidently in the same field. A corpus based analysis shows that they frequently co-occur in texts, then they are also related to the same linguistic game. On the other hand, it is not clear what would be the relation among *water* and *algorithm*, if any. They are totally unrelated simply because the concrete situations (i.e. the linguistic games) in which they occur are in general distinct. It reflects on the fact that they are often expressed in different texts, then they belong to different fields.

Our proposal is then to merge the notion of linguistic game and that of Semantic Field, in order to provide an objective criterion to distinguish and delimitate fields from a corpus based analysis of lexical co-occurences in texts. We refer to this particular view on Semantic Fields by using the name Semantic Domains. The concept of Semantic Domain is the main topic of this work, and it will be illustrated more formally in the following section.

4 Semantic Domains

In our usage, Semantic Domains are common areas of human discussion, such as ECONOMICS, POLITICS, LAW, SCIENCE, which demonstrate lexical coherence. The Semantic Domain associated to a particular field is the set of domain specific terms belonging to it, and it is characterized by a set of *domain words* whose main property is to co-occur in texts.

An approximation to domains are Subject Field Codes, used in Lexicography (e.g. in (Procter, 1978)) to mark technical usages of words. Although this information is useful for sense discrimination, in dictionaries it is typically used only for a small portion of the lexicon. WORDNET DOMAINS (Magnini and Cavaglià, 2000) is an attempt to extend the coverage of domain labels within an already existing lexical database, WORDNET (Fellbaum, 1998). As a result WORDNET DOMAINS can be considered an extension of WORDNET in which synsets have been manually annotated with one or more domain labels, selected from a hierarchically organized set of about two hundred labels.

WORDNET DOMAINS represents the first attempt to provide an exhaustive systematization of the concept of Semantic Field and its connections to the textual interpretation depicted in section 3. It allowed us to start an empirical investigation about the connections between the textual and the lexical counterparts of Semantic Domains. First we concentrated on corroborating a lexical-coherence assumption, claiming that a great percentage of the concepts expressed in the same text belong to the same domain. Lexical coherence is then a basic property of most of the texts expressed in any natural language and it allows us to disambiguate words in context by associating domain specific senses to them. Otherwise stated, words taken out of context show domain polysemy, but, when they occur into real texts,

their polysemy is solved by the relations among their senses and the domain specific concepts occurring in their contests.

Intuitively, texts may exhibit somewhat stronger or weaker orientation towards specific domains, but it seems less sensible to have a text that is not related to at least one domain. In other words, it is difficult to find a "generic" text. This intuition is largely supported by our data: all the texts in SemCor [2](Landes et al., 1998) exhibit concepts belonging to a small number of relevant domains, demonstrating the domain coherence of the lexical-concepts expressed in the same text. In particular, 34.5 % of nouns in co-occurring in the same texts in SemCor are annotated with the same domain label, while about 40% refer to generic concepts. The conclusion of this experiment is that there exists a strong tendency for the lexicon in texts to be aggregate around a specific domain. As we will see later in the paper, such a tendency should be presupposed to allow lexical disambiguation.

Then we investigated the relations between Semantic Domains and lexical ambiguity and variability, the two most basic and pervasive phenomena characterizing lexical semantics. The different senses of ambiguous words should be necessarily located into different domains, because they are characterized by different relations with different words. On the other hand, variability can be modeled by observing that synonymous terms refer to the same concepts, then they will necessarily belong to the same domain. Thus, the distribution of words among different domains is a relevant aspect to be taken into account to identify word senses. Understanding words in contexts is mainly the operation of locating them into the appropriate semantic fields.

To corroborate these assumptions we developed a Word Sense Disambiguation (WSD) procedure relying on domain information only, named Domain Driven Disambiguation (DDD) (Magnini et al., 2001; Gliozzo et al., 2004). The underlying hypothesis of the DDD approach is that information provided by domain labels offers a natural way to establish associations among word senses in a certain text fragment, which can be profitably used during the disambiguation process. DDD is performed by selecting the word sense whose Semantic Domain maximize the similarity with the domain of the context in which the word is located. For example, the word *virus* is ambiguous between its `Biology` and `Computer Science` senses, and can be disambiguated by assigning the correct domain to the contexts where it actually occurs. Results clearly shows that domain information is crucial for WSD, allowing our system to improve the state-of-the-art for unsupervised WSD.

The main conclusion of that work was that Semantic Domains play a dual role in linguistic description. One role is characterizing word senses (i.e. *lexical-concepts*), typically by assigning domain labels to word senses in a dictionary or lexicon. On the other hand, at a text level, Semantic Domains are clusters of texts regarding similar topics/subjects. They can be perceived as collections of domain specific texts, in which a generic corpus is organized. Examples of Semantic Domains at the text level are the subject taxonomies adopted to organize books in libraries.

The generality of these results encouraged us to extend the range of applicability of our assumptions, leading to the definition of a large number of NLP techniques relying on the common theoretical framework provided by Semantic Domains in computational linguistics (Gliozzo, 2005). For brevity, we will not describe into details all these results, limiting ourselves to enumerate the range of applicability of domain driven techniques in NLP: Word Sense Disambiguation (Gliozzo et al., 2005b), Text Categorization (Gliozzo and Strapparava, 2005b), Term Categorization (D'Avanzo et al., 2005), Ontology Learning (Gliozzo, 2006) and Multilinguality (Gliozzo and Strapparava, 2005a).

In all those tasks state-of-the-art results have been achieved by following the common methodology of acquiring Domain Models from texts by means of a common corpus based technique, inspired and motivated by the Trier's theory and by its connection to the meaning-is-use assumption. In particular we adopted an approach based on Latent Semantic Analysis to acquire domain models from corpora describing the application domain, and we assumed the principal components so acquired be mapped to a set of semantic domains. Latent Semantic Analysis

[2]Semcor is a subportion of the Brown corpus annotated by WordNet senses.

has been performed on a term-by-document matrix capturing only co-occurrency information among terms in texts, with the aim of demonstrating our meaning-is-use assumptions. Then we exploited domain based representations to index both terms and texts, adopting a semi-supervised learning paradigm based on kernel methods. Empirical results showed that domain based representations performs better than standard bag-of-words commonly adopted for retrieval purposed, allowing a better generalization over the training data (i.e. improving the learning curve in all the supervised tasks in which they have been applyied), and allowing the definition of hybrid similarity measures to compare terms and texts, as expected from the notion of Semantic Domain.

5 Conclusion

In this paper we explicitly depicted the connections between the use of Semantic Domains in NLP and the linguistic theory motivating them. Understanding these relations provided us an useful guideline to lead our research, leading to the definition of state-of-the-art techniques for a wide range of tasks. Having in mind a clear picture of the semantic phenomena we where modeling allowed us to identify the correct applications, to predict the results of the experiments and to motivate them.

Nonetheless, several questions arises when looking at semantic domains from an epistemological point of view:

1. is the concept of semantic domain a computational theory for lexical semantics?
2. do we have enough empirical evidence to support our linguistic claims?
3. is the task based evaluation a valid epistemological framework to corroborate linguistic theory?

My personal point of view is that the task based evaluation is probably the only objective support we can provide to linguistic theory, and especially to all those issues that are more intimately related to lexical semantics. The basic motivation is that computational linguistics is also a branch of Artificial Intelligence, and then it is subjected to the behavioral Turing test. The *meaning-is-use* assumption fits perfectly this view, preventing us from applying the traditional linguistic epistemology to computational linguistics. In fact, we are interested in exploiting the language in concrete and situated linguistic games rather than representing it in an intensional way. From this point of view, the task based support we have given to our claims is a strong evidence to conclude that Semantic Domains are computational models for lexical semantics.

Anyhow, my opinion is just a minor contribution to stimulate a larger epistemological debate involving linguists, cognitive scientists, philosophers, computer scientists, engineers, among the others. I hope that my research will contribute to stimulate this debate and to find a way to escape from the "empasse" caused by the vicious distinction between empirical and theoretical methods characterizing the research in computational linguistics in the last decade.

References

E. Coseriu. 1964. Pour une sémantique diachronique structurale. *Travaux de Linguistique et de Littérature*, 2(1):139–186.

E. D'Avanzo, A. Gliozzo, and C. Strapparava. 2005. Automatic acquisition of domain information for lexical concepts. In *Proceedings of the 2^{nd} MEANING workshop*, Trento, Italy.

F. de Saussure. 1922. *Cours de linguistique générale*. Payot, Paris.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. MIT Press.

A. Gliozzo and C. Strapparava. 2005a. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *n Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

A. Gliozzo and C. Strapparava. 2005b. Domain kernels for text categorization. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 56–63.

A. Gliozzo, C. Strapparava, and I. Dagan. 2004. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech and Language*, 18(3):275–299.

A. Gliozzo, C. Giuliano, and C. Strapparava. 2005a. Domain kernels for word sense disambiguation. In *Proceedings of the 43^{rd} annual meeting of the Association*

for Computational Linguistics (ACL-05), pages 403–410, Ann Arbor, Michigan, June.

A. Gliozzo, C. Giuliano, and C. Strapparava. 2005b. Domain kernels for words sense disambiguation. In *to appear in proc. of ACL-2005*.

A. Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, University of Trento.

A. Gliozzo. 2006. The god model. In *proceedings of EACL-06*.

S. Landes, C. Leacock, and R. I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press.

J. Lyons. 1977. *Semantics*. Cambridge University Press.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, pages 1413–1418, Athens, Greece, June.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2001. Using domain information for word sense disambiguation. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation System*, pages 111–114, Toulose, France, July.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

G. Miller. 1990. An on-line lexical database. *International Journal of Lexicography*, 13(4):235–312.

W. Porzig. 1934. Wesenhafte bedeutungsbeziehungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 58.

Procter. 1978. *Longman Dictionary of Contemporary English*.

J. Trier. 1931. *Der deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg.

J. Trier. 1934. Das sprachliche feld. eine auseinandersetzung. *Neue Fachrbücher für Wissenschaft und Jugendbildung*, 10:428–449.

S. Ullmann. 1957. *The Principles of Semantics*. Blackwell, Oxford.

L.M. Vassilyev. 1974. The theory of semantic fields: a survey. *Linguistics*, 137:79–93.

L. Weisgerber. 1939. Vom inhaltlichen aufbau des deutschen wortschatzes. *Wortfeldforschung*, pages 193–225.

L. Wittgenstein. 1965. *Philosophical Investigations*. The Macmillan Company, New York.

Computational linguistics meets philosophy: a Latent Semantic Analysis of Giordano Bruno's texts

[1,2]Simonetta Bassi, [3]Felice Dell'Orletta, [1,2]Daniele Esposito, [4]Alessandro Lenci

[1]Università di Pisa, Dipartimento di Filosofia
P.za Torricelli 3/A, 56126, Pisa, Italy
sbassi@fls.unipi.it, frogproduction@gmail.com

[2]SIGNUM, Scuola Normale Superiore
Via della Faggiola 19, 56126, Pisa, Italy

[3]Università di Pisa, Dipartimento di Informatica
Largo B. Pontecorvo 3, 56126, Pisa, Italy
felice.dellorletta@ilc.cnr.it

[4]Università di Pisa, Dipartimento di Linguistica "T. Bolelli"
Via Santa Maria 36, 56126, Pisa, Italy
alessandro.lenci@ilc.cnr.it

Abstract

Traditional methods such as concordance lists are not able to provide us with a real access to the semantic space of a philosophical text and to its dynamics. In this paper, we will argue that new insights on these points can come from the application of more sophisticated mathematical and computational methods of meaning representation, based on the automatic construction of text-driven word similarity spaces out of the way word distribute and co-occur in texts. More specifically, we report on a current research in which semantic spaces are dynamically built by applying a variation of Latent Semantic Analysis to the *Eroici furori* by Giordano Bruno, the XVI century Italian philosopher.

1. Introduction

One of the major challenges in the analysis of philosophical texts is how to track the variation of meaning of some notable word and concept within a single text and across different texts of the same author. In a philosophical context, word polysemy is actually a highly value-loaded phenomenon and forms an inherent part of the way an author designs his thought and interacts with the philosophical tradition. Moreover, the same word can be used in very different ways by the same author in different periods of his production, as a consequence of the evolution of his thought. Every single philosophical text thus defines its own semantic space in which words and concepts can be variously located depending on the meaning similarities and associations they acquire in the text. Focusing on the way semantic spaces emerge out of texts and dynamically change can give us new instruments for text interpretation and philosophical investigation in general. This point is crucially related to the issue of how we can use information-based technology to increase our way to explore text content. In fact, it is clear that traditional methods such as concordance lists are not able to provide us with a real access to the semantic space of a text and to its dynamics. In this paper, we will argue that new insights on these points can come from the application of more sophisticated mathematical and computational methods of meaning representation, based on the automatic construction of text-driven word similarity spaces out of the way word distribute and co-occur in texts. More specifically, we report on a current research in which semantic spaces are dynamically built by applying *InfoMap* - a variation of standard Latent Semantic Analysis (Landauer and Dumais 1997) - to one of the most important works of Giordano Bruno, the XVI century Italian philosopher.

2. Bruno's text and philosophical analysis

Bruno's texts provide an important vantage point from which the pros of using computational text analysis can clearly be observed. A real difficulty in the attempt to study the way a word is used and the meanings it assumes in a text (or across a group of texts) is due to the rhetorical strategy used by the author. This problem is particularly evident in the text we used in our experiments: *Eroici furori* (1585)[1]. This text is a dialogue (with two or more characters with different points of view), and it is composed by sonnets, descriptions of allegorical devices, explanations of these sonnets and devices and philosophical dissertations on the topics alluded to in the sonnets and the devices. One of the characters explains to the others the gneoseological experience of love fury, using poems, devices and allegories, and distinguishes between the kind of truth achievable by the wise and the one achievable by the *furioso* "furious", between different kinds of love and intellect, etc. The result is a process, unfolding along the whole text, that produces dramatic

1 The whole text is freely avalaible at BiViO (http://www.bivionline.it), the digital library of Humanism and Renaissance texts developed by Signum (http://www.signum.sns.it) and the Istituto Nazionale di Studi sul Rinascimento (http://www.insr.it).

changes in the meaning of a word, or and that turns a simple concept (e.g. love or intellect) into a series of more complex concepts. Devices and sonnets are not only examples or starting points, but sometimes represent the core of the explanation itself and of the new meaning, as in the metaphor of the *caccia* "hunting", a word directly connected to the research of wisdom (*sapienza*). The myth of the hunter Atteon, who sees the goddess Diana and is transformed in the hunted and devoured by his dogs, is the metaphor of the *furioso*'s fate when he wants to approach the divine truth (*Eroici Furori*, I, IV). In this case we have a metaphorical space that spans across different pages, without being restricted to a single word or sentence. Thus, it is important to focus not only on explicit philosophical meanings, but also on the metaphors they are connected to (e.g. explicit metaphors, as the "hunting" one, or less explicit, as the connection, in the first half of the text, between truth and hearing, and in the second half, between truth and sight). [2]

Bruno's text is a highly complex and polymorphous semantic space: every single word can be used within a few pages with a meaning that is, from time to time, poetical, figural, philosophical, etc. Moreover, *Eroici Furori* is a philosophical text: Bruno tries to redefine words and concepts he uses, but he achieves that not with definitions, axioms, etc., but through discussions and explanations across the whole text. The redefinition of the relevant words is one of the final achievements of every philosophical text, but this redefinition in Bruno is achieved through polysemy. Moreover, the dialogue form allows the author to leave some of the discussed points without a final definition, or ending in an "aporia", making polysemy not only as the way to achieve a result but as the result itself. A further challenge is represented by Bruno's language, a XVIth century variety of Italian – often intermixed with Latin words and sentences – that is quite different from standard contemporary Italian under the lexical and the grammatical point of view. This obviously represents a big obstacle for the application of *Natural Language Processing* (NLP) tools to texts analysis, since such tools are usually fitted to deal with contemporary language.

The aim of our research is to extract information from Bruno's text beyond standard lists of collocations or concordances, with the purpose of tracing the polysemy of relevant words and drawing the semantic spaces emerging from Bruno's text. In fact, our goal is to investigate the way a concept evolves throughout the text, the way it is used, and the metaphors it is connected to. This the reason why we do not want and we can not resort to a fixed and *a priori* determined ontology of concepts. No lexical ontology can in fact hope to keep track of the full conceptual dynamics of Bruno's lexicon, both within a single text and across different works. For instance, in the *Eroici furori* words like *giogo* "yoke, *catene* "chains", *cattività* "slavery" can change their meaning (and their gnoseological and ethic value) within a few pages span. In lexical ontologies, typically word meanings are never shaped or changed depending on the context of usage. Conversely, there is a radically different approach to word meaning in which a word information content is assumed to be inherently rooted in its *contexts of use*. In this model, hinging on the so-called *distributional hypothesis* of lexical meaning, an alternative view of semantic representations emerges, that allows us to approximate the idea of a *context-sensitive lexicon* as a way to account for semantic changes and meanings shifts.

3. Exploring conceptual spaces through word distributions

Since Harris (1968), distributional information about words in context has been taken to play an important role in explaining several aspects of the human language ability. The role of distributional information in developing representations of word meaning is now widely acknowledged in the literature. The distributional hypothesis has been used to explain various aspects of human language processing, such as lexical priming (Lund *et al.*, 1995), synonym selection (Landauer and Dumais, 1997), retrieval in analogical and judgments of semantic similarity. It has also been employed for a wide range of natural language processing tasks, including word sense and syntactic disambiguation, document classification, identification of translation equivalents, information retrieval, automatic thesaurus/ontology construction and language modeling (see Manning and Schütze, 1999 for a comprehensive overview), and has been taken to play a fundamental role in language acquisition.

According to the distributional hypothesis, the meaning of a word is thought to be based, at least in part, on usage of the word in concrete linguistic contexts. If this is the case, it should then be possible, in principle, to automatically acquire the meaning properties reflecting the distributional behaviour of a word, by inspecting a sufficiently large number of its contexts of use. The set of context-sensitive properties provides us with a text-based characterisation of the possible meaning facets of a lexical unit. This way, it is possible to model the conceptual content of a word as a semantic space emerging from its modes of combining with other words in a text. Word co-occurrence in a certain textual environment can thus be taken as key phenomenological evidence to reconstruct its semantic properties. Actually, the distributional paradigm represents a thoroughly new mode of accessing and exploring texts with computational methods. These can be used to keep track of the combinatorial properties of words and to draw from them a "*cartography*" of text-driven semantic spaces. In our work, we have adopted the distributional paradigm to carry out a computational semantic analysis of Bruno's *Eroici furori*, with the specific aim to build semantic maps of Bruno's lexicon, which are automatically derived from word distributions in the text.

Various computational models implementing the distributional hypothesis currently exist, most notably *Latent Semantic Analysis* (LSA; Landauer and Dumais 1997), *Hyperspace Analogue to Language* (HAL; Lund *et al.* 1995), and more recently *Correlated Occurrence Analogue to Lexical Semantics* (COALS; Rohde *et al.* in press). These different instantiations of the distributional paradigm all share the fundamental assumption that

2 We will come back to this metaphor later on: in this case, in fact, LSA has proven to have a real heuristic value, creating a virtuous hermeneutic circle.

measuring the semantic similarity between any two words is equivalent to measuring the degree of overlapping between their sets of linguistic contexts. This is to say that *two words that tend to be selected by similar linguistic contexts are closer in the semantic similarity space than two words that are not as distributionally similar*. Distributional semantic models typically formalize this assumption as a computational system formed by the following three components:

vector-based representation of words – a word w_i is represented as a *n-dimensional vector*. Each dimension records the number of times that w_i occurs in a given context. Actually, models may differ for the definition of context they adopt. For instance, in LSA each document in a given collection represents a specific context, and consequently each dimension of a word vector records the frequency of that word in a certain document. Conversely, in HAL, vectors records *word-co-occurrence* in a given context: a word w_i is represented as a vector in which each dimension d_{ij} records the number of times w_i occurs within a window of n words before or after a certain word w_j, where n is an empirically fixed parameter.

vector dimension reduction – typically, distributional word vectors have a very high number of dimensions, generally, in the order of thousands or tens of thousands. This may depend on the size of the collection used to build the word vectors, or in the case of word-co-occurrence vectors on the size of the text vocabulary (i.e. the number of distinct word types). Moreover, word vectors are usually very "sparse". This is obviously related to the fact that words – in particular those semantically richer – have highly selective combinatorial properties, and tend to co-occur with a restricted set of other words (or in a restricted subset of documents). Since the high number of zeros in a vector may negatively impact on their comparison, thereby altering the actual similarity relations, distributional methods typically apply some form of vector dimension reduction, in order to compress the number of dimensions of word vectors, while preserving their statistic distribution. For instance, LSA reduces the number of word vector dimensions through *Single Value Decomposition* (SVD) (Landauer and Dumais 1997). Vector dimension reduction is motivated by the fact that not all the contextual combination of a word in a text are equally relevant to characterize its semantic properties. Reduced vectors are then meant to extract those contextual dimensional which are truly constitutive of a word semantic space. If the original vector dimensions simply record the global distributional history of a word in a text, after applying SVD words are represented as points in a vector space whose reduced dimensions can be interpreted as a sort of "latent" semantic axes, which are implicit and hidden into the original word contextual distribution.

vector comparison – According to the distributional hypothesis, if two words w_i and w_j have close values for the same dimensions d_i and d_j in their vectors, we can regard w_i and w_j as semantically similar with respect to such dimensions. This in turn implies that to determine the global semantic similarity between two words it is necessary to compare their vectors with respect to the whole number of their dimensions. Intuitively, the greater the number of dimensions for which w_i and w_j present close values, the higher the similarity between them. This in turn presupposes the definition of a distance function $d(v_x, v_y)$, associating a scalar to any pair of target words. Similarity is then defined as an inverse function of d, whose values range between 0 (no similarity) and 1 (maximum similarity). One of the most common distance similarity measure between vectors is the *cosine*:

$$\cos(v_x, v_y) = \frac{v_x \bullet v_y}{\|v_x\|\|v_y\|} = \frac{\sum_{i=1}^{n} v_{x_i} v_{y_i}}{\sqrt{\sum_{i=1}^{n} v_{x_i}^2} \sqrt{\sum_{i=1}^{n} v_{y_i}^2}}$$

The cosine can be interpreted as a correlation coefficient, saying how similar the two vectors are. If the vectors are geometrically aligned on the same line, the angle between them is 0 and its cosine 1. Conversely, if the two vectors are independent, their angle is close to 90± and its cosine close to 0.

We can imagine distributionally defined lexical concepts arranged in what Gärdenfors (2000) calls *conceptual spaces*. The latter are defined by a certain set of quality dimensions that impose a topological structure on the stream of our experiential data. In Gärdenfors' classical example, the structure of the "color space" is defined by three dimensions: hue, brightness, and saturation. The meaning of each color term is then identifed with a three-dimensional vector of values locating its position on each of the space axes. The conceptual space model predicts that semantically similar color terms will then be located closer in the color space. In much the same way, the distributional hypothesis suggests that we can use n-dimensional frequency vectors to define a lexical semantic space derived by the average use of words in context. These represent classes of semantically similar words as *clouds* of n-dimension points, which occur close in the semantic space relatively to some prominent perspectives.

4. Building semantic spaces in Bruno

For the specific purposes of our research on Bruno's lexicon, we have adopted *InfoMap* (Widdows 2004), a recent variation of classical LSA. Differently from the latter, in InfoMap a word is represented as a vector in a *word-by-word co-occurrence matrix*: each vector dimension thus represents the number of times a given word co-occur with another word within a certain text span. InfoMap has instead adopted from LSA Single Value Decomposition as the vector dimension reduction algorithm. Finally, vector cosine is used to rank words with respect to their semantic similarity.

In a first experiment to acquire a more fine-grained and text-driven characterization of Bruno's semantic lexicon, InfoMap has been applied to the *Eroici furori*. A major parameter in distributional approaches to word meaning is the way text is represented. Usually, LSA is applied to tokenized documents, without any previous step of morphological normalization (e.g. stemming or

verità	1.000000	**infinito**	1.000000	**intelletto**	1.000000
apolline	0.622923	*infinitamente*	0.834759	*intelligibile*	0.773289
ingiuria	0.622744	*finito*	0.826582	*sensibile*	0.730790
veritade	0.618802	*infinita*	0.738326	*moltivario*	0.671390
solitudine	0.603709	positivo	0.736982	*moltiforme*	0.662998
inaccessibile	0.594529	*potenza*	0.655363	almanco:	0.608959
inobiettabile	0.593932	*interminato*	0.638099	*volontade*:	0.606160
infallibile	0.590822	accio	0.636445	esercitare	0.590920
visione	0.589427	rebecchina	0.625724	*consegliare*	0.570937
empedocle	0.576162	*finitamente*	0.621887	rovere	0.568231
sopranaturale	0.573116	*perfettivo*	0.617568	*potenza*	0.561274
empire	0.563613	vano	0.615645	materiale	0.559904
misura	0.563482	affirmazione	0.593107	applicazione	0.556474
bontà	0.558930	pario	0.592869	locale	0.544754
deserto	0.553126	circolazione	0.579869	*universale*	0.537373

Table 1 – the words with the highest cosine wrt *verità* "truth", *infinito* "infinity" and *intelletto* "intellect"

lemmatization). According to this approach, two morphological forms of the same word (e.g. *furore* and *furori*) receive two distinct and independent vector representations. Similarly, word forms act as distinct dimensions in the word-by-word co-occurrence matrix that is used to build the distributional vectors. This type of approach has the clear advantage of simplicity, since no linguistic analysis of the text is required beyond tokenization. Conversely, the price to pay for this choice is the loss of important semantic generalizations. In fact, typically different morphological forms of a word share the same semantic properties. Since these are extracted from word co-occurrences in context, the distribution of the different morphological forms of a word should rather be used to compute a unique word vector associated with their abstract lemma. Therefore, semantic similarity spaces should be built from the comparison of "lemma vectors" rather than "word form vectors". Obviously, this problem is even more urgent in the case of a highly inflected language like Italian.

In order to overcome the limits deriving from applying distributional methods only to the "raw" text, in our experiments we have used a more abstract and linguistically grounded representation of Bruno's work. In fact, we applied InfoMap to a lemmatized version of the *Eroici furori*. The original text has been processed with *Italian NLP*, an existing tool for morphological analysis and PoS tagging of contemporary Italian (Bartolini *et al.* 2004). The output has then been manually revised to cope with processing failures. These were mostly due to gaps in the morphological lexicon used by *Italian NLP*, deriving from orthographical variations or lexical archaisms typical of the XVI century Italian of the *Eroici furori*. The lemmatized text has then been used to build word semantic spaces with InfoMap. In our experiments, we empirically fixed the context window to 30 words to the left and 30 words to the right of the target word. It is worth remarking that the choice of the context window size may greatly affect that type of distributionally derived semantic spaces. This parameter strongly interacts with the type of semantic associations that can be extracted. A too narrow window leads to loss of potentially relevant correlations between words, whereas a too large window may compute irrelevant correlations. Determining the proper size of the context window is always an empirical problem and greatly depends on the type of the text and on the goals of the analysis, i.e. on the type of semantic associations that we are looking for.

In order to evaluate the ability of InfoMap to identify proper semantic similarity spaces from the *Eroici furori*, a team of philosophy historians[3] selected a set of words that have a key role in Bruno's lexicon. These words belong to four major categories, which so-to-speak represent a portion of the "top-ontology" of Bruno's thought: *Amore e Libertà* (Love and Freedom), *Amore e Facoltà* (Love and Faculties), *Intelletto* (Intellect), *Vicissitudine* (Vicissitude). For each word w_i (henceforth *target word*), we have queried InfoMap to obtain the list of the words w_j with the highest cosine value wrt to w_i. The results have been manually inspected to evaluate whether the returned words actually included items semantically related to the target word. A sample of the results is reported in Table 1, which illustrates the words with the highest cosine value (over 0.5) wrt to *verità* "truth", *infinito* "infinity" and *intelletto* "intellect". The words that have true semantic association with the target word are underlined and in italic. As can also be judged from these examples, results evaluation has shown that the cosine values computed by InfoMap nicely correlate with semantic associations that are highly relevant with the context of Bruno's thought. For instance, among the closest words to *intelletto* "intellect" we find *intellegibile* "intelligible", *moltivario* "multifarious", *multiforme* "multiform", *volontade*, "will", *consegliare* "to advise", etc. In the overall, the resulting semantic space is highly composite, but it surely includes words - either nouns, adjectives or verbs - that contribute to characterize the conceptual space expressed by *intelletto*. Similar considerations hold for *verità* "truth", in

3 The group is formed by Simonetta Bassi, Francesca Dell'Omodarme, Fabrizio Meroi, Olivia Catanorchi, Elisa Fantechi, Daniele Esposito.

whose semantic neighbourhood we find a near-antonym like *ingiuria* "insult", an orthographical variant like *veritade* "truth", but also a set of adjectives - *inaccessibile* "inaccessible", *inobiettabile* "indubitable, *infallibile* "infallible" and *sopranaturale* "supernatural" - that nicely characterize the Brunian concept of truth. The results reported in Table 1 have been obtained by training InfoMap on the lemmatized text. We can also evaluate the effect of text linguistic analysis by comparing the output produced by InfoMap when trained onto the tokenized text. Table 2 reports the words with a cosine value higher than 0.5 wrt the target word *caccia* "hunting": the left column shows the results produced by a distributional model obtained by training InfoMap on the tokenized text, while the results in the right column have been obtained by training InfoMap on the lemmatized version of the *Eroici furori*. As we claimed in section 2, *caccia* has a highly symbolic role in Bruno's figurative language, with particular regard to the myth of Atteon. First of all, we can observe that the lemmatized model is able to identify a higher number of words that actually belong to the semantic neighbourhood defined by *caccia*: e.g. *cacciatore* "hunter", *selva* "wood", *gemito* "cry", *venazione* "hunting". Moreover, the lemmatized model also assigns a fairly high cosine value to two verbs - *predare* "to prey upon" and *abbattere* "to kill" – that are strongly related to hunting events. Notice that these same verbs are instead missing from the words identified by the tokenized model. This also confirms the fact that the use of linguistically analyzed texts can decisively improve the accuracy of distributionally carved semantic spaces. Due to the different inflected forms in which a verb can appear in the text, only the use of "lemma vectors" can grant us with the possibility to draw a more complete semantic "cartogaphy" of lexical spaces, representing not only the associations between different entities in a domain, but also the prototypical events and actions involving such entities.

tokenized model		*lemmatized model*	
caccia	1.000000	**caccia**	1.000000
cacciator	0.715892	*cacciatore*	0.772613
venazione	0.699857	*selva*	0.640169
cacciatore	0.645397	*gemito*	0.629727
preda	0.603427	inenarrabile	0.629727
inebriato	0.593269	*predare*	0.625944
venaggione	0.570792	*venazione*	0.612536
corporal	0.568281	*abbattere*	0.612242
gesti	0.564450	*sapienza*	0.545181
avvegna	0.561344	allargare	0.533222
cattivare	0.560402	penetro	0.528437
vergini	0.558592	comprensibile	0.523042
apelle	0.554898	pelle	0.522736
approvare	0.554297	marcire	0.522024
aggrade	0.541634	intisichire	0.520585

Table 2 – words to which InfoMap has assigned highest cosine values wrt *caccia*

A further observation concerns the presence of the word *sapienza* "wisdom", among the words that the lemmatized model has identified as closely associated with *caccia*. Although *prima facie* this association does not seem to be correct, it is actually a very interesting result within the specific context of Bruno's work. In fact, Bruno explicitly claims that Atteon is the representation of the human intellect that aims at reaching wisdom (*Eroici furori*, part I, dialogue IV). Therefore, within Bruno's concept space, actually *sapienza* belongs to the semantic space centred on *caccia*, exactly because knowledge and wisdom are what Atteon's hunting aims at. This also allows us to do a more general remark on the use of semantic distributional methods, and on the shape of the semantic spaces individuated by InfoMap. In fact, the words that are singled out with the highest cosine values wrt the target word actually belong to a semantic similarity space that appears to be highly various and multifaceted. If we limit ourselves to the semantic neighbourhood of *caccia*, we can find true synonyms or near-synonyms - *venazione* -, words identifying typical participants to hunting events and that are also morphologically related to *caccia –cacciatore* "hunter" -, names that denotes a typical place for hunting –– *selva* "wood" -, events related to hunting - *predare* "to prey upon", *abbattere* "to kill" –, up to words like *sapienza* which have a metaphorical and allegorical link with the target word only within the specific setting of the *Eroici furori*. This range of words represents a fairly prototypical example of the shape of semantic spaces extracted from texts by distributional models. Such spaces can in fact simultaneously include words linked to the target word by semantic relations as different as synonymy, hyperonymy, antonymy up to metaphor and allegory. Distributionally designed semantic spaces are therefore fairly distant from the structure of traditional symbolic ontologies or lexical taxonomies. On the other hand, such spaces seem to be able to reproduce in a more precise way the inherent multidimensional and protean nature of the lexicon. A term like *caccia* is actually the centre of a network of multifarious semantic associations, some of which are highly text-specific and determine the particular shade of meaning that this lexeme acquires in the *Eroici furori*.

A further example can show us the heuristic value that distributional computational models might have for the research of philosophy historians. As shown in Table 3, the words most closely associated with *causa* "cause" by InfoMap trained on the whole text (left column), are significantly different from those produced by InfoMap once trained on the first half of the *Eroici furori*. According to the first model, among the closest words to *causa* we find *ciechi* "blind men", *vedere,* "to see", *occolte,* "hidden", and *visiva,* "visual". These types of association can be easily explained because in the second half of the *Eroici furori* the metaphor of the blindness and of its causes is the key metaphor to understand the relation between the *furioso* and knowledge, between man and god, between man and the endless universe, etc. The results of the model trained on the first part are instead more hard to interpret: the first word after *causa* is *sordo,* "deaf man", and there are other words that are related to hearing: *ubedire* "to obey" and *armonico*, "harmonic". Thus, we find a significant difference in the metaphorical space used to define the same word in the first half and in the whole text. Actually this contrast may represent an

interesting clue for philosophy historians. In fact, in the second half of the *Eroici furori* we find a platonic conception of knowledge. On the other hand, in the first half of the text there is the search for (i.e. the hunting) of this upper level and a great use of quotations taken from the Ancient Testament, in which the connection with God is based not on sight but on hearing (Scholem 1998). Although this interpretation surely needs more careful investigations, we believe it to be a good example of the interesting synergies deriving from the cooperation between computational methods and traditional philosophical analysis.

whole text		*first half*	
causa	1.000000	**causa**	1.000000
ciechi	0.674393	sordo	0.537429
attitudine	0.588798	quasi	0.505394
inclinazione	0.586934	umiliare	0.505080
parlava	0.583797	nativita	0.492546
ultimamente	0.581050	ubedire	0.491164
circe	0.580207	nullamente	0.489278
appresa	0.578190	contrasto	0.480551
parturita	0.575564	ripugnanza	0.478364
vedere	0.572744	provocare	0.474981
atto	0.572678	antitesi	0.472766
occolte	0.569679	convenienza	0.472755
visiva	0.564488	incanto	0.467928
consultava	0.557463	armonico	0.467667
incognite	0.555445	prossimo	0.452789
reprimere	0.544914	dannare	0.452688
feccia	0.544379	cecita	0.447493
persuadersi	0.541207	diviso	0.447167
potenza	0.539055	variamente	0.446063
afflige	0.536174	conclusione	0.445437

Tabella 3 - words to which InfoMap has assigned highest cosine values wrt *causa*

5. Semantic maps

Context-vector representations of words can be inspected with the aid of computational methods that group semantically similar words. *Clustering*, *Principal Component Analysis* and *Multidimensional Scaling* are just some of the many techniques that can be used to draw topological pictures of the semantic similarity spaces of words, as determined by their distributional properties. In our research we have used *Self-Organizing Maps* (SOMs; Kohonen 2001) to build semantic maps of Bruno's lexicon, thereby evaluating the real ability of a distributional model like InfoMap to make semantic regularities emerge out of text.

SOMs are unsupervised neural networks in which learning is achieved entirely by the system's self-organization in response to the input. In their simplest form, SOMs consist of two layers: an input and an output layer. In our work, the input consisted of the 100-dimensional word vectors produced by InfoMap through SVD. The output layer is represented by a two-dimensional topological map, where each processing unit (*neuron*) is a location on the map that can uniquely represent one or several input patterns. Before training the weights on each output unit were set randomly. At the end of the training regimen, the output layer presented a topographic organization which developed on the basis of regularities among the input word vectors. The training, which is unsupervised, consisted of presenting the word vector in random order. Self-organization in the system arises through a ''winner takes all'' mechanism: the output unit with the largest input wins and inhibits all other active units through lateral connections. For the winning unit, the weight vector is changed toward the input pattern so that the same unit is more likely to respond to the same pattern in the future. A neighbourhood function ensures that not only is the weight of the single winning unit adjusted, but so are the weights of the neighbouring units. In the output layer, the representation of each word was taken to be the unit that responded maximally to a given vector by the end of the training regimen. As a result of the self-organizing process, the statistical structures implicit in the high-dimensional space of the input are extracted as topological structures and represented on a two dimensional space.

If the context-vectors produced by InfoMap actually encode "latent" semantic properties of a word, then we can expect that semantically similar words will be mapped closer on the SOM than less similar ones. This means that we can use SOMs as a probe to investigate the quality of the text-driven semantic spaces built through InfoMap. In the current research we have trained SOMs on the four sets of words we mentioned in section 4. The network output is then intended to represent a distributionally derived *semantic map* of a particular conceptual area of Bruno's lexicon. Figure 1 reports the SOM generated from the context vectors produced by InfoMap for the set of words belonging to the category *Amore e Facoltà*. It is interesting to observe how the SOM has been able to derive semantically coherent word clusters. For instance, close to the left border there is an area occupied by words referring to different facets of the faculty of will: *voluntade* "will", *concupiscenza* "desire", *affetto* "affect", *appetito* "appetite". The central area of the map has instead been colonized by words referring to other typs of faculties mostly referring to senses: *immaginazione* "imagination", *senso* "sense", and *fantasia* "fantasia". Interestingly, *intelletto* "intellect" and *intenzione*

concupiscenza			**vizio**				**virtu**
affetto				**senso**			**fuoco**
appetito		**intelletto**					**acqua**
				imaginazione			
voluntade							**vita**
		intenzione					**morte**
memoria			**fantasia**		**amore**		**speranza**

Figure 1 – Semantic map for the conceptual category *Amore e Facoltà*

"intention" has been mapped half-way between the "will-area" and the "sense-area", prompting suggestive hypotheses about the particular interpretations that these concepts may acquire in Bruno's philosophy. Moving to the right-hand side of the map, we can notice the close position of two near-antonyms like *vita* "life" and *morte* "death", similarly to the case of *fuoco* "fire" and *acqua* "water". Finally, it is worth mentioning the suggestive spatial proximity of the two words *amore* "love" and *speranza* "hope", the former in turn appearing close to *fantasia*. All in all then, this shows that actually computational distributional methods are able to bootstrap interesting semantic spaces from the text. Besides, SOMs and similar data-analysis techniques provide a powerful tool to inspect text derived conceptual spaces, thereby offering new perspectives and potentialities for philosophical investigations.

cattivita			giogo				**giustizia**
							nodo
			catena				
desio							
							laccio
speranza			**felicita**				**liberta**

Figure 2 – Semantic map for the concepual category *Amore e Libertà*

As a further experiment we have manually subdivided the above word categories into semantically coherent sub-categories. For instance, the category of *Amore e Libertà* has been subdivided into two conceptual clusters, that are particularly relevant within the context of the *Eroici furori*: 1. *words expressing positive concepts*, i.e. *libertà* "freedom", *speranza* "hope", *desio* "desire, and *giustizia* "justice"; 2. *words expressing negative concepts*, i.e. *catena* "chain", *nodo* "knot", *laccio* "string", *cattività* "captivity", *giogo* "yoke". The InfoMap vectors corresponding to these two word sets were given as input to the SOMs, to evaluate to what extent the resulting semantic map could reproduce the above partition. The output SOM has been reported in Figure 2. It is worth noticing that the words belonging to sub-category 1. (represented in bold face) are located in the bottom area of the map, with the notable exception of *giustizia* "justice"; the whole top-right part is instead occupied by the words expressing negative concepts. Moreover, even in this case, highly semantically related words are located in closer positions in the SOM: this is for instance the case of *nodo* "knot" and *laccio* "string", as well as of *cattività* "captivity" and *giogo* "yoke", the latter being typically used as a metaphor for oppression and humiliation. Therefore, the distributionally built semantic spaces are able to approximate relevant partitions of a given lexical area. Obviously, semantic maps are not noise-free and word clusters do not perfectly coincide with natural semantic categories, like the ones that could be found in standard linguistic ontologies or thesauri, such as for instance WordNet (Fellbaum 1998). Although this might appear as a limit of distributional models, actually it should be regarded as a direct consequence of the inherent fuzziness of lexical concepts. This appears to be even more true once we stop considering word meanings as immutable and abstract entities, and we observe how they are continuously reshaped and moulded in context.

Figure 1 – Semantic maps built from the first part (left) and from the second part (right) of the *Eroici furori*

Since distributional models build word representations that are inherently context-sensitive, they can also be applied to keep track of semantic dynamics, that is to say of how semantic spaces change along the "time dimension" defined by the narrative flux in the text. This is a particularly important point, especially within the context of philosophical analysis, since it is often the case that the same word may acquire different semantic shades in different parts of the same text. In order to test the potentiality of the distributional approach with respect to this issue we trained two models with InfoMap, using respectively the first and the second part of the *Eroici furori*. Then, for a set of selected words, we trained two SOMs using as input respectively the context vectors produced by the two models of InfoMap. The output SOMs are reported in Figure 3. Notice that in the SOM derived from the first half of the text (left map), *Diana* is located in the same area of *furore* "fury", *furioso* "furious", and *beltà* "beauty". *Atteone* is instead near *voluntade* "will" and *preda* "hunted". Actually, in the first half of the text, Atteon - the hunter - becomes the hunted after he sees Diana naked (the beauty). The experience of Atteon is the same as the one of the wise who approaches the truth and the beauty. In this experience, will is really important, but not as much as the fury, i.e. the transformation of the lover in the beloved object. In the second half of the text (*Eroici furori*, II, 2), the name of Atteon is used only twice. In this part Bruno explains a theory of knowledge similar to the platonic one: the lover stares at the ideas and loves the chains of love that made him a slave. Interestingly, in the right map *Dana* is now significantly distant from *furore* and in the same area of *unità* "unity" and *destino* "fate". *Atteone* is more distant from the word *preda* "hunted", and in the same area of *Diana* and unità. This actually finds a nice correspondence in the second half of the *Eroici furori*, in which the hunter has gone beyond the experience of rage and can only receive the "revelation" of Diana, and the possibility of becoming the hunted is part of his fate.

6. Conclusions and future developments

In this paper we have argued that semantic distributional models can offer new interesting perspectives for philosophical text analysis. The reported results actually confirm the heuristic value and the potentialities of these types of computational methods that allow us to acquire a really dynamical view of the text, monitoring the changes in meanings of single words appearing in different parts of the *Eroici furori*. Further developments of our research involve extending the analysis to other Bruno's texts such as for instance to the *Cabala del Cavallo Pegaseo*, and thus performing a comparative study of the word similarity spaces corresponding to different phases of the philosopher's production. Moreover, we are also planning to apply InfoMap methodology to Bruno's Latin works.

In the overall, we believe that our work is an interesting example of the possible and fruitful synergies that can be obtained by combining standard philosophical studies with more advanced computational linguistics techniques. Methods and tools for NLP and statistical text analysis that are widely used in applicative contexts such as ontology learning or document indexing, can be profitably adapted to cope with the challenges of philosophical text analysis.

7. References

Bartolini, R., Lenci, A., Montemagni, S., Pirrelli V. (2004), "Hybrid Constraints for Robust Parsing: First Experiments and Evaluation", in *Proceedings of LREC 2004*, Lisbona, Portogallo: 859-862.

Fellbaum, C. (ed.) (1998), *WordNet. An Electronic Lexical Database*, Cambridge, The MIT Press.

Gärdenfors, P. (2000), *Conceptual Spaces. The Geometry of Thought*, Cambridge, The MIT Press

Kohonen, T. (2001), *The Self-Organizing Maps*, (3rd ed.). Berlin, Springer.

Landauer T. K., Dumais S. T. (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge", Psychological Review, 104, pp. 211-240.

Lund, K., Burgess, C., Atchley, R.A. (1995), "Semantic and associative priming in high-dimensional semantic space", in *Proceedings of the Cognitive Science Society*, Hillsdale, N.J., Erlbaum Publishers: 660-665.

Manning, C.D. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MA, MIT Press.

Rohde, D. L. T., Gonnerman, L., Plaut, D. C. (in press), "An improved model of semantic similarity based on lexical co-Occurrence", *Cognitive Science*.

Scholem, G. (1998), *Il nome di Dio e la teoria cabbalistica del linguaggio*, Adelphi, Milano.

Widdows, D. (2004), *The Geometry of Meaning*, Stanford, CSLI.

A geometrical approach to literary text analysis

Roberto Basili, Paolo Marocco

University of Rome Tor Vergata,
Department of Computer Science, Systems and Production,
00133 Roma (Italy),
{basili,marocco}@info.uniroma2.it

Abstract

According to (Eco, 1960), texts are *open* as they imply a global understanding by the reader's community. Readers' task is "*to fill the interstices of the text*". In this paper we propose a model to capture the implicit semantics used by readers to achieve this goal. A geometric paradigm is applied and a structured corpus-based approach based on a target novel plus complementary text (i.e. critical reviews on the novel) is defined. The inductive nature of the model and its unsupervised charactr suggest its application as an advanced tool for literary analysis. Experimental evidence has been acquired via its extensive application to the novel "*Gli Indifferenti*" by Alberto Moravia (Moravia, 1929). Results are more than promising and confirm the general feasibility of the aproach to literary anaylsis.

1. Introduction

Models of textual cooperation have been proposed about forty years ago by Umberto Eco that in *Opera Aperta* (Eco, 1960) argued that texts are *open* in the sense that they imply a global potential understanding of the reader's community, whose task is "*to fill the interstices of the text*". The readers are thus seen as active subjects between the mind and the distributed social notion of meaning, i.e. the society. Readers reproduce the social encyclopedic knowledge while interpreting the text in their daily lives and communicating experiences. We will try to put forward these ideas in a computational perspective, by relying on a dynamic modeling of the text meaning and from an inductive perspective rooted in the machine learning tradition.

In this paper, we present a framework for modeling lexical semantics via a geometrical approach then suggesting a dynamic acquisition process from a literary work. The emerging textual semantic correlations are here analysed and aspects related to the full automation of the outlined model are discussed. Our perspective focuses not just on narrative aspects as observed in the text, but also on a paradigmatic level of sense description: the aim is to simulate the behavior of an ideal group of readers, that suggest and converge on a specific set of relations between topical concepts as a result of a latent cooperation phenomenon. For example, in the sentence "*He took a walk along the bank of the river*", a reader finds a combination of words *bank* and *river* at the textual level, but he also accesses at the paradigmatic knowledge needed to associate *river* whith absent terms, like *sea, hill, water*: any textual interpretation of the source sentence always involves both levels.

2. LSA and Semantic domains

Latent Semantic Analysis (LSA) has been proposed in early nineties as a geometrical model for the representation of textual semantics (Landauer and Dumais, 1997) . The general claim is that meanings can be captured by exploiting the relationships between their lexical realizations (word) and the targeted contexts (e.g. a paragraph or an entire document) . A mathematical notion (i.e. the principal component) is here employed to the occurrences (or frequencies) of words in texts aiming to determine a new space where semantic phenomena (e.g. *similarity*, *antinomy* as well as *textual relevance*) are better expressed.

2.1. A brief digression on Latent Semantic Analysis

In traditional models for information retrieval, a geometrical "space" is defined such that each dimension of the space is a term (i.e. a word in general) occurring in a goven collection of documents collection. Each document is represented in the space as a vector with a coordinate for each one of its terms: the value of each coordinate (or term) is a *weight* intended as a measure of how important is the term in the target document.

While this approach is an effective first approximation of the statistical properties of the collection, it is nevertheless an oversimplification. Its major limitation is that it assumes that terms are independent, orthogonal dimensions of the document space. Relationships among terms, e.g., the fact that certain terms are likely to co-occur in documents about a given topic because they all refer to aspects of that topic, are ignored. Similarly (and more subtly), the traditional term vector approach ignores the fact that a term A and a term B may occur in similar contexts in two distinct documents because they are synonyms. Finally, traditional vector space models are characterized by a very large number of dimensions, as many as the large number of terms appearing in the collection.

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) approaches the above problems by capturing term-term statistical relationships and cluster together terms expressing similar information. In LSA, the document space is replaced by a lower dimensional document space called k-space (or LSA space) in which each dimension is a derived concept, a "conceptual index," called an LSA "factor" or "feature."

While term-based vector space models assume term independence, natural languages are characterized by strong associations between terms, so that this assumption is never satisfied (Hull, 1994). LSA attempts to capture the semantic term dependencies using a purely automatic method, a matrix decomposition process called Singular Value Decomposition (SVD). The original term-by-document matrix M describing the traditional term-based document space is

transformed in the product of three new matrices: T, S, and D such that their product $TSD^T = M$. They capture the same statistical information than M in a new k-dimensional space where each dimension represents one of the derived LSA features (or concepts). These may be thought of as artificial concepts and represent emergin meaning components from many different words and documents (Deerwester et al., 1990).
D is the document matrix. Each of the D's columns is one of the k derived concepts. Rows in D represents documents in the new space, i.e. in terms of the k concepts. Similarly, T is the term matrix whose columns are the k derived concepts. Rows in T are vectors in the k-space describing a term of the original collection. Terms in this matrix are then characterized by vectors of weights indicating their strength of association with the underlying concepts (Deerwester et al., 1990). In other words, each term vector (i.e., row) in T is a weighted average of the different meanings of the term (Hull, 1994).
S is a diagonal matrix whose non zero values (called "singular values") express the decreasing significance of the k LSA factors. User has the the control over how many dimensions should the k-space be done of. SVD decomposition algorithms guarantee that the factors are presented in order of their importance (as measured by the diagonal of S). Therefore, the least important factors can be easily neglected by truncating matrices T, S, and D. The first k columns are called the LSA factors (Hull, 1994). The usual number of dimensions k needed for effective performances obviously depends on the collection and the task. Experiments show that improvement starts at about 10 or 20 dimensions, with peaks between 70 and 100, and then the impact decreases (Berry et al., 1995). Other reports say that the optimum number of dimensions is usually between 100 and 200.
LSA maps documents from a vector space representation to new space with a low number of dimensions. Terms are also mapped into vectors in the reduced space. Following the usual vector space similarity models, e.g., calculating cosine similarities in the new space, we can evaluate term-by-term, document-by-document, but also term-by-document similarity. LSA has several interesting implications:

- First, similarity between vectors in the reduced space is better than the similarity measured in the original term space as it established at the semantic level rather than just at a lexical level. Two related documents may be quite similar even though they do not share any keywords. This occur, for example, when different (thus independent) words in the two target documents co-occur frequently in other documents.

- A duality principle holds so that words and documents lies in the same space. Lexical and textual properties (e.g. similarity) can be thus explored in parallel. Although each row of matrix T is called a "*term vector*", the phrase is used quite differently in the LSA terminology than in conventional vector space terminology. Conventionally a "term vector" is a vector in the document space describing a document in terms of weights assigned to each term for the given document. In LSA, both terms and documents are described in the LSA k-dimensional space.

- LSA does not pose any particular constraints on the type of preprocessing needed as it only relies on the SVD transformation. This means that it discovers in a fully unsupervised manner most of the underlying semantic correlation among terms and can be applied to any text. In the next section we will see how this properties opens interesting perspectives towards computational models of literary phenomena as emerging properties of literary and critical collections.

- Traditional learning algorithms do not work effectively when applied to the large dimensional vector spaces of document collections, due to insufficient training data and computational complexity restrictions. Therefore, the dimensionality reduction implied by Latent Semantic Anaysis is a viable approach as it supports low-dimensional linear combinations of orthogonal indexing variables, i.e. the k concepts.

2.2. LSA and literary texts

LSA represents a paradigm alternative to logical and metalinguistic approaches to meaning (e.g. predicative structures in generative linguistics or logical formalisms for ontology representation and reasoning). LSA pushes for an analytical and geometrical view on meaning within a Vector Space Model paradigm largely used in Information Retrieval. The concepts emerge from texts, as a consequence of a similarity metrics grounded on the relations among texts and lexical items. Figure 1 depicts an example of LSA-based space where regions express word clusters as emerging concepts: in the example, a context[1] of a word, *bank*, is shown as a point in the LSA space. Its surrounding includes other lexicals like *river*, *hill* or *gate* that naturally trigger the proper "river bank" sense of $bank$ and characterize the micro-domain of the source sentence.

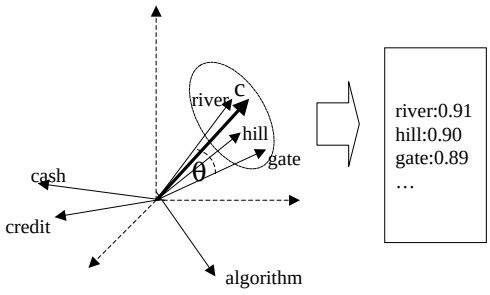

Figure 1: An example of geometric representation of semantically associated lexicals

Distance in the the latent semantic space gives rise to a natural notion of *semantic domain*(Gliozzo, 2005; Vigliocco and Vinson, 2005), fully expressed on a lexical basis. *Semantic Domains are clusters of terms and texts that exhibit*

[1] "*He took a walk along the bank of the river*"

a high level of lexical coherence. They are also characterized by sets of domain words, which often occur in texts about the corresponding domain" ((Gliozzo, 2005)).
The idea here pursued is to exploit the automatic acquisition of semantic domains as a form of intelligent support to the critical interpretation of narrative texts. It is worth to be noticed that the variety of evidence made available by geometric methods ranges across a number of information levels, each representing a specific knowledge dimension:

- an *encyclopedic* dimension at the highest. It captures the textual encyclopedic competence by expressing an *ideal social* knowledge about texts
- the *knowledge about the novelist* (the text's author) and *the novel*, in the middle. It expresses a more local competence related to the author derivable from LSA-based analysis of critical essays, interview, etc.
- the *knowledge related to target literary work* itself in the lowest level as the representation of the concepts determined only by the opera and the relations there emerging

A structured LSA-based analysis is a way to capture and integrate the different levels above. In our view this modularization better support the definition of dynamically emerging language phenomena in texts as earlier investigated in semiotics studies. Notice how while the firsts two levels can be seen as *paradigmatic*, the lowest level is *syntagmatic* [2]. The last one represents the plane of encyclopedic knowledge of the reader's community. In our case, we consider the level of critical essay as paradigmatic, and not syntagmatic, because it embodies a frame for meta-languistic and denotative word definitions. This model structure resumes some concepts of the *Opera aperta* previously described, in a newer machine learning perspective: in this sense, a semiotic vision of the cooperation among the ideal readers can be reflected and implemented.

3. A geometrical view on lexicalized narrative concepts

Narrative analysis is usually fed with the collocational evidence as it is found in the target texts. However, structured knowledge about a novel is not directly captured realized in atomic lexicalized phenomena. For example, when studying a work like "*Gli Indifferenti*" by Moravia (Moravia, 1929), the notion of $noia$ ($boredom$) is central to the analysis of some of the novel's characters. It is not straightforward to capture such structured notion only by means of simple atomic lexical information, i.e. words. No collocational analysis would be comprehensive without the effort of the critic to enumerate lexically this notion in order to detect proper subsections of the novel dealing with it. Notice that the word "*noia*" itself is not so frequently used in the novel (it appears just 18 times and is the 634-th words in the frequency ranking). Second most of its morphological variants (e.g. *annoiare* (*to bore*), *annoiato* (*bored*)) are not captured collocationally with *noia*. Third collocational analysis has no way to capture most of its topically associated words, like *esistenza* (*existence*), *avventura* (*adventure*), falsita' (*falsity*, *pretence*), ... An in depth knowledge of the novel itself is required to exploit collocational analysis and this is an inherent bias to the results: tautologically, reviewers just discover what they already know about the work.

3.1. Capturing social knowledge as LSA-based semantic domain

In an explorative perspective, a better model should characterize a target concept (like "*noia*") by means of structured representations able to support automatic matching in the underlying novel. The geometrical perspective opened by LSA is useful. Proximity in the LSA space can be considered a measure of the topical associativity among words and gives rise to a structured notion of *semantic domain*. The semantic domain of a word like *noia* can be modeled as the neighborhood of a specific point in the LSA space, properly related to the word *noia*. Fig. 1 suggests a lexicon related to a sentence: any textual information can be mapped into the LSA space and determine a region where specialized lexicons can be found. This implies a way to map a text (or simply one word) into a specalized lexicon. Let us define such a lexicon as a surrogate semantic domain generated by the underlying target notion, in our example "*noia*".
Notice how LSA can be run over any text so that it may give different responses: if run only on the novel it will suggest a lexicon only internal to the novel. In this case the semantic domain will be charaterized by the situations, people and places related to the occurrences of the word *noia* in the novel. However, this is not the only evidence available. A further body of textual evidence is represented by the critical reviews of the opera, i.e. the paradigmatic and social knowledege, the first and second levels, of Section 2.2.. When LSA is run against such an extended collection, it will reproduce the typical critical argumentation, situations and analogies involving the word noia as they can be detected in the extended corpus.
The model we propose to determine lexically a semantic domain for a notion of narrative interest (like *noia*) by relying on three textual entities:

- relevant words, that we call *cue words*
- critical texts associated to the opera that we call extended corpus, C_E
- the opera itself C characterized by paragraphs and chapters as *textual units* t

The LSA space model is build on the matrix words-paragraphs that insist on the extended corpus C_E or on the opera C, that will be referred LSA_E and LSA_C, respectively. It models the relationships between words and their contexts within different knowledge levels. It allows to capture a particular semantic domain of a cue word triggered by a specific text portion.

[2] "The opposition syntagmatic/paradigmatic proposed by Saussure, identifies two orthogonal axis, the first one involve the composition of the words in textual bodies (e.g. a text, a discourse), whilst the second refers to the body of relations which connect the originating text to other texts.

Now, in order to characterize a semantic notion, like *noia*, we have two possibilities: study its behaviour in the opera (i.e. over the corpus C) or associating to its discussion the bundle of social evidence also given by critical reviews of the opera itself (i.e. ove the corpus C_E). Notice how the first choice is tight to the author's view on the concept, that is the fact and the narrative evidence intrinsic to the work, this including people, events and locations, discussed in the text. This view can be partial as it does not capture the implicit role of readers that make reference to a wider evidence, i.e. their experience and knowledge of the world. The adoption of the extended corpus augment the generalization power of the system as it may refer to every situation (i.e. piece of textual evidence) available. This is more general and expressive of the overall semantics underlying the target narrative concept suggested by the cue word.
The adopted process can thus be formally expressed as follows. Given a cue word c and the extended corpus C_E:

- First, run LSA on to the C_E and make available the transformation matrices $TS^{1/2}$ and $S^{1/2}D^T$ that map term and document vector in the transformed LSA_E space.
- Map the cue word c in the LSA_E space, i.e. compute the vector $\vec{c}$ in the LSA space
- Select words w that are close enough in LSA_E to the cue word c, as the lexicon L_c characterizing the semantic domain $S(c)$. More precisely

$$L_c = \{w \in C_E \text{ such that } ||\vec{c} - \vec{w}|| < \tau\}^3 \quad (1)$$

  where $||.||$ is the cosine similarity distance in LSA_E and τ is a positive constant aiming to control the generalization trade-off, required not to introduce too much noise in the process.

Equation 1 defines mathematicaly the notion of neighborhood of a word, as depicted in Fig. 1, aiming to capture lexical cohesion among topically related words of a semantic domain. The result of the above process is a subset of the overall dictionary L_c (i.e. words belonging to the opera as well as words used in the critical reviews) that characterize the semantic domain of the cue word c. For example from the cue word *noia*, we obtained L_{noia} = {*esistenza*, *atto*, *avventura*, *fatalita'*, *falsita'*, *familiare*, *ragazza*, *abitudini*, ...}[4].

3.2. Capturing paradigmatic knowledge about semantic domains

A lexicon L_S is used here to characterize a semantic domanin c. It suggests a variety of concepts and relations underlying the target literary notion (e.g. *noia*). A way to explicit the underlying conceptualization may proceed by an interpretation of individual words and by the selection of their intended meaning (senses) locally to each L_c. Notice how this implies a form of word sense disambiguation (WSD). Unsupervised methods for WSD have been widely studied in the computational linguistics literature usually based on the sense repository called Wordnet ((Miller, 1990). A model introduced in (Basili et al., 2004) will be here used to derive the proposed senses and extract them from the Wordnet taxonomic hierarchy. It is based on an n-ary similarity estimation that, when applied to a set A paradigmatically related words, detect the subset of senses for words $w \in A$ that are mostly appropriate for A. This methodlogy can be here applied by assuming as the set A the lexicon L_c characterizing a semantic domain S. The result is a multi-rooted taxonomy that described the most impostant concepts (as senses higher in the hierarchy) that generalize all the words in L_c: the taxonomic relations are here retained so that the resulting semantic network is a general explanation of c. The method in fact detect the most specific generalizations in Wordnet able to cover all the words in L_c, thus providing an abstraction process that result in an explanation of c. In Fig. 2 the network of concepts activated by the semantic domain generated by the cue word *noia* is reported[5]. The presentation of Wordnet is (Miller, 1990), while technical details on the disambiguation process adopted here can be found in (Basili et al., 2004).

3.3. Studying the behaviour of a semantic domain

While semantic domains are captured from the extended corpus making use of the semantic distance established in the latent semantic space, a further type of analysis of each domain can be directly done against the opera itself. Notice how each textual unit of the opera is a sort of pseudo document and is also represented as a point into the LSA space. Again distance in this space can be assumed here as a narrative information. Semantic similarity (the dual notion of semantic closeness) suggests how much a textual unit t (e.g. a paragaph) is related to a semantic domain c, i.e. at what extent a critical analysis of the target opera should take t in consideration as an embodiment of the notion c.
Moreover, textual units are either individual paragraphs or entire chapters of the book. They are strictly ordered and give naturally rise to a syntagmatic view on the target narrative work. In this way similarity can be established not only locally but as a dynamic notion that proceeds across individual units and follows the narrative development. Notice that whenever a quantitative notion of similarity among a narrative concept and a paragraph (a text portion enclose between the beginning and the ending of the line) is available the narrative development can be expressed graphically as its function along the totally ordered set of units. A graphical expression of a complex semantic domain is thus achievable and can be made fully available operationally.
The above two aspects require from one side an expressive

[3] It should be noted here that threshold τ can be made dependent on individual words w, so that words more relevant for the corpus are given some preference. Technically in our experiments, $\tau_w = \frac{\tau}{ln(tf_w)}$ where tf_w is the term frquency of w.

[4] L_{noia} = {*existence*, *act*, *adventure*, *deceit*, *falsity*, *relative*, *girl*, *habits*, ...}

[5] The hierarchy is in English as the American Wordnet 1.7.1 has been used as a reference and the italian lexicon L_{noia} has been translated to trigger the conceptual density function: *fatality*, *act* and *existence* are, for example, English translations of *fatalita'*, *atto* and *esistenza*, respectively.

definition of a semantic distance (or dually of a similarity function). On the other side an additional model that sees the opera as a sequence of possibly structured units is needed. So, paragraphs will be assumed as atomic notions. Chapters are sequences of paragraphs so that similarity at the level of chapters is an aggregation function of the similarity function over individual paragraphs. Finally the entire opera can be seen as a sequence of chapters. The graphical methaphor can depict similarity along the linearly organized sequence of chapters, or along the sequence of paragraphs internal to a chapter. An analysis at different degrees of granularity is thus made possible.

The semantic distance function, adopted for this stage of the analysis, is defined as the cosine similarity within the LSA space generated over the opera. In this context, given a concept c, its lexicon as derived from the C_E corpus, and given a textual unit t of the original opera, i.e. $t \in C$, the semantic similarity between c and t is the cosine similarity among their vectors, as they are represented in the LSA space generated over the only opera, i.e. LSA_C. More precisely,

$$sim(c,t) = cos\_sim(\vec{c},\vec{t}) = \frac{\sum_i c_i t_i}{||\vec{c}||||\vec{t}||} \quad (2)$$

where $\vec{c} = \sum_{w \in L(c)} \vec{w}$, $\vec{t} = \sum_{w \in t} \vec{w}$ and c_i t_i are the i-th components of the vectors ($\vec{c}$, $\vec{t}$). $\vec{}$ is here always to be intended as the representation in the LSA_C space[6]. Notice how the lexical items in $L(c)$ are derived from an LSA-based analysis in the extened corpus C_E. Here their representation restricted to C is used, so that LSA_C is intended.

A quantitative representation of the narrative development of c in a text can be here obtained by a discrete function $f : \aleph \times T \rightarrow \mathcal{R}$, where T is the opera, i.e. the sequence of textual units t_i, and $\aleph$ is the abstract space of narrative concepts. f can be defined as follows:

$$f(t_i, c) = \frac{sim(c,t_i) - \mu}{\sigma} \quad (3)$$

where t_i represents the i-th unit of the opera, μ and σ are the mean and standard deviation values of the $sim(c,t_i)$ distribution, respectively. Here different distribution can be assumed with respect to the locality adopted. Different grains can be targeted so that the mean (or standard deviation) can be obtained over a chapter Ch_i (by averaging across paragraphs $t_j \subset Ch_i$) or over the entire opera T (i.e. by averaging across chapters $Ch_j \subset T$) The plot of the function $f()$ provides a graphical representation of the behaviour of the relevance of c across different groupings of textual units in the entire opera, i.e. paragraphs or chapters.

Given a chapter $Ch \subset T$, as a subsequence of length $n < N$ of the original $T = t_1, ..., t_N$, the overall semantic similarity between Ch and c requires an aggregation function Ψ the maps individual contibutions local to paragraphs into a global score. More precisely, given a narartive concept $c \in \aleph$ and a chapter Ch, the similarity function among the two is given by:

$$f(c, Ch) = \Psi_{t_i \in Ch}(f(t_i, c)) = \frac{1}{N} \sum_{t_i \in Ch} f(t_i, c) \quad (4)$$

Equation 4 expresses the aggregation as the standard mean value of the discrete distribution of values $f(t_i, c)$. Experimental evidence as acquired from the analysis of "*Gli Indifferenti*" di A. Moravia will be discussed in Secton 4.3..

4. Studying semantic domains within "*Gli Indifferenti*"

In order to validate and experiment the above defined model for narrative analysis we made a quantitaive study of the novel *Gli Indifferenti* by Alberto Moravia (Moravia, 1929). The book (corpus C) is made of about 16 chapters and about 91059 words (tokens). The number of different words in the novel is 3273. Additionally, we created the extended corpus C_E by including critical reviews up to a total size of 13041 tokens with 3920 different words. Individual pseudo documents have been created from the opera based on paragraphs: each pseudo document consists of a single paragraph in the opera. We found about 1854 total paragraphs and 116 paragraphs per chapter on average.

Different weighting schemes can be adopted for the assignment of initial values to the term-document matrices that triggers LSA. The score adopted in all the experiment discussed in here is the simple lexical frequency tf_{ij} of words w_i in in the pseudo-documents t_j. The dimensions used by the LSA on both corpora have been limited to (the first) 100 dimensions (i.e. principal components).

Finally, syntactic filters have been imposed on the lexicons so that discrimantion between verbs, nouns, adjectives and adverbs is possible. Also proper (e.g. *Leo*) and common nouns (e.g. *falsity*) are taken separate. Most of the following discussion has been centered around nouns as the analysis of the other categories is still in progress at the time of writing this paper.

4.1. Case I: extracting meaningful semantic domains

The generation of the semantic domains has been obtained through the notion of distance in the LSA_E space. In the different runs a threshold value (i.e. τ in Eq. 1) of 0.5 has been applied: in general about 35 different lemmas are obtained in the lexicons L_c, of which about 10 are nouns.

Although the system can be activated with every abstract concept as the originating *cue word*, we show in the following the evidence obtained around a set of meaningful cue words (c):

- *c=noia*. L_{noia} = { *esistenza, atto, avventura, fatalita', falsita', familiare, ragazza, abitudini, ...* }
- *c=indifferenza*. $L_{indifferenza}$ = { *vita, noia, scena, prova, vero, volonta', esistenza, proposito, ambiente, mancanza, incapacita', vanita'*}
- *c=Carla*: L_{Carla} = { *osservare, baciarono, torpore* }
- *c=Leo*: L_{Leo} = {: *suonare, ingiunse, stiro', cammina, fastidio, signor*}

[6] $\vec{w_i}$ is obtained by multiplying the i-th row of the original term-document matrix with the mapping matrix $TS^{1/2}$, derived according to the SVD transformation.

The resulting lexical descriptions are very interesting as a number of semantic phenomena are captured in a fully automatic way. First, words strongly correlated with the cue word are derived (e.g. *noia*/*boredom* vs. - *abitudini*/*habits*, *esistenza*/*existence*). Second, correlation at the level of the typical plot of the novel are also obtained, like the *noia-falsita'*/*falsity* pair. Notice here that the notion of *falsity* is a strong connotation of the typical middle class family described in the novel: it is a sort of originating state of the boredom itself.

Notice how the semantic domain built around the concept of *indifferenza* (*indifference*) includes *noia* and this pair is quite important as a narrative element, where the first is a sort of side effect of the first. On the other hand, LSA seems not to capture properly the semantics underlying proper nouns, as the L_{Carla} and L_{Leo} show. The distribution of proper noun seems not to be well modeled by our lexicalized approach. Reasons for this are diverse. First, proper nouns are often expressed by anaphorical references. As no method for anaphora resolution (or guessing) is applied, we suspect that most of the occurrences of proper nouns are not captured. Moreover, the novel characters are very few. They are rarely mentioned explicitly and mostly alluded. Due to this, LSA applied to the partial information does not converge to significant results so that a different treatment of proper nouns is required.

4.2. Case II: Explaining semantic domains paradigmatically through Wordnet

In order to analyze the internal structure of a semantic domain we tried to found a paradigmatic explanation of the derived L_c by using a reference semantic network, i.e. Wordnet, as discussed in Section 3.2.. We modeled the paradgmatic interpretation of a semantic domain c as a sense disambiguation problem local to the lexicalizations L_c obtained with respect to the extended corpus C_E.

Figure 2 reports the hierarchy of senses derived from the interpretation of the lexical representation L_{noia} of the semantic domain *noia*. Red boxes are the Wordnet topmost (maximally general) senses while leaves of the hierarchy are the originating words. Intermediate levels are shown when they are common generalizations of more than one term in L_c.

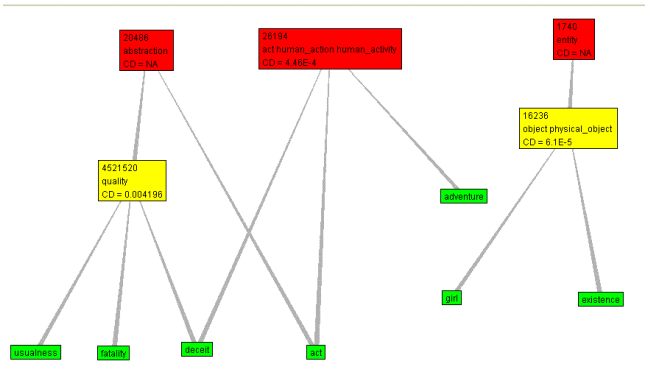

Figure 2: Wordnet network for: la *noia*

It can be sees here that the LSA analysis induces a context (the semantic domain L_c) where individual and ambiguous nouns (e.g. *atto*/*act* that can be an action or a legal document) are correctly mapped into the proper generalizations (i.e. *act, human act, human activity* and not an *object, physical object*). An example is *deceit* where the sense "*misrepresentation, deceit, deception*" and only the two meanings "*dissembling, dissimulation, deceit*" (abstraction) and "*fraudolence, deceit*" (quality) are retained.

As a further example Fig. 3 reports the ontological structure underlying the notion of *indifferenza*. Here basic concepts emerging are *states*, *psychological features* and *entities*, where basic entities are environments (e.g. words like *scena* (*setting*) or *ambiente* (*environment*)) where the feelings manifest. Notice how the word *environment* is also disambiguated correctly: it only preserves its "*location*" sense "*environment, environs, surroundings, surround*" and looses its "*state*" sense (first, i.e.more general, sense in Wordnet) .

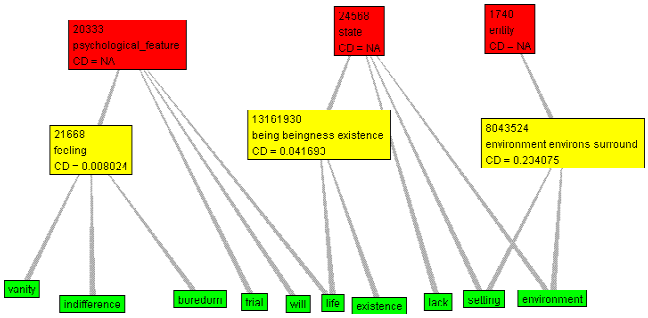

Figure 3: Wordnet network for the semantic domain: *indifferenza*

4.3. Analysing the development of a semantic domains in the opera

In the study of the dynamics of a semantic domain in the novel, we evaluated the relevance of the L_c derived in the first phase from the extended corpus C_E across the novel's portions. Individual chapters and paragraphs have been studied. Fig. 4 plots the behaviour of cue word $noia$ through the relevance (Eq. 4) of its lexicon L_{noia} across the 16 chapters. Individual words $w \in L_{noia}$ are also shown.

First it should be noticed that lexical members of the semantic domain have in general a common behaviour. All the words (contributing to the semantic domain) tend to have analogous strong or weak relevance in individual chapters. This indirectly suggests that their latent semantics has a general validity even though it has been originated in a different corpus, C_E. Second, some chapters are more correlated with the intended narrative concept (*noia*) than others. Chapter 13 for example shows a relatively low correlation at the level of the entire semantic domain but a very high standard deviation: some but not all words are very relevant. Figure 5 plots the relevance of individual paragraphs of the chapter 13.

The focus here is on paragraph 5 where most words have high correlation. While the individual word *noia* here has a low relevance the overall score $f()$ of other terms, like

abitudini/*habits*, *fatalita'*/*fatality*, is significantly above the mean value. Appendix 5. reports the entire transription of the 5th paragraph, where the troubled thinking of *Michele*, a young character of the novel, is closely followed. Here the only nouns *existence* and *deceit* of L_{noia} are found, while other words (like *habits*, *fatality* or *adventure*) are missing. However, as Fig. 5 suggests, the latter can be even more strongly correlated with the paragraph. In some sense they are evoked from the text via the connection through the latent semantic space dimensions. This proofs that the semantic domain *noia* is captured and works as an attractor more powerful than other purely lexicalized approaches (e.g. collocational analysis).

5. Conclusions

This paper is an early attempt, although still limited in scope and depth, to bridge a gap existing bewteen narrative analysis and Natural Language Processing (NLP). Semiotics, in the last century, has traced the paths to define and formalize linguistic and narrative concepts. Current NLP plays a double role here. It is a framework providing technologies and tools to support empirical validation of semiotic theories over a larger scale. On the other hand, the applicability of validated semiotics models through advanced NLP tools, will give rise to a new generation of technologies, characterized by higher levels of abstraction and better suited for supporting effective and natural interactivity within human-machine interfaces.

The empirical evidence discussed in this paper already shows that a geometrical approach can provide systematic definitions of quantitative models of meaning with a strong lexical emphasis. Conceptualizations here are driven by lexical information as it is distributionally observable in texts and geometrically modelled in vector spaces. We explored here syntagmatic and paradigmatic dimensions in meaning as a combination of constraints across different levels of knowledge: a social level provided by extended collection of texts, a paradigmatic level provided by the Wordnet semantic network and syntagmatic evidence as observable in the target literary work. The proposed analysis achieves impressive results as a significant amount of semantic evidence is captured in a fully automatic way. Future work will allow to assess these results over other collections and other literary genres. It is certainly true that the perspectives opened by the technology proposed in this paper are huge. This challenge is worth of a carefull and passionate research in view of more proactive paradigms of computer-assisted literary analysis.

Appendix 1: Example from *"Gli Indifferenti"*

Paragraph 5 in Chapter 13

La prima ipotesi era chiara; si trattava di isolarsi con poche idee, con pochi sentimenti veramente sentiti, con poche persone veramente amate, se ce n'erano, e ricominciare su queste basi esigue ma solide una vita fedele ai suoi principi di sincerita'. La seconda, eccola qui: nulla sarebbe mutato se non nel suo spirito sconfitto; avrebbe aggiustato alla meglio la situazione come una brutta casa in rovina, che si rifa' qua e la', non essendo possibile per mancanza di denari fabbricarne una nuova: avrebbe lasciato che la sua famiglia andasse in rovina o che si facesse mantenere da Leo, e si sarebbe risolto a sua volta (benche' molto l'umiliasse accontentarsi di una tale consolazione) a far la sua piccola sudiceria con Lisa; porcherie, piccole bassezze, piccole ***falsita'****, chi non ne depone in tutti gli angoli dell'****esistenza*** *come in quelli di una grande casa vuota? Addio vita chiara, vita limpida: sarebbe diventato l'amante di Lisa.*

6. References

Roberto Basili, Marco Cammisa, and Fabio Massimo Zanzotto. 2004. A semantic similarity measure for unsupervised semantic disambiguation. In *Proceedings of the Language, Resources and Evaluation LREC 2004 Conference*, Lisbon, Portugal.

M.W. Berry, S.T. Dumais, and G.W. O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595.

S. Deerwester, S. Dumais, G. Furnas, and R. Harshman T. Landauer. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Umberto Eco. 1960. *Opera Aperta*. Bompiani, Milano.

Alfio M. Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. University of Trento.

D. Hull. 1994. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

T. K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

G. Miller. 1990. An on-line lexical database. *International Journal of Lexicography*, 13(4):235–312.

Alberto Moravia. 1929. *Gli Indifferenti*. Bompiani.

G. Vigliocco and D.P. Vinson. 2005. Semantic representation. In G. Gaskell, editor, *Handbook of Psycholinguistics*. Oxford University Press, Oxford.

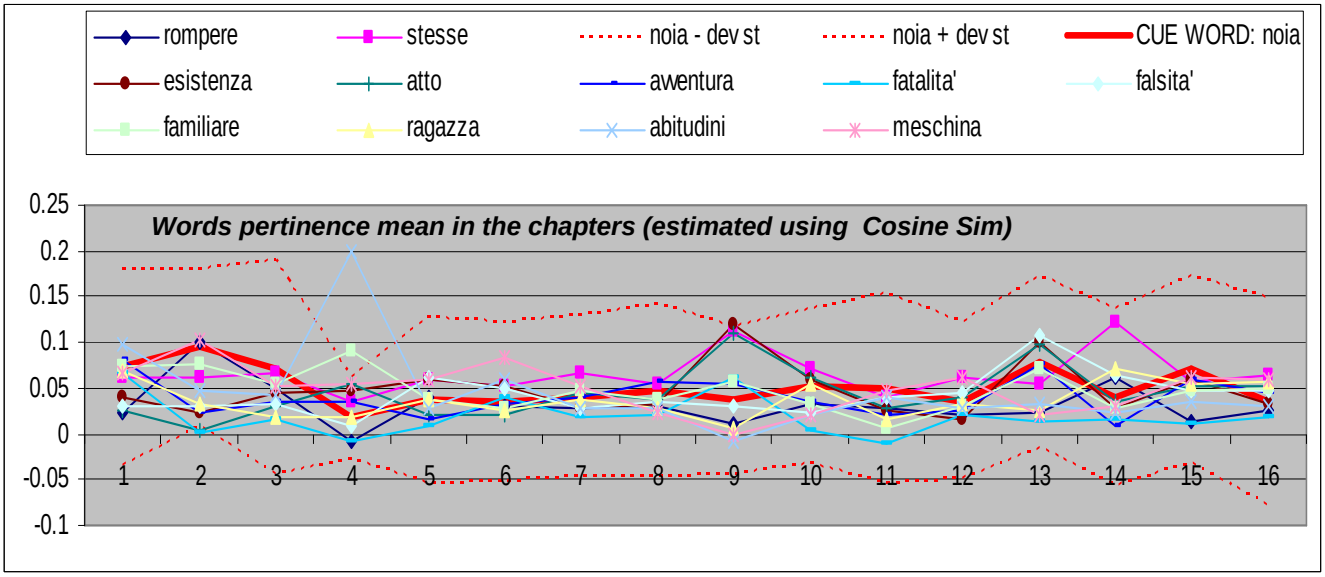

Figure 4: The plot of the semantic domain "*noia*" across the novel chapters

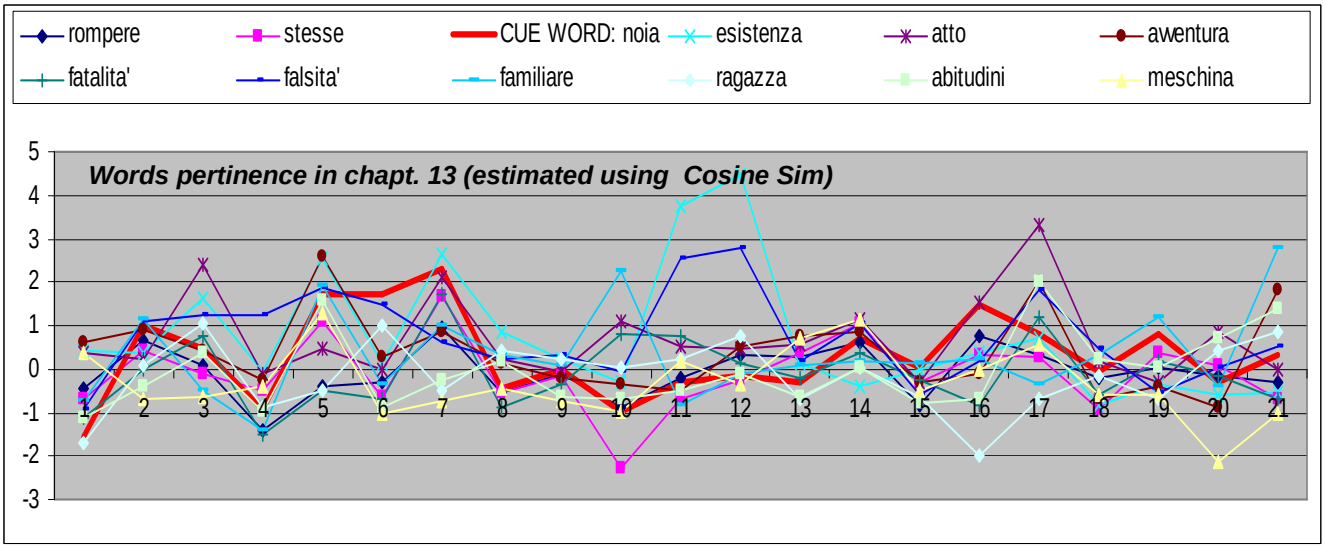

Figure 5: The plot of the semantic domain "*noia*" across chapter 13

The encoding model of Puccini's Correspondence Project

Elena Pierazzo

University of Pisa, Centro Studi Giacomo Puccini (permanent consultant)
Dipartimento Studi Italianistici, via del Collegio Ricci 10, 56126 Pisa (Italy)
pierazzo@ital.unipi.it

Abstract

The present paper focuses on the project of the edition of Giacomo Puccini correspondence held by the Centro Studi *Giacomo Puccini* (Lucca- Italy), that consists in the publication of 5.800 to 10.000 letters. The edition will consist in a digital and a printed edition, generated by an XML document through the application of different XSL stylesheets. The choice of a suitable encoding schema is discussed as well as the advantages and disadvantages of the most used encoding schemas such as TEI (P4 and P5) and DALF, the latter being the one selected by the project. The object of the project is discussed as well, and letters are distinguished from other documents on the basis of a practical definition of what is a letter. The paper presents the encoding model adopted by the project and discusses the three main parts of the encoded documents: metadata, envelope and postal matters, body of the letter. Finally, the two outputs (web and printable) are presented in their main points.

1. The correspondence of Puccini: printed or digital edition?

The study of correspondence of authors and leading people of the past is important for several reasons that would be too long to consider in depth. However, it is worth mentioning that correspondence is fundamental to reconstruct the biography and the relationships of the author, and it can provide important information about the circumstances of the artistic creations. Furthermore, correspondence is an indispensable historical and linguistic document, and can provide the historian, literary, textual critic and linguistics scholars with invaluable information. For all these reasons, the edition of correspondence has always held a big role in literary and historical studies.

Started at the Centro Studi *Giacomo Puccini* the *Progetto Epistolario* (PE) is a demanding project due to the consistence of the preserved material. At present, about 5.800 letters have been traced, but such number increases sensibly each year (about 400 units per year) thanks to a thorough inspection of libraries, private and public archives, collections and antiquaries' catalogues. The existence of at least 10.000 preserved letters can be easily supposed.

This publication is the Centro Studi main focus project over the next 8 years; the first two volumes of the correspondence – letters written before the 1900 – are scheduled for completion in 2008, in time for the end of the Puccini Celebrations (which began in 2004 in coincidence with the centenary of the writing of *Madama Butterfly* and end in 2008 in coincidence with the 150 years birth's anniversary).

The peculiarity of PE is that the publication will be in a double print and digital format. This choice fulfil different needs of both academic and general audience.
A print publication in fact:

1. provides physical consistence to the work, i.e. it produces a physical tangible object;
2. is easily and permanently quotable;
3. can be benefited by users not acclimatized with a digital environment;
4. has a stronger settlement in academic world.

On the other side a digital publication:

1. can be easily updated even after its initial publication, point of capital importance considering the constant increase in the number of known letters;
2. can be easily and quickly accessed from all over the world with just the cost of the internet connection;
3. can be used for linguistic and lexical inspections;
4. can be easily indexed according to different needs of users.

On the other hand, the twofold medium carries many difficulties in the production phase, the main laying in the synchronization of versions. A print edition and a digital edition, in fact, organize differently information and editorial interventions; for example a footnote related to the name of a person in a print edition, can be substituted by an hyperlink to an ancillary Index of Names document in a digital edition; the same for internal cross references: a footnote in print can correspond to a link in digital.

For these reasons we decide to produce both print and digital versions from the same master document, the more suitable format being a digital edition based on encoding.

Practice of encoding is at present widely based on XML language that allows the encoding in a single document of entities that can be handled differently by different stylesheets (like XSL and FO), producing printable or browsable versions.

The design of the encoding model for PE needed to consider firstly two main points:

5. the selection of the object of the edition, including a workable definition of letter
6. the adoption of an existing encoding schema vs. the creation of a new custom schema.

2. Object of the edition

Documents of different nature have been taken into consideration: letters, telegrams, postcards, illustrated postcards and cards; we discussed whether to include documents like petitions (submitted to several institutions) in which we found Puccini's autographic signature and dedicated photographs and scores.

The articulate debate among scholars on "What is a letter?" told us that this is not an odd question and that there is no unique answer.[1] For practical purposes (we

[1] For a comprehensive panorama on such topic, see Vanhoutte & Van der Branden 2005.

have no pretension of providing a theoretical answer to the question) we define "letter" a written document that:

1. is an a-synchronic form of communication
2. is written by a sender (singular or collective)
3. is addressed to a receiver (singular or collective)
4. has an informative content, i.e. the message contained in it is not known to the receiver.

Such restriction given, we considered that, even if petitions and dedications can help in reconstruct some moments of Puccini's biography and give important evidence of the relationships between the composer and a particular person or cultural sphere, they cannot be considered as correspondence without some forcing: petitions fail because they are a sort of a collective public act, dedications are evidence of a more or less deep connection between two persons with a small or inexistent informative content. However, in consideration of their historical and cultural relevance, we decided to include such documentation in the PE but with a different statement, i.e. creating an appendix of documents to complete the edition.

3. The encoding schema

The choice of the encoding schema has been a very important point. From the very beginning we discarded the idea of creating a new encoding schema that would have precludes us from the possibility of sharing our work with the international community.

Looking at existing schemas, we initially considered the adoption of TEI P4,[2] maybe the most known encoding schema in the Humanities Computing world. Unfortunately, such DTD gives not enough evidence to some peculiar characteristics of the correspondence, such as postmarks, headed papers, envelopes. The same remarks can be extended to the new TEI release, the P5 schema.[3]

We finally settled on the DALF DTD, a customization of the TEI DTD suited for correspondence, developed by the Centre for Scholarly Editing and Document Studies (CTB), a research centre of the Royal Academy of Dutch Language and Literature in Belgium.[4] Such DTD, in fact, includes all the features we pointed out as distinctive of correspondence. However, the provided DTD had to be extended in a couple of instances (see below). This resulted in a cooperation with the DALF team that will soon provide a new joined release of the updated encoding schema.

For the encoding of ancillary documents (as petitions and dedications) we decided to adopt the TEI P5 (Sperberg-McQueen and Burnard, 2005) that sensibly improves the manuscript transcription section with respect to the P4 version. The P5 is actually in a draft status, but the module for the description of a manuscript (intending with 'manuscript' any kind of handwritten primary sources) is considered by the TEI as stable and nearly definitive.

Since the letter description included in the DALF DTD has been inspired by the Master Project DTD[5] as well as the P5 manuscript description section, it will be quite easy to manage, index and query both letters and document with the same tools.

4. The encoding model

The choice of the object of the encoding and the choice of the encoding schema does not exhaust the modelling issues: it is also necessary to elaborate a specific encoding model able to provide all the characteristics considered important by the editors.

The creation of the encoding model is crucial to the design of any project, because the choices operated in such phase will determine from the beginning the answers that the encoded text will be able to give to scholars and final users.

Two main aspects must be taken into consideration at this level:

1. the specificities of the object
2. the wished output, including both the visualization aspects and the querying needs.

In elaborating the encoding model particular attention was paid to the editorial habits of the scholars and of the target consisting in Academics and Puccini fans.

The encoding model includes:

1. an accurate set of metadata
2. transcription of the envelop and of other postal matters
3. transcription of the body of the letter.

4.1. Metadata

4.1.1. Identification of the letter

The identification of the letter is performed through its localisation, including information about country and city where the letter is preserved, the institution responsible of the preservation of the document, the eventual collection in which the letter is included and the eventual identification number or call number of the document. A field of notes contains evidence of the kind of source from which the transcription has been made, including values as: 'from the autograph', 'from the non autographic original' (i.e. in case of telegrams), 'from a facsimile', 'from a former publication' (when the letter is untraceable at present).

```
<letIdentifier>
   <country>Italia</country>
   <settlement>Lucca</settlement>
   <repository key="I-Lmp"> Museo casa natale
Puccini</repository>
   <collection>Assente</collection>
   <idno>Assente</idno>
   <note place="in-text"> dall'autografo </note>
</letIdentifier>
```

4.1.2. Credits (communicative participants)

[2] See the TEI (Text Encoding Initiative) web site: http://www.tei-c.org. For an edition of correspondence based on the P5 encoding schema see Schreibman, Gueguen, Kumar and Saddlemyer (2005).

[3] Anyhow the P5 has been adopted for the encoding of ancillary documents (petitions, dedications); see further.

[4] See the DALF (Digital Archive of Letters by Flemish authors and composers) web site: http://www.kantl.be/ctb/project/dalf/

[5] MASTER is a European Union funded project to create a single on-line catalogue of medieval manuscripts in European libraries. This project developed a single standard for computer-readable descriptions of manuscripts. MASTER is funded under the Framework IV Telematics for Libraries call.' http://www.cta.dmu.ac.uk/projects/master/ (Vanhoutte & Van der Branden, 2005).

The sender and the addressee are fully described, specifying if their names are explicitly written in the document or they have been inferred by the editors. Names of sender and addressee are linked to the Index of Name, a document that also provides a brief biography of each participant to the communicative process.

The same section gives details about the letter's place and date of creation, declaring if the information is taken from the document itself or is the result of a critical reconstruction.

An optional notes field contains an editorial declaration explaining how the eventual missing information has been reconstructed.

```
<letHeading>
   <author attested="yes"><name>Giacomo Puccini
</name> </author>
   <addressee attested="yes">
        <name>Elvira Bonturi</name>
   </addressee>
   <placeLet attested="no">Bruxelles</placeLet>
   <dateLet attested="no">
        <date value="09-10-1900">9 ottobre 1900
</date>
   </dateLet>
   <note place="in-text">la lettera è sicuramente
da collocare prima del matrimonio di <xref
to="FGEMIGNANI"> <name reg="Gemignani Fosca">
Fosca Gemignani</name></xref>, quindi prima del
<date value="16-07-1902">16 luglio 1902</date>;
<name reg="Puccini Giacomo">Puccini</name> prima
di allora fu a <placeName reg="Bruxelles">
Bruxelles </placeName> solo nel 1900, dal 6 al 26
ottobre; dato che <q>le prove sono indietro</q> e
che Puccini ritiene impossibile che l'opera vada
in scena prima del 20, non resta che martedì 9
ottobre, tra i due martedì possibili </note>
</letHeading>
```

4.1.3. Physical description

This section describes the physical aspect of the document. When available, descriptions include:

1. type of support, to be chosen from a controlled vocabulary (including: letter, letter with envelope, postcard, telegram, etc.)
2. type of paper
3. colour of the paper
4. number of sheets
5. ink
6. seals
7. dimensions.

A specific optional, discursive field describes the presence of eventual music notation in the letter.

Decorations eventually found in the document are described by two different fields, the first describing any kind of decoration produced by the act of writing (draws, scrawls; music notation is not considered as a decoration), and the second describing decorations originated independently from the writing act, i.e. pre-printed draws, (as in headed paper) and similar matters.

A database has been provided to manage headed papers and heading stamps,[6] as they can have a fundamental importance in the localization and dating of a letter. Such database collects hundreds of still images together with the full transcription of the included words. The database provides simple querying facilities and, is currently available only to the project staff, due to its rapid update and evolution. We plan to transform it in a web tool available to anyone in the near future.

The physical description ends with the explanation of the document's status of preservation, including information such the presence of drops, ink transparencies, humidity traces and mould.

```
<physDesc>
   <type>Lettera</type>
   <support>
        <p>
         <seg type="tipo di supporto">carta</seg>
          <seg type="paginazione">2 fogli piegati
a metà nel senso della larghezza, contenuti l'uno
nell'altro</seg>
         <seg type="colore carta">seppia</seg>
      <seg type="colore inchiostro"> nero</seg>
     </p>
   </support>
   <extent>148x200</extent>
   <layout>
      <p><seg type="foliazione">di otto facciate
su carta intestata <xref to="B-Bghcol">Le Grand
Hotel,  Bruxelles</xref></seg>.  La  scrittura
comincia a c. 2r e continua ininterrotta fino a
c. 4r.</p>
   </layout>
   <paraphernalia>
     <paraphList>
       <paraphItem id="carta_intestata">
         <paraphDesc>
           <p><xref to="B-Bghcol">carta intestata
            </xref></p>
         <paraphText>
           <p><seg>Grand Hotel</seg>
              <seg>rue du Temple, 4</seg>
              <seg>Bruxelles</seg></p>
         </paraphText>
       </paraphDesc>
     </paraphItem>
    </paraphList>
   </paraphernalia>
   <condition>
        <p>Buono stato di conservazione</p>
   </condition>
</physDesc>
```

4.1.4. Provenance of the letter

When the actual physical location of a letter is unknown, we provide a specific field in which to include its last known locations. A typical example is given by those letters transcribed from the reproductions once available from the catalog (printed or web) of auction houses, and later sold to unknown private collectors. Another example is given by letters transcribed and published by someone in the past, but untraceable at present.

```
<history>
   <provenance>
        <p>In data <date value="23-10-2001">23-
10-2001</date> era presente nel sito web di
Sotheby, non è attualmente noto se è tuttora
presso la casa d'asta oppure è stata venduta.</p>
   </provenance>
```

[6] Stamps containing the logo of hotels, restaurants, theatres or publishing houses, able to transform white paper into headed paper.

```
</history>
```

4.1.5. Bibliography

A selected list of short references of eventual previous publications of the letter is provided. The list includes:

1. first publication of the letter
2. normally quoted edition (if any)
3. first translation
4. first quotation of the letter.

Complete references are given in a Bibliography List ancillary document, pointed by the short references.

```
<listBibl>
   <bibl>
     <author>
      <xref to="CARNER">Carner</xref>
     </author>
     <biblScope>106, 255</biblScope>
   </bibl>
   <bibl>
     <author>
       <xref to="MAREK">Marek</xref>
     </author>
   <biblScope>92–3, 94, 209–10</biblScope>
   </bibl>
</listBibl>
```

4.2. The envelope and other postal matters

For envelope we intend the part of a letter that provides postal information. Frequently such information are, in fact, not on a separate envelope, but included on the back side of the letter sheet, while, in case of postcards or illustrated postcards, they are written in designed field of the postcard itself. In case of telegrams they are normally provided at the beginning of the sheet, before the telegram's content.

The envelope section provides the full transcription of addresses (the sender as well as the addressee addresses), postmarks and reference numbers.

We distinguish from departure (d), arrival (a), and transit (t) postmarks (provided in such order) by means of an attribute.

Reference numbers are commonly used in Italian public institutions, meaning the reception of any correspondence or document, they represent a sort of internal (to a public or private institution) postmark. They are eventually accompanied by annotations (declaring the transition from an internal department to another) or drafts of the answer.

The DALF DTD doesn't provide – at present – a specific element to encode reference numbers, even if, in our opinion, they are a peculiar correspondence's feature; under the suggestion of the PE team the creation of an element to encode reference numbers will be included in the next release of the DALF DTD. In the mean time we encode reference numbers with the help of a generic element specified by an attribute.

```
<envelope>
   <envPart side="back">
     <address type="receiver">
       <addrLine><abbr  expan="Signora">  Sig<hi
rend="superscript">a</hi> </abbr> Elvira Puccini
       </addrLine>
       <addrLine>Verdi 4</addrLine>
       <addrLine><hi rend="underlined">Milano
       </hi></addrLine>
     </address>
     <postmark n="d">
      <placeName>Napoli Ferrovia</placeName>
      <date> 21 / 1 – 06 5 <abbr expan="Sera">S
      </abbr></date>
     </postmark>
     <postmark n="a">
      <placeName> Milano (Centrale)</placeName>
      <date> 23 / 1 – 06 5 <abbr expan="Mattina">
      M</abbr></date>
     </postmark>
   </envPart>
   <envPart>
     <div>
       <p> <seg type="reference number"> <name
type="institution"> Comune di Lucca </name>, <num
type="reference          number">          12097
</num> <date value="12-07-1901"> 12 lug. 1901
</date></seg>
       </p>
     </div>
   </envPart>
</envelope>
```

4.3. The body of the letter

The letter is semi-diplomatically transcribed, i.e. we tried to preserve most of the original's characteristics, the main exception being not recording the original's line interruptions; full stop and new line or presumably voluntary line interruption are obviously recorded.

Postscripts are encoded with specific elements; we considered as 'postscript' a letter's section explicitly introduced by a 'P.S.' (or similar) formula and everything written after the author's signature.

```
<ps>
<label>P.S. </label><lb/>
 Ho  bisogno  di  te  per  <title  reg="Madama
Butterfly"  type="opera"  n="Madama  Butterfly">
Butterfly</title>  -    scrivimi  a  <placeName
reg="Torre del lago">Torre </placeName> per saper
dove ti trovi - si tratta di piccoli accomodi
<lb/>
Ciao
</ps>
```

Transcription includes the reproduction of the main graphical issues of the text, such as: emphasizing, additions and deletions, eventual gaps. A limited number of editorial manipulations are also possible (correction of material errors, editorial notes, expansion of abbreviations and regularization of names).

4.3.1. Graphical issues

Emphasizing, single or double underlined words, the use of italics, capital letters are recorded, as well as minor graphical issues such as superscripted words or lines isolating paragraphs.

```
Sono stato dopo le prove a far visite al <title
type="newspaper">  <hi  rend="underlined">mattino
</hi> </title>
```

An added text is inserted in the place where the author intended to insert it, suppressing the eventual cross-reference mark; the encoding of additions defines the place where text was added (e.g. left margin, bottom margin, following page, etc.) also.

```
Sabato sera <add place="underlinear"> anzi domenica mattina</add>
```

Deletions produced by the author in the act of writing are encoded as well; in case of an unreadable deleted text, we insert an "xxx" string.

```
dopo fatte le prove, che la <sic>velocita</sic> di <del>1000 400</del><add>10400</add> fu raggiunta (equivalente a 6 miglia e mezzo)
```

4.3.2. Damages and correction of material errors

Damages or unreadable words in the original are recorded together with the probable extension of the gap.

```
così: spargiamo<lb/>
intorno april <gap reason="unreadable" extent="1 w"/> invece del seminiamo
```

In case of evident material errors (e.g. lapsus calami, spelling, etc.) we mark the error without providing a correction, except in case of possible misunderstandings.

```
Il Signore Giapponese che tenta Cho Cho San è cambiato in <sic>miliardajo</sic> debosciato americano
```

```
Ti si piomba addosso Evira <corr sic="Posca">Fosca</corr> Leonardi e Giacomo.
```

4.3.3. Abbreviations

The use of abbreviations in writing is typical of the private correspondence, and, in particular, it is usual for Puccini who used to write many letters a day.

For that reasons we adopted a slightly conservative policy. We distinguish, in fact, between abbreviations to be expanded and abbreviations not to be, i.e. we expand only abbreviations that can possibly give interpretation problems to the reader and preserve the original face of abbreviated words that can easily be understood by a modern reader because they either have a unique expansion or the word is still used abbreviated.

```
Poi, dice il Sig. Giulio, manda a lui o a me
```

```
Ricevo <abbr expan="telegramma">teleg</abbr> da Milano
```

4.3.4. Regularization of names and dates

Correspondence has a peculiar historical importance because it relates to facts and people. The encoding model must take into account this fundamental point.

In order to allow automatic inspections or to generate automatic indexes, names and dates are exhaustively encoded, regularized and classified.

Dates, even if incomplete, are regularized to a standard notation system: DD-MM-YYYY; in case of missing data, we supply 0 digits.

```
<date value="21-01-1906">21 gen. 1906</date>
```

```
<date value="00-01-1906">gen. 1906</date>
```

Date ranges are also considered and conveniently encoded:

```
<dateRange from="15-03-1898" to="31-03-1898">la seconda metà del mese</dateRange>
```

About names, we distinguish:

1. Names of person, classified (when applicable) in:
    a. singer
    b. conductor
    c. composer
    d. character
    e. author
    f. scenographer
    g. costume designer
    h. musician
    i. librettist
    j. editor
    k. director
    l. manager
    m. journalist
    n. critic
    o. publisher

```
<persName type="librettist" reg="Illica Luigi">Illica</persName>
```

```
<persName reg="Bonturi Elvira"> Topisia </persName>
```

2. Names of place

```
passerò a <placeName reg="Torre del Lago">Torre </placeName> domani
```

3. Other names of relevance, classified in:
    a. publishing house
    b. car
    c. institution
    d. theatre
    e. engine (i.e. boats, motorcycles)

```
alla <name key="theatre" reg="Teatro alla Scala">Scala</name>
```

4.3.5. Quotations

Quoted text is marked both in the body of the letter and in editorial notes.

Quoted titles are encoded and regularized twice: the first regularization traces the titles to the normal quotation form; the second regularization traces it to a form that put in the first position the main semantic word (e.g. excluding an eventual article) in order to enable automatic indexing.

Titles are classified in:

1. opera
2. journal
3. newspaper
4. drama
5. music

```
<title type="opera" reg="La Tosca" n="Tosca, La">Tosca</title>
```

4.3.6. Changes of hands and pre-printed text

Changes of hands are marked with a generic element actualized by an attribute. We chose not to adopt the specific TEI/DALF element <hand> because it is an empty element that precludes the selection of the different hands contributions. These lack also has been submitted to the DALF team.

```
<seg type="Puccini Antonio"> <xref to="h2" targType="hand"/> un bacione anche da me</seg>
```

Pre-printed words (e.g. in telegram forms) are marked with a specific DALF element.

```
<p rend="right"> <print> <placeName reg="Milano"> Milano</placeName> </print> <date value="07-03-1901">7. Marzo 901 </date></p>
```

5. Outputs

The project provides two different outputs for the encoded texts: a browsable digital output (hypertext) and a static print version suited for an inclusion in a volume.

The two outputs are obtained thanks to the application of XSL and XSL:FO stylesheets. Each element of the encoding model is handled differently by the two stylesheets in consideration of the needs of the two outputs (Landow 1992).

The hypertext (HTML) output includes a number of links to other texts of the corpus:

1. to other letters concerning the same topics
2. to editorial notes
3. to the Indexes (of Names, of References, of Addressee).

These links are substituted mostly in the print version by footnotes and cross-references; links to names are substituted by the Index of Names (that includes the page numbers of each reference) provided at the end of each volume.

Recurrent notes (e.g. biographical references about some frequently mentioned people) are managed in hypertext with an editorial note each time that they recur, as hypertexts allows (and, in some case, encourages) non-consequential access to the texts and it is impossible to preview set courses; for that reason a number of redundant information must be provided.

In print text recurring notes are rendered with a single footnote per volume containing the explanations, and a number of footnotes giving a cross-reference to that note.

Credits for each letter (scientific responsibilities, encoding, editorial processes) are given in hypertexts in a box loaded by an hyperlink, while in the print version every volume gives the same information in single comprehensive prefatory note.

5.1. Printable output

An initial print model was selected for the publication of the correspondence of Puccini with Luigi Illica and Giuseppe Giacosa regarding the composition of the *Madama Butterfly* included in Gross & Bernardoni, Biagi Ravenni, Schickling (2005).

The PE model revised such model including some more information. The new print model consists of the following sections:

1. Heading, given by:
    a. Letter code
    b. Addressee
    c. Place and date of letter's composition
2. Metadata, including:
    a. Type of the letter
    b. Synthetic physical description
    c. Addressee's address
    d. Postmarks
    e. Type of source, location of the letter
    f. Source of the provided transcription
    g. Notes
3. Body of the letter
4. Editorial footnotes.

5.2. Web output

The Web output model was designed based on the print model, the main differences given by the different treatment of links, cross-references and footnotes, as mentioned above.

The web output model gives more details on the physical description of the source document and the credits than the print one. These details do not modify the main layout structures as they are loaded in pop-up boxes opened by hyperlinks, with the result that at a first sight web output and printable output are very similar

6. References

Gross, A. and Bernardoni, V., Biagi Ravenni, G., Schickling, D. (2005). *Madama Butterfly. Fonti e documenti della genesi*. Lucca, Centro studi Giacomo Puccini & Maria Pacini Fazzi.

Landow, G.P. (1992), *Hypertext : the convergence of contemporary critical theory and technology*. Baltimore & London, The John Hopkins U.P.

Schreibman, S., Gueguen, G., Kumar A., Saddlemyer, A. (2005). Letters and lacunae: editing an electronic scholarly edition of correspondence. In: The Association for Computers and the Humanities/The Association for Literary and Linguistic Computing, *ACH / ALLC 2005. The International Conference on Humanities Computing and Digital Scholarship. The 17th Joint International Conference*, University of Victoria (Canada), June 15 - June 18.

Sperberg-Mcqueen, C.M. and Burnard, L. (2002). *TEI P4 Guidelines for Electronic Text Encoding and Interchange: XML-compatible Edition*. (P4). Oxford, Providence, Charlottesville, & Bergen. Available also at http://www.tei-c.org/P4X/.

Sperberg-Mcqueen, C.M. and Burnard, L. (2005). *TEI P5 Guidelines for Electronic Text Encoding and Interchange: Revised and re-edited*. (P5). Oxford, Providence, Charlottesville, & Nancy. Available at http://www.tei-c.org/release/doc/tei-p5-doc/html/.

Vanhoutte E. and Van der Branden, R. (2005). Describing, Transcribing, Encoding, and Editing Modern Correspondence Material: a Textbase Approach. In: Fred Unwalla & Peter Shillingsburg (Eds.), *Computing the edition*. Toronto, Toronto University Press. Preprint available at (link visited on 01-04-2006): http://www.kantl.be/ctb/pub/2004/comedvanvanfig.pdf.

Vanhoutte E. and Van der Branden, R. (2003). *DALF guidelines fort he description and encoding of modern correspondence material. Version 1.0*. Gent, Centrum voor Teksteditie en Bronnenstudie.

Stylogenetics: Clustering-based stylistic analysis of literary corpora

Kim Luyckx, Walter Daelemans, Edward Vanhoutte

University of Antwerp
Faculty of Arts
Universiteitsplein 1
B-2610 Antwerp
Belgium
{kim.luyckx, walter.daelemans, edward.vanhoutte}@ua.ac.be

Abstract

Current advances in shallow parsing allow us to use results from this field in stylogenetic research, so that a new methodology for the automatic analysis of literary texts can be developed. The main pillars of this methodology - which is borrowed from topic detection research - are (i) using more complex features than the simple lexical features suggested by traditional approaches, (ii) using authors or groups of authors as a prediction class, and (iii) using clustering methods to indicate the differences and similarities between authors (i.e. stylogenetics). On the basis of the stylistic genome of authors, we try to cluster them into closely related and meaningful groups. We report on experiments with a literary corpus of five million words consisting of representative samples of female and male authors. Combinations of syntactic, token-based and lexical features constitute a profile that characterizes the style of an author. The stylogenetics methodology opens up new perspectives for literary analysis, enabling and necessitating close cooperation between literary scholars and computational linguists.

1. Introduction

Recently, language technology has progressed to a state of the art in which robust and fairly accurate linguistic analysis of lexical, morphological, and syntactic properties of text has become feasible. This enables the systematic study of the variation of these linguistic properties in texts by different authors (author identification) (Baayen et al., 1996; Gamon, 2004), different time periods, different genres or registers (Argamon et al., 2003), different regiolects, and even different genders (Koppel et al., 2003; Kelih et al., 2005).

We see this trend as potentially providing new tools and a new methodology for the analysis of literary texts that has traditionally focused on complex and deep markup (McCarty, 2003) and the statistical assesments of concordances and word-count applications (Raben, 1965; Burrows, 1987; Lancashire, 1993; Bucher-Gillmayr, 1996) for the analysis of rhyme and sound patterns (Wisbey, 1971; Robey, 2000), the investigation of imagery and themes (Corns, 1982; Fortier, 1989; Fortier, 1996; Ide, 1986; Ide, 1989), the structure of dramatic works (Potter, 1981; Potter, 1989; Steele, 1991; Ilsemann, 1995), stylometrics and authorship attribution (Hockey, 2000, 104-123), (Craig, 2004). See (Rommel, 2004) for an overview of computational methods in literary studies. The methodology we propose is borrowed from the text categorization literature (Sebastiani, 2002) where simple lexical features (called a bag of words) are used to characterize a document with some topic class. Statistical and information-theoretic methods are used to select an informative bag of words to distinguish between documents with different topics. Machine Learning methods are then used to learn to assign documents to one of the predefined topics on the basis of examples. We generalize this methodology in three ways:

- i. By extending the simple lexical features with more complex features based on distributional syntactic information about part of speech tags, nominal and verbal constituent patterns, as well as features representing readability aspects (average word and sentence length, type/token ratio etc.). The statistical and information-theoretic methods can then be applied to more complex features than individual words for stylistic analysis.

- ii. By using individual authors or groups of authors as classes to be predicted rather than topics. It can then be investigated which features are predictive for author identity, gender, time period etc. See (Koppel et al., 2003) for work on this approach for gender prediction.

- iii. By using the vectors of complex features, computed on a sufficiently large sample of the work of an author as a signature for the style of that author and using similarity-based clustering methods to develop a stylogenetic analysis of differences and similarities between authors, periods and genders. We define stylogenetics here as an approach to literary analysis that groups authors on the basis of its stylistic genome into family trees or closely related groups from some perspective.

Tree classification as a tool for the study of proximity and distance between texts and authors has recently been explored by few studies which take the whole vocabulary of the texts which are compared into consideration. (Julliard and Luong, 1997; Julliard and Luong, 2001; Spencer et al., 2003; Labbé and Labbé, to appear 2006). Central in these studies, however, are not the complex features as proposed in our methodology, but the lexical and lexicographical standardization of the vocabulary that is the qualitative basis for proximity measurements between pairs of texts.

2. Corpus

In this paper we report on explorative stylogenetic work using a large corpus of literary works. From three online text archives (viz. The Oxford Text Archive, the Electronic Text Center of the University of Virginia and to a minor extent Project Gutenberg) we collected representative samples of 100,000 words of 50 English and American authors, half of them male, half of them female, from 12 time periods between 1525 and 1925 (we worked with 25-year periods). The appendix provides an overview of the authors, genders, and periodization of the samples used (cf. Tables 1, 2).

3. Feature Extraction

Four types of features that have been applied as style markers can be distinguished: token-level features (e.g. word length, readability), syntactic features (e.g. part-of-speech tags, chunks), features based on vocabulary richness (e.g. type-token ratio) and common word frequencies (e.g. of function words) (Stamatatos et al., 2001). While most stylometric studies are based on token-level features, word forms and their frequencies of occurrence, syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Baayen et al., 1996; Diederich et al., 2000; Khmelev and Tweedie, 2001; Kukushkina et al., 2001; Stamatatos et al., 1999). Thanks to improvements in shallow text analysis, we can extract syntactic features to test their relevance in stylogenetic research.

In a first step, we developed an environment which enables the automatic production of profiles of se samples in the Stylogene corpus. A profile consists of a vector of 208 numerical features representing automatically assigned information about the following features:

- **Type-token ratio**: The type-token ratio *V/N*, *V* representing the size of the vocabulary of the sample, and *N* the number of tokens, is a measure indicating the vocabulary richness of an author.
- **Word length**: The distribution of words of different lengths has been used as a feature in authorship attribution studies (Diederich et al., 2000). Words with a length of 15-19, 20-24 and 25+ were combined in separate categories.
- **Readability**: The readability feature is an implementation of the Flesch-Kincaid metric which indicates the readability of a text, using mean word and sentence length.
- **Distribution of parts-of-speech**: Syntax-based features are not under the conscious control of the author and therefore reliable style markers. Somers suggests that

  > A more cultivated intellectual habit of thinking can increase the number of substantives used, while a more dynamic empathy and active attitude can be habitually expressed by means of an increased number of verbs. (Holmes, 1994, 89)

- **Distribution of frequent function words**: Traditional approaches to stylometry research use content words rather than function words, assuming that the latter occur to frequently to be of any relevance for style. Nevertheless, function words (e.g. determiners, conjunctions, prepositions) are not under the conscious control of the author and therefore meaningful for stylogenetic studies (Holmes, 1994, 90-91).
- **Distribution of frequent chunks**: Similarly to parts-of-speech, chunks are also reliable features for stylogenetic research. We automatically extracted frequencies of noun phrase, verb phrase, prepositional phrase, adjectival phrase, adverbial phrase, conjunction, interjection, verb particle, subordinated clause and preposition-noun phrase chunks.
- **NP and VP chunk internal variation**: The internal organisation of NP and VP chunks is subject to variation, which can reveal the subconscious preference of the author.

The resulting profiles can be used in applications like author or gender identification, but also in a stylogenetic analysis for the discovery of stylistic relationships between authors that may not be evident on the basis of a more superficial comparison. As a representation of contemporary non-literary language, we added a profile based on 100,000 words of Wall Street Journal text.

In order to be able to extract these features automatically, we used shallow parsing software developed in our lab (Daelemans and van den Bosch, 2005) to automatically assign parts of speech and constituent structure to the 51 x 100,000 word corpora. The pos tag set and chunk label set used are those of the Penn Treebank project (Marcus et al., 1993).

4. Cluster Analysis and Interpretation

The clustering method used is the one implemented in the *cluster* program of Andreas Stolcke, which is an instance of Euclidea distance based centroid clustering. Initially, all data points are treated as clusters and the most similar clusters are iteratively merged into larger clusters, building up a hierarchical tree.

Figure 1 shows the family tree produced by applying hierarchical clustering with Euclidean distance as similarity metric to the full profiles of each author.

In further exploratory research, we used information-theoretic analysis (i.e. Gain Ratio) of the relevance of each feature in the profile in predicting the gender of the author as a heuristic to select a new profile to cluster for gender-related stylistic family trees. We selected the 43 features that turned out to be the most relevant for characterizing style differences between genders.

Figure 2 shows the family tree after feature selection in which we find five groups of gender clusters.

The tree in Figure 1 shows that the Wall Street Journal (WSJ) profile is clearly separated from the rest of the corpus and that within the latter, Defoe, Hobbes, Mill, Behn, and More are stylistic outliers. The interrelation between genre and period may explain their distance from the rest

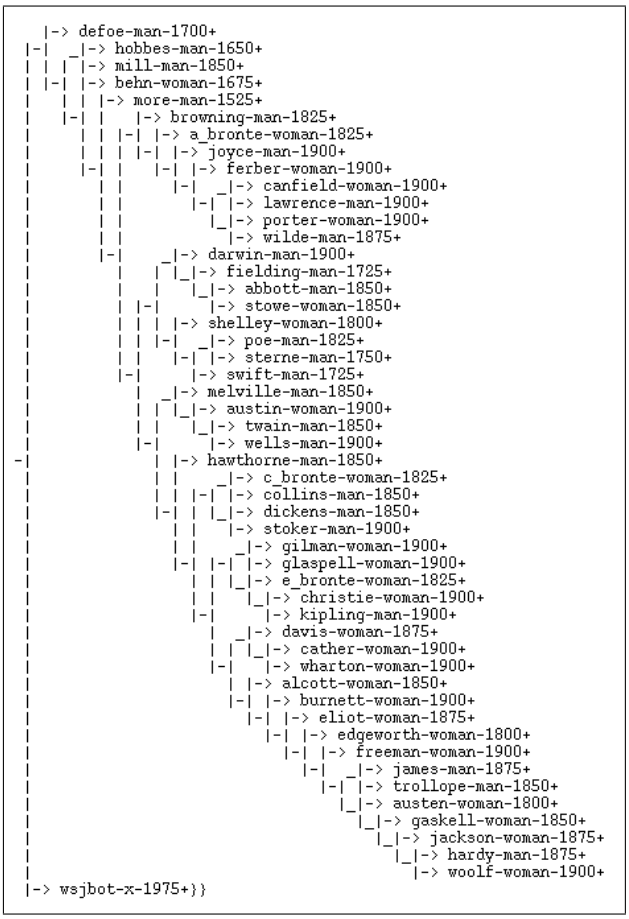

Figure 1: Family tree based on entire feature set

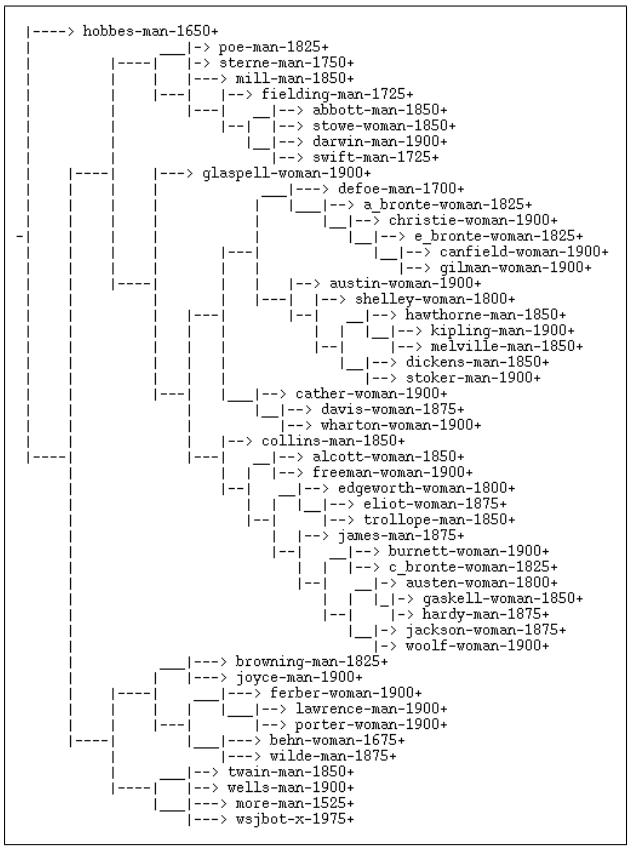

Figure 2: Family tree after feature selection on gender clustering

of the stylogene corpus. Hobbes, Behn, More and Defoe-as a borderline case-are significantly earlier texts, whereas the samples by Hobbes, Mill, and More all come from philosophical essays. As an early female playwright, Behn is also and understandably an outsider. Furthermore, clustering for gender seems to be quite successful. The family tree presents itself naturally in two parts, the upper part of which (from Defoe to Stoker) is predominantly populated by male authors (21 out of 30 or a score of 70%) and the lower part is strongly populated by female authors (16 out of 20 or a score of 80%). Since up to the end of the Victorian period, that is up to the beginning of the twentieth century, female authors are generally observed to adopt the prevailing male style of writing, the reason why four male authors (Kipling, James, Trollope, and Hardy) appear in the female part of the tree might be more interesting to study. In the second tree that shows the family tree after feature selection we can distinguish five groups of gender clusters with 11 exceptions (or 22%); six women writers (Stowe / Austin, Shelley / Ferber, Porter, Behn) and five male authors (Defoe / Collins, Trollope, James, Hardy). Aggregating the results from the first tree with the results from the gender-related stylistic family tree presented in Figure 2 reduces the initial female gender problem from 9 to 3 cases (only A. Brontë, Canfield, and, C. Brontë are correctly clustered within female groups after feature selection) and the male gender problem from 4 to 3 (James, Trollope, and Hardy). However, this clustering introduced two new problematic names: Defoe and Collins which, together with the remaining names, deserve further research.

5. Conclusions and Further Research

Without claiming any relevance for these particular family trees, it seems clear to us that specific literary style hypotheses can be tested using similar approaches. Close cooperation between literary scholars and computational linguists is essential for this.

We have shown that robust text analysis can bring a new set of tools to literary analysis. Specific hypotheses can be tested and new insights can be gained by representing the work (or different works) of authors as profiles and applying clustering and learning techniques to them. In future work we will investigate more specific literary hypotheses, and generalize the appoach to the analysis and comparison of individual books of authors rather than random samples of their work.

6. References

S. Argamon, M. Koppel, J. Fine, and A. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3).

H. Baayen, H. Van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

S. Bucher-Gillmayr. 1996. A computer-aided quest for allusions to biblical text within lyric poetry. *Literary and Linguistic Computing*, 11(1):1–8.

J. Burrows. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

T. Corns. 1982. *The development of Milton's prose style*. Oxford: Clarendon Press.

H. Craig, 2004. *A Companion to Digital Humanities*, chapter Analysis and authorship studies, pages 273–288. Malden, MA/Oxford/Carlton, Victoria: Blackwell Publishing.

W. Daelemans and A. van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2000. Authorship attribution with Support Vector Machines. *Applied Intelligence*, 19(1-2):109–123.

P. Fortier, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Analysis of twentieth-century French prose fiction, pages 77–95. Philadelphia: University of Pennsylvania Press.

P. Fortier, 1996. *Research in Humanities Computing 5: Papers from the 1995 ACH-ALLC Conference*, chapter Categories, theories, and words in literary texts, pages 91–109. Oxford: Oxford University Press.

M. Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of COLING 2004*, pages 611–617.

S. Hockey. 2000. *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

D. Holmes. 1994. Authorship Attribution. *Computers and the Humanities*, 28(2):87–106.

N. Ide, 1986. *Méthodes quantitatives et informatiques dans l'étude des textes. In honour of C. Muller*, chapter Patterns of imagery in William Blake's 'The Four Zoas', pages 497–505. Geneva: Slatkine.

N. Ide, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Meaning and method: computer-assisted analysis of Blake, pages 123–141. Philadelphia: University of Pennsylvania Press.

H. Ilsemann. 1995. Computerized drama analysis. *Literary and Linguistic Computing*, 10(1):11–21.

M. Julliard and X. Luong. 1997. Words in the hood. *Literary and Linguistic Computing*, 12(2):71–78.

M. Julliard and X. Luong. 2001. On consensus between tree-representation of linguistic data. *Literary and Linguistic Computing*, 16(1):59–76.

E. Kelih, G. Antíc, P. Grzybek, and E. Stadlober, 2005. *Classification of author and/or genre? The impact of word length*, pages 498–505. Heidelberg: Springer.

D. Khmelev and F. Tweedie. 2001. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(4):299–307.

M. Koppel, S. Argamon, and A. Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

O. Kukushkina, A. Polikarpov, and D. Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatsii*, 37(2):96–108. Translated as Problems of Information Transmission.

C. Labbé and D. Labbé. to appear, 2006. A tool for literary studies: intertextual distance and tree classification. *Literary and Linguistic Computing*. Advance access: October 27, 2005.

I. Lancashire, 1993. *The Digital Word: Text-Based Computing in the Humanities*, chapter Computer-assisted critical analysis: a case study of Margaret Atwood's Handmaid's Tale, pages 291–318. Cambridge, MA/London: The MIT Press.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics: Special Issue on Using Large Corpora*, 19(2):313–330.

W. McCarty. 2003. Depth, markup and modelling. *Text Technology*, 12(1).

R. Potter. 1981. Character definition through syntax: significant within-play variability in 21 English language plays. *Style*, 15:415–434.

R. Potter, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Changes in Shaw's dramatic rhetoric, pages 225–258. Philadelphia: University of Pennsylvania Press.

J. Raben. 1965. A computer-aided study of literary influence: Milton to Shelley. In *Literary Data Processing Conference Proceedings*, pages 230–274. White Plains: IBM.

D. Robey. 2000. *Sound and Structure in the Divine Comedy*. Oxford: Oxford University Press.

T. Rommel, 2004. *A Companion to Digital Humanities*, chapter Literary Studies, pages 88–96. Malden, MA/Oxford/Carlton, Victoria: Blackwell Publishing.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

M. Spencer, B. Bordalejo, P. Robison, and C. Howe. 2003. How reliable is a stemma? An analysis of Chaucer's Miller Tale. *Literary and Linguistic Computing*, 18(4):407–422.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic authorship attribution. In *Proceedings of EACL 99*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

K. Steele, 1991. *A TACT Exemplar*, chapter 'The Whole Wealth of Thy Wit in an Instant'. TACT and the explicit structures of Shakespeare's plays, pages 15–35. Toronto: Centre for Computing in the Humanities.

R. Wisbey, 1971. *The Computer in Literary in Linguistic Research*, chapter Publications from an archive of computer-readable literary texts, pages 19–34. Cambridge: Cambridge University Press.

Female authors	Works	Number of Words	Period
Louisa-May Alcott	*Little Women*	100,000	1850+
Jane Austen	*Mansfield Park*	100,000	1800+
Mary Austin	*The Trail Book*	83,918	1900+
	The Land of Little Rain	16,082	
Aphra Behn	*The Rover*	75,673	1675+
	The City Heiress	24,327	
Anne Brontë	*The Tenant of Wildfell Hall* 100,000		
Charlotte Brontë	*Jane Eyre*	100,000	1825+
Emily Brontë	*Wuthering Heights*	100,000	1825+
Frances Burnett	*The Secret Garden*	97,863	1900+
	A Little Princess	2,137	
Dorothy Canfield	*The Brimming Cup*	100,000	1900+
Willa Cather	*The Song of the Lark*	100,000	1900+
Agatha Christie	*The Secret Adversary*	95,852	1900+
	The Mysterious Affair at Styles	4,148	
Rebecca Davis	*Frances Waldeaux*	45,173	1875+
	Margret Howth	24,179	
	Life in the Iron-Mills	18,501	
	One Week an Editor	8,843	
	Walhalla	3,304	
Maria Edgeworth	*The Parent's Assistant*	100,000	1800+
George Eliot	*Silas Marner*	100,000	1875+
Edna Ferber	*Fanny Herself*	100,000	1900+
Mary Freeman	*The Heart's Highway*	85,980	1900+
	Copy-Cat and Other Stories	14,020	
Elizabeth Gaskell	*Sylvia's Lovers*	100,000	1850+
Charlotte Gilman	*What Diantha Did*	69,762	1900+
	Herland	30,238	
Susan Glaspell	*The Visioning*	100,000	1900+
Helen Jackson	*Ramona*	100,000	1875+
Eleanor Porter	*Just David*	100,000	1900+
Mary Shelley	*Frankenstein*	75,530	1800+
	Mathilda	24,470	
Harriet Stowe	*The Key to Uncle Tom's Cabin*	100,000	1850+
Edith Wharton	*The Age of Innocence*	100,000	1900+
Virginia Woolf	*Night and Day*	100,000	1900+

Table 1: Stylogene Literary Corpus: Female authors

Male authors	Works	Number of Words	Period
Jacob Abbott	*History of King Charles the Second of England*	65,076	1850+
	Aboriginal America	34,924	
Robert Browning	*Dramatic Romances*	57,541	1825+
	Sordello	42,459	
Wilkie Collins	*The Woman in White*	100,000	1850+
Charles Darwin	*The Voyage of the Beagle*	100,000	1900+
Daniel Defoe	*Moll Flanders*	100,000	1700+
Charles Dickens	*Dombey and Son*	100,000	1850+
Henry Fielding	*The History of Tom Jones, a Foundling*	100,000	1725+
Thomas Hardy	*Tess of the D'Urbervilles*	100,000	1875+
Nathaniel Hawthorne	*The Marble Faun*	100,000	1850+
Thomas Hobbes	*Leviathan*	100,000	1650+
Henry James	*The Portrait of a Lady*	100,000	1875+
James Joyce	*Ulysses*	100,000	1900+
Rudyard Kipling	*Actions and Reactions*	83,648	1900+
	Captains Curageous	16,352	
D.H. Lawrence	*Women in Love*	100,000	1900+
Herman Melville	*Moby Dick*	100,000	1850+
J.S. Mill	*On Liberty*	53,773	1850+
	The Subjection of Women	46,227	
Thomas More	*Dialogue of Comfort against Tribulation*	100,000	1525+
E.A. Poe	*A Descent into the Maelstrom*	100,000	1825+
	The Gold-Bug		
	Mellonta Tauta		
Laurence Sterne	*The Life and Opinions of Tristram Shandy*	100,000	1750+
Bram Stoker	*Dracula*	100,000	1900+
Jonathan Swift	*Gulliver's Travels*	100,000	1725+
Anthony Trollope	*Can You Forgive Her?*	100,000	1850+
Mark Twain	*The Innocents Abroad*	100,000	1850+
H.G. Wells	*The World Set Free*	73,522	1900+
	The War of the Worlds	26,478	
Oscar Wilde	*The Picture of Dorian Gray*	95,213	1875+
	Lord Arthur Savile's Crime	4,787	

Table 2: Stylogene Literary Corpus: Male authors

Sentiment Classification Techniques for Tracking Literary Reputation

Maite Taboada[1], Mary Ann Gillies[2] and Paul McFetridge[1]

[1]Department of Linguistics, [2]Department of English

Simon Fraser University

8888 University Dr., Burnaby, BC, V5A 1S6, Canada

E-mail: mtaboada@sfu.ca, gillies@sfu.ca, mcfet@sfu.ca

Abstract

The initial stages of a project tracking the literary reputation of authors are described. The critical reviews of six authors who either rose to fame or fell to obscurity between 1900 and 1950 will be examined and we hope to demonstrate the contribution of each text to the evolving reputations of the authors. We provide an initial report on the use of the semantic orientation of adjectives and their rough position in the text to calculate the overall orientation of the text and suggest ways in which this calculation can be improved. Improvements include further development of adjective lists, expansion of these lists and the consequent algorithms for calculating orientation to include other parts of speech, and the use of Rhetorical Structure Theory to differentiate units that make a direct contribution to the intended orientation from those that are contrastive or otherwise make an indirect contribution.

1. Introduction

The objective of our research is to extract information on the reputation of different authors, based on writings concerning the authors. The project aims to create a database of texts, and computational tools to extract content automatically.

Research on opinion and subjectivity in text has grown considerably in the last few years. New methods are being created to distinguish objective from subjective statements in a text, and to determine whether the subjective statements are positive or negative with respect to the particular subject matter. We believe that the methods currently being used to extract subjective opinion, or sentiment, from movie and consumer product reviews (e.g., Gamon, 2004; Hu & Liu, 2004; Turney, 2002) can be applied to literary reviews and other texts concerning author's works.

In this paper, we describe some of the methods currently being used to extract sentiment from text, and explain how we are applying those methods to literary reviews, letters to the editor, newspaper articles, and critical and scholarly publications concerning six authors who were active in the 1900-1950 period. Section 2 provides some background on literary reputation, and how we plan to quantify it. Section 3 discusses sentiment detection, as it has been applied to movie reviews and other present-day reviews of consumer reports. In Section 4, we address the issue of document structure: how important it is to identify the most important parts of the text, and what methods we can use to that end. This project is in its initial stages, and we do not have conclusive results yet. We present, however, the current state of the system in Section 5, and illustrate it with two examples in Section 6. Finally, conclusions and a discussion of future work are found in Section 7.

2. Background

The question of why writers' works, and by extension their literary reputations, fall in and out of critical and popular favour has long fascinated literary critics. In 1905, Marie Corelli was the best-known and most successful novelist in Britain. By 1950 she had been consigned to literary obscurity and few read her books. In 1910, T.S. Eliot was an unknown American poet in Paris, dreaming of "belonging in a great centre of artistic and intellectual innovation" (Gordon, 1977: 33). By 1950 Eliot, a Nobel Laureate, stood at the very centre of Western aesthetic and intellectual culture. Why had these two writers' reputations suffered such dramatically opposite fates? How do we account for such shifts in literary reputation? These two questions form the core of our project, on literary reputation in Britain between 1900 and 1950.

Scholarly discussions of publishing, readership, canon construction, and the various institutions of literature have proliferated in recent years, most of which attempt to map out how "our experience of the work" (Herrnstein Smith, 1988: 16) relates to its critical or popular value (Fromm, 1991; Guillory, 1993; Lecker, 1991; Remplin, 1995). And yet in literary studies, few of these discussions attempt to combine a quantitative analysis of data with a qualitative analysis. An exception is Gaye Tuchman & Nina Fortin's *Edging Women Out* which sets out to answer the question "Why does some literature supposedly transcend the ages and so constitute 'culture' while other once-popular books languish in disuse?" (Tuchman & Fortin, 1989: 1). Tuchman & Fortin focus on one publisher, Macmillan, from 1867-1917. They designed a quantitative study of Macmillan's records, identifying four distinct data sets and applying a systematic analysis of the records in order to derive conclusions about the "literary opportunities" of women at the turn of the century. Tuchman & Fortin admit, however, "Although our data about the literary

opportunities of most women novelists are substantial, our conclusions are based on inferences." (Tuchman & Fortin, 1989: 18). Our project asks similar questions to Tuchman & Fortin and Herrnstein Smith, but we have designed it so that it permits us to combine the aesthetic and evaluative concerns raised by the former with the kinds of quantitative methodology employed by the latter.

The quantitative aspects of the project are based on research in information retrieval and text categorization. We are scanning documents pertaining to the authors in this study into a computer database designed to store them, and we will then analyze these documents automatically for positive and negative content, i.e., the document's overall *sentiment*. This problem has been characterized as one of determining whether the text is "thumbs up" or "thumbs down" (Turney, 2002).

A number of techniques have been proposed for the problem of automatic sentiment classification, based on adjective classification (Hatzivassiloglou & McKeown, 1997), extraction of subjective content (Wiebe et al., 2004), or through the use of machine learning methods (Bai et al., 2004; Gamon, 2004; Pang et al., 2002). In all cases, the most difficult problem consists of finding the relevant parts of the text, those that contain subjective evaluation. We propose to apply our knowledge of text structure, and to use discourse parsing, a method that parses the discourse structure of the text, establishing main and secondary parts.

We are currently conducting a pilot project with two authors: John Galsworthy and D.H. Lawrence. We have in mind a larger project, with more authors. For the larger project, we have selected six writers: three who were very successful in the public discourse (financial and/or critically) in the early years of the 20th century and who had largely been consigned to the margins of literary study by 1950—John Galsworthy, Arnold Bennett, and Marie Corelli; and three who were less well known at that time but who came to occupy central places in the literary canon by 1950—Virginia Woolf, Joseph Conrad, and D.H. Lawrence.

We selected the time period 1900-1950 for two reasons. First, the advent of mass market publications around the turn of the century created new ways of producing and disseminating literature—for example, cheap paperback novels and tabloid newspapers helped transform the very definition of literature; at the same time, they focused ever greater attention on individual authors. Writers and readers came to view literature as something very different than had their Victorian parents thus making 1900 a marker of a crucial sea change in literary studies. Second, another major shift occurred around 1950. Here technology also played a leading role: the advent of television and vinyl recordings brought writers into people's homes in ways never before possible, thereby solidifying the celebrity status of authors. The influence of the educational establishment in post war society is also important; university syllabi, designed by writers and critics whose vested interests were served through creating a canon that fit their definitions of what "great" literature was, created a publishing demand for these very writers. The result was a wholesale shift away from the writers who were prominent at the beginning of the century towards those who were notable for their marginal status in the 1900-1920 period.

Our specific concern will be to create a database of English language published material on each of the six writers in the period 1900-1950. We are not concerned with "creative" or "imaginative" literature written by the six, but with reviews, newspaper articles, magazine or periodical press articles (critical or scholarly) either written by the six or on the six. We will enter/scan all items into the database thereby creating a very large data set of information. The database will also house the bibliographical information on each item we obtain. This information will then be mounted on the Simon Fraser University Library's Electronic Document Centre where it will be available for use by other scholars. This part of the project will require that the text already scanned into the database be coded—using either HTML or XML—so that it can be made available on the web.

The next few sections describe how we process the texts once they have been scanned, and how we are extracting information from the texts that we hope will shed light on how literary reputation is built or destroyed.

3. Sentiment Classification: Semantic Orientation of Words

The problem of extracting the semantic orientation (SO) of a text (i.e., whether the text is positive or negative towards a particular subject matter) often takes as a starting point the problem of determining semantic orientation for individual words. The hypothesis is that, given the SO of relevant words in a text, we can determine the SO for the entire text. We will see later that this is not the whole or the only story. However, if we assume that SO for individual words is an important part of the problem, then we need lists of words with their corresponding SO, since such information is not typically contained in a traditional dictionary. The expressions "semantic orientation", "sentiment", and "opinion" are used in this paper to refer to the subjective evaluation conveyed by a word, a phrase, a sentence, or an entire text.

One approach is to manually compile a list of words that are known to express sentiment, and annotate them according to whether the sentiment is positive or negative. One such list is the one contained in the General Inquirer, a content analysis program (Stone, 1997; Stone et al., 1966). The General Inquirer contains lists of words, classified according to specific categories, such as "self-reference", "strong", "active", or abstract concepts (words relating to objects, places, institutions, etc.). Of interest to sentiment detection are two tags that indicate whether the word is positive or negative. These have been used to determine whether the majority of words in a text are either positive or negative.

Whitelaw et al. (2005) use a semi-automatic method to create a dictionary of words that express appraisal. Appraisal is a functional framework for describing evaluation in text: how personal feelings, judgement

about other people, and appreciation of objects and art are expressed (Martin & White, 2005; White, 2003). Whitelaw and colleagues compiled a list of appraisal words from the literature on appraisal, and extended it automatically by extracting synonyms and related words from WordNet (Fellbaum, 1998) and on-line thesauri. Other researchers have explored this avenue, extracting synonyms using either Pointwise Mutual Information (Turney, 2001) or Latent Semantic Analysis (Landauer & Dumais, 1997). It is unclear which method provides the best results; published accounts vary (Rapp, 2004; Turney, 2001). Word similarity may be another way of building dictionaries, starting from words whose SO we already know. For this purpose, WordNet is a valuable resource, since synonymy relations are already defined (Kamps et al., 2004). Esuli and Sebastiani (2005) also use synonyms, but they exploit the glosses of synonym words to classify the terms defined by the glosses.

Manual and semi-automatic methods, although highly accurate, are not ideal, given that it is time-consuming and labour-intensive to compile a list of all the words that can possibly express sentiment. Researchers have turned to automatic methods to "grow" dictionaries of sentiment words, out of a few words. Most research in this area has focused on adjectives. Adjectives convey much of the subjective content in a text, and a great deal of effort has been devoted to extracting SO for adjectives. Hatzivassiloglou & McKeown (1997) pioneered the extraction of SO by association, using coordination: the phrase *excellent and X* predicts that *X* will be a positive adjective. Turney (2002), and Turney & Littman (2002; 2003) used a similar method, but this time using the Web as corpus. In their method, the adjective *X* is positive if it appears mostly in the vicinity of other positive adjectives, not only in a coordinated phrase. "Vicinity" was defined using the NEAR operator in the Altavista search engine, which by default looked for words within ten words of each other. The contribution of Turney & Littman was to find a way to not only extract the sign (positive or negative) for any given adjective, but also to extract the strength of the SO, expressed in a number (e.g., 2.2 is more positive than 1.3). They use Pointwise Mutual Information (PMI) for that purpose. PMI calculations do not have to be limited to adjectives. In fact, Turney (2002) used two-word combinations that included Adjective+Noun, Adverb+Noun, and Adverb+Verb.

Pang et al. (2002) propose three different machine learning methods to extract the SO of adjectives. Their results are above a human-generated baseline, but the authors point out that discourse structure is necessary to detect and exploit the rhetorical devices used by the review authors. Machine Learning methods have also been applied to the whole problem, i.e., the classification of whole text as positive or negative, not just the classification of words (Bai et al., 2004; Gamon, 2004).

We have tested a number of methods for creating SO dictionaries, in part motivated by the fact that Altavista no longer allows searches with the NEAR operator (Taboada et al., 2006). We tested whether an AND search, where the two words can be found anywhere in a document, not just close to each other, would be useful for the task. The AND searches were performed using the Google search engine. Our results show that NEAR-created dictionaries outperform AND-based ones in the task of extracting sentiment. The tests were performed on reviews of movies and other consumer products. However, our results indicate that variability in the number of hits returned by Google (since it indexes a dynamic space) affects the quality of the dictionary.

In summary, SO dictionaries are actively being created. Although no perfect method for compiling one exists, progress is being made, and we can expect better methods and larger dictionaries in the near future.

4. Document Structure

Research in subjective evaluation of text has not taken into account text structure, most of it relying on the content of adjectives, such as *great* or *poor* (e.g., Turney, 2002). However, adjectives have different meanings according to their linguistic context, whether immediate: *a huge disaster* vs. *a huge success*, or more remote: *The movie is great, if you're looking for reasons to be depressed*. In the latter example, it is important to know that the positive evaluation (*the movie is great*) is hedged by a condition on it. Previous work on movie reviews has revealed a common argumentation device, whereby authors list a number of positive aspects, to end with a negative summary. Example (1) illustrates the strategy[1]: the author lists a number of positive qualities for the movie "The Last Samurai". He or she, however, finishes with a clear negative evaluation. The concession structure ("good in some aspects, but overall bad") is very common in reviews, especially those found on-line.

(1) [1] It could have been a great movie. [2] It could have been excellent, and to all the people who have forgotten about the older, greater movies before it, will think that as well. [3] It does have beautiful scenery, some of the best since Lord of the Rings. [4] The acting is well done, [5] and I really liked the son of the leader of the Samurai. [6] He was a likeable chap, [7] and I hated to see him die. [8] But, other than all that, this movie is nothing more than hidden rip-offs.

It is obvious that we need to understand the overall structure of the text, and especially the concessions and conditions that authors attach to their opinions. For that purpose, we need to parse the entire structure of the text. Discourse parsing is analogous to sentence parsing: elements of the text are tagged, and incorporated into a tree that captures the dependencies found in the text.

Discourse parsing in this project is based upon Rhetorical Structure Theory (Mann & Thompson, 1988). RST is one of the most successful theories of discourse structure, in part because it lends itself well to computational implementations: it has been used in parsing and natural language generation, and in text summarization. A rhetorical, or discourse, relation is one that holds between two non-overlapping text spans, called nucleus and

[1] From the website Epinions.com. The text is reproduced verbatim. We have only added unit numbers (in square brackets).

satellite. Some relations are also multinuclear, consisting of two spans that are equal in importance. The nucleus contains the most important information, whereas the satellite supports or enhances that information. Spans are typically clauses in their minimal composition, but they are also built incrementally, so that a span may consist of different clauses, with their own internal structure. Multinuclear relations are analogous to paratactic or coordinate structures, whereas nucleus-satellite relations resemble hypotactic or subordinate relations.

There are different types of relations, based on the type of information or intention expressed: Condition, Contrast, Concession, Cause, Background, etc. Rhetorical relations can be represented in the form of trees, which have the following properties: completeness, uniqueness, connectedness and adjacency. Trees represent contiguous text, and the tree schemas can be applied recursively, to represent an entire text of arbitrary length.

The whole text in Example (1) above can be captured in a single relation: spans 1-7 are the satellite (i.e., the subordinate or less important part) to the nucleus presented in 8. The overall relation is one of Concession, as shown in Figure 1. The arrow pointing from 1-7 to 8 indicates that 8 is the nucleus, the most important part in the Concession relation. Spans 1-7 have further internal structure, which we could also analyze using RST.

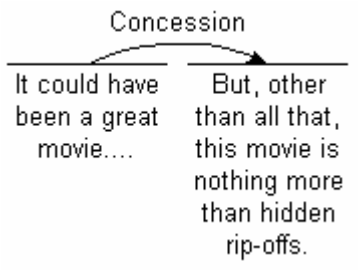

Figure 1. General structure for Example (1)

Unfortunately, a full discourse parser based on RST (or any other theory) does not exist yet. Soricut & Marcu (2003) created a sentence-level parser, trained on data from the RST Treebank (Carlson et al., 2002), a collection of articles from the Wall Street Journal. We have been testing this parser, which creates trees for individual sentences (but not for the full text). Our results are quite poor so far, probably due to the very different text genres. Current research aims to improve sentence-level parsing, and to create a corpus of manually-annotated reviews, in order to train a full whole-text parser.

The results of such parsing would help distinguish main from secondary parts of the text. There is a significant amount of research on how RST can be used to summarize text, exploiting the discourse structure to prune the less important parts (Marcu, 2000). Our plan is to use it for a dual purpose: (i) to pinpoint the most important parts of the text; and (ii) to calculate the aggregation of nuclei and satellites. In Example (1), that would mean, first of all, to identify spans 1-7 and span 8 as the main parts of the text, with span 8 as the nucleus of the relation between the two. In addition, the analysis tells us that the relation between those two spans of text is one of Concession. That means that there is a discrepancy in the situations, events or opinions expressed by each span. In the example, we see that the first part of the text contains a large number of positive words and phrases (*great, excellent, beautiful, some of the best, well done, likeable*), but the weight of those must be decreased in the final aggregation, because they are in the satellite of a Concession relation, and the most important part, what the author wanted to convey, is that the movie contains *hidden rip-offs*, a negative phrase.

RST classifies parts of a text according to a number of relations. The number and types of relations are often based on those proposed by Mann and Thompson (1988), but extensions and modifications are possible [2]. In addition, a higher-level classification could be imposed, dividing the text into stages, or parts, typically determined by the text genre (Eggins & Martin, 1997). For example, in present-day reviews of movies, there is usually a clear structure: introduction of the movie, plot, actors, director, background (e.g., other movies by the same director or cast), and evaluation. Segmenting each text into these stages would help identify the parts that contain an actual evaluation of the work, and not of the characters. RST has been integrated into genre analysis for other genres (Taboada, 2004a, 2004b), and could be easily integrated into the literary review genre and other genres in this project.

5. Processing Documents

The documents are first tagged with parts of speech (adjective, noun, verb). The words with subjective content are extracted and compared to a custom-built lexicon of words annotated with evaluation tags (i.e., positive for the word *excellent*, negative for the word *poor*). This electronic dictionary (or lexicon) assigns numeric values to words in the text (e.g., 5 for *outstanding*, -5 for *appalling*). The lexicon is being built partly automatically, based on the context of those words in documents found on the Internet (Turney & Littman, 2002). We are testing different methods of creating the dictionary (Taboada et al., 2006). We have already applied some of these methods to the problem of extracting sentiment from reviews about movies and consumer products (Taboada & Grieve, 2004). Our current dictionary contains 3,314 adjectives, whose semantic orientation was calculated using AND searches on Google. As described in our previous work, the values in the dictionary are normalized, so that 0 is the median value for the entire dictionary.

The final step in the process is to devise an algorithm to aggregate the negative and positive words in the document. We are currently using a weighted average of the adjectives in the text. Weights are assigned according to whether the adjective appears in the first, second, or last third of the text, as shown in Figure 2 (Taboada &

[2] Each relation in RST has a formal definition. Definitions and examples for the most common relations can be found on the RST website (Mann, 2005).

Grieve, 2004). The intuition behind these weights is that authors tend to summarize or repeat their opinions towards the end of the text. We also take negation into account, changing the sign of an adjective in the scope of a negating word (e.g., *not, no, nor, neither*). Negating words are considered within scope if they are found up to five words to the left of the adjective.

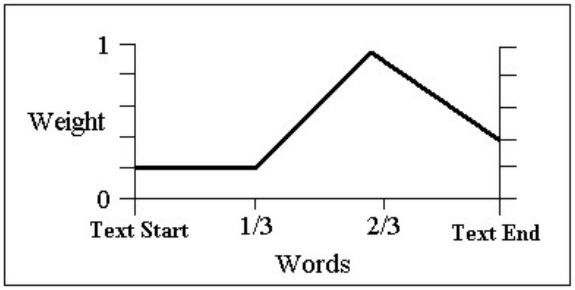

Figure 2: Weights given to adjectives

Future work involves a discourse analysis of the texts, to examine the types of patterns that signal the presence of subjective content; and a method to determine the contribution of different rhetorical relations to a text's sentiment. Words other than adjectives will also be considered, as long as they convey sentiment. The final goal of our project is to be able to determine what in a reviewer's text seems to influence the literary reputation of a particular author, and whether what reviewers say can be mapped to the author's reputation trajectory.

6. Two Examples

Since we are describing work in progress, we do not yet have large-scale quantitative results. In this section, we show a detailed analysis of two documents, one for each author, explaining what processing was carried out, and the current results.

The documents are reviews of (at the time) recently published works by the two authors. The review of John Galsworthy's plays (*A Bit o' Love, The Foundations* and *The Skin Game*) was published June 26, 1920, in the *Saturday Review* (Anonymous, 1920). The second document is a review of D.H. Lawrence's *The White Peacock*, published March 18, 1911, in *The Academy and Literature* (Anonymous, 1911). The Galsworthy text comments on the work of an established artist, and issues quite a damning criticism of his work. The text on Lawrence, on the other hand, is about an up-and-coming artist, who, up to that point, had been known only as a poet. The first one is 1,018 words long, whereas the D.H. Lawrence text contains 629 words.

6.1. Semantic Orientation for Adjectives

Space precludes a full examination of the entire texts. We will concentrate on some passages. The Galsworthy text starts with a simple statement: "For many years Mr. Galsworthy has been consistently overpraised." It ends with a summary of that opinion: "Mr. Galsworthy, in fact, remains the second-rate artist he always was." The entire text is organized around those two statements, with a lengthy elaboration of the first by way of a general criticism of Galsworthy's work (lack of creativity; he is ridden by ideas, but creates no real characters; his views are too present), and a specific example of how this is evident in the play *A Bit o' Love*.

The first process applied to the text (apart from normalization of punctuation and spacing) is the part of speech tagging. Each word is assigned a part of speech (noun, verb, adjective, determiner, etc.). Tagging is performed automatically, using Eric Brill's freely available tagger (Brill, 1995). After tagging, all words tagged as adjectives are extracted and their semantic orientation extracted from our dictionary. Example (2) shows in bold type the words that were tagged as adjectives in the first few sentences of the text, with the SO values according to the dictionary in square brackets.

(2) For many years Mr. Galsworthy has been consistently overpraised. His admirers, detecting in his **imaginative** [2.13] work—and particularly in his plays—the quality of **moral** [-2.06] earnestness, have taken him to their **susceptible** [0.03] hearts as one of the **supreme** [-0.41] artists of our time; but it is as a **creative** [4.001] artist, **pure** [-0.35] and **simple** [1.01], that he fails. He has many gifts, many qualities—**technical** [4.57] ability, imaginativeness, sympathy, experience of life, ideas, ideals; but the one **supreme** [-0.41], **essential** [2.95] gift—the ability to create living men and women working out their destinies in the grip of fate—is not his. He is ridden by his ideas, harried by his ideals; he has no spaciousness, no ease, no geniality; and his characters are invariably irritatingly **true** [0.65] to type and the instruments for their author's views on sociology, politics and what not.

One could disagree with some of the adjective values. They were calculated automatically, and according to their context in web pages indexed by Google(Taboada et al., 2006). What we would like to point out here is that many other words convey opinion: *earnestness, ability, imaginativeness* (all nouns), or *fails* (a verb). Note also that one of the most important words, *overpraised*, is not tagged as an adjective. The tagger interpreted it as a verb (a past participle), which is, strictly speaking, correct in this case. It is also clear from the example that the context, and the person or object being evaluated, are quite relevant. For instance, the word *susceptible* is applied to Galsworthy's admirers; it does not necessarily reflect upon him or his work; *pure* and *simple* are used to emphasize a statement and do not refer to any entity in the text. Finally, the word *creative* (one of the most positive in this fragment) is negated through the verb *fails*. All of those aspects (words beyond adjectives, context and sentence topic) are part of our future work.

Applying this same method to the entire text, we extracted all the adjectives, and produced a weighted average, with the final number of 0.19. This is a positive number, but quite close to the 0 level, reflecting the fact that many of the statements in the text were negative in nature.

The same procedure was carried out on the Lawrence text, of which we show a portion in (3). This text starts with a contrast between Lawrence's previous work as a poet, and what the reviewer sees as a promising novelist career.

It describes *The White Peacock* in detail, and concludes by saying that "…he has given us a book of considerable achievement and infinite promise." Example (3) shows the adjectives detected by the tagger. As with the Galsworthy text, some crucial words are missing, such as the verbs *surprises* and *charms,* and *disillusioned,* which was tagged as a past participle. The final number for the entire text was 0.25, a slightly more positive value than for the Galsworthy text.

(3) Hitherto we have only known Mr. D. H. Lawrence as being one of the many **interesting** [1.411] poets discovered by the English Review. Henceforth we shall certainly know him as the author of "The White Peacock," for it is beyond all argument an **admirable** [0.58] and **astonishing** [0.38] piece of work. We use the word "**astonishing**" [0.38] advisedly, for, like most **new** [3.59] books of **uncommon** [0.49] merit, "The White Peacock" surprises even while it charms. There are pages in it that made the **present** [1.01] reviewer, a **sophisticated** [1.62] and disillusioned reader of novels, lay down the book and rub his eyes in wonder at the author's individuality and courage.

6.2. Rhetorical parsing

The texts are next processed through a rhetorical, or discourse parser. As explained in Section 4, there is no available parser for entire texts. The only existing parser (Soricut & Marcu, 2003) is one that analyzes individual sentences, classifying their parts (main and subordinate clauses, clausal adjuncts and other clausal components) into nuclei and satellites, and then defining the type of relation between those. The parser was designed for newspaper articles, and does not work as well for these texts. Future work involves adapting it to our purposes. Let us examine, however, its current output.

The sentences in Example (2) were segmented. The first one is a simple sentence, and did not undergo further segmentation. The second sentence is quite complex, and was divided into 6 spans, as shown in Example (4), with span numbers in square brackets. The structure of the text, according to the parser, is displayed in Figure 3.

(4) [1] His admirers, [2] detecting in his imaginative work [3] —and particularly in his plays—the quality of moral earnestness, [4] have taken him to their susceptible hearts as one of the supreme artists of our time; [5] but it is as a creative artist, pure and simple, [6] that he fails.

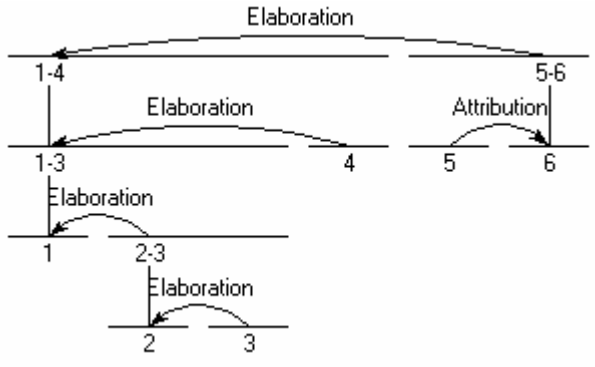

Figure 3. Rhetorical structure of one sentence

There are quite a few problems with the analysis. Its main failure is that the relation between the two main parts is too abstractly captured as an Elaboration relation, whereas a Contrast relation would be more appropriate, rephrased as: "his admirers think of him as creative; he fails as a creative author." The segmentation itself is problematic, especially around the parenthetical remarks between dashes.

We hope that the example is sufficient to illustrate the type of analysis that we want to perform, even though the results are far from perfect at this point. The important aspect of this analysis is that it identifies nuclei and satellites in the text. As we mentioned in Section 4, we plan to use this analysis for two purposes: to extract nuclei, and to aggregate the semantic orientation of individual spans according to the relation that joins them. In the example, there are quite a few elaboration relations. The semantic orientation of the words in each of those spans (1-4) can be simply added, since they are all contributing to the same idea. However, the contrast between 1-4 and 5-6 cannot be simply aggregated.

At the present time, we are not using relations to aggregate (given the fact that the parser does not yet capture them accurately). We are extracting the nuclei in the text, and calculating semantic orientation for those. For the text in Example (4), the nuclei are the fragments show in (5).

(5) [1] His admirers
[4] have taken him to their susceptible hearts as one of the supreme artists of our time;
[6] that he fails.

Nuclei for the entire text are extracted, and then the semantic orientation calculation is performed again, this time using adjectives found only in the nuclei. The Galsworthy text goes down in overall semantic orientation to -0.01. This probably reflects the fact that many of the positive adjectives are found in the satellites, or les important parts of the text. However, the same method applied to the Lawrence text yields an overall semantic orientation of 0.14, lower than the original 0.25. Such number is not an accurate reflection of the semantic orientation in the Lawrence text, since it is a generally positive review.

As is obvious from these two examples, our current system requires much further development. We are in the process of error-checking and improving each of the components, from the tagger to the adjective list (including other words than adjectives). The rhetorical parser is a very important part of that effort. We believe it can be made more efficient by improving the segmentation, and training it on examples drawn from our corpus.

7. Conclusions

This paper describes the initial stages of a project tracking the literary reputations of six authors between 1900 and 1950, and the applicability of existing techniques for extracting sentiment from texts that

discuss and criticize these authors.

One of the techniques for calculating sentiment and semantic orientation that has been developed is the analysis of adjectives from the text. This can give useful results but is limited by the size and accuracy of the list of adjectives used, the accuracy of the algorithm used to identify adjectives, the ability of the algorithm to recognize the context in which the adjective appears (including the presence of negating elements and where the adjective appears in the text), the contribution to the sentiment of the text by words of other parts of speech, and the overall discourse structure of the text. Each of these limitations suggests fruitful avenues of research.

We are engaged in developing algorithms for automatically developing adjective dictionaries. Future research will expand this effort to include semantic orientation dictionaries for nouns and verbs as well. As these are developed, algorithms for integrating their contribution to the orientation of the text as a whole can be investigated.

An accurate identification of semantic orientation requires analysis of units larger than individual words; it requires understanding of the context in which those words appear. To this end, we intend to use Rhetorical Structure Theory to impose on the text a structure that indicates the relationships among its rhetorical units. In particular, we want to distinguish units that are nuclei from those that are satellites so that their respective contributions can be appropriately calculated.

Finally, since the overall structure of a text is often correlated with the genre of the text, we must often be sensitive to the bias that machine learning techniques can inadvertently bring. Freely available texts such as newspapers that often provide the corpus for machine learning algorithms have a consistent structure that is different from the critical reviews that we are analyzing.

8. Acknowledgments

This project is funded through a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and an institutional SSHRC grant from Simon Fraser University. Thanks to Katia Dilkina, Jack Grieve, Caroline Anthony and Kimberly Voll for creating some of the tools used here, and Robert Outtrim and Lee Simons for collecting, scanning and preparing the documents.

9. References

Anonymous. (1911). Review of D. H. Lawrence's 'The White Peacock'. The Academy and Literature, March 11, 328.

Anonymous. (1920). Ridden by ideas. Review of John Galsworthy's 'Plays: Fourth Series'. Saturday Review, 129, 590.

Bai, X., Padman, R. & Airoldi, E. (2004). Sentiment Extraction from Unstructured Text Using Tabu Search-Enhanced Markov Blanket (Technical Report CMU-ISRI-04-127). Pittsburgh: Carnegie Mellon University.

Brill, E. (1995). Transformation-based error-driven learning and Natural Language Processing. Computational Linguistics, 21(4), 543-565.

Carlson, L., Marcu, D. & Okurowski, M.E. (2002). RST Discourse Treebank [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Eggins, S. & Martin, J.R. (1997). Genres and registers of discourse. In T.A.v. Dijk (Ed.), Discourse as Structure and Process. Discourse Studies: A Multidisciplinary Introduction (pp. 230-256). London: Sage.

Esuli, A. & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany.

Fellbaum, C. (Ed.). (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fromm, H. (1991). Academic Capitalism and Literary Value. Athens: University of Georgia Press.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of COLING 2004 (pp. 841-847). Geneva, Switzerland.

Gordon, L. (1977). Eliot's Early Years. Oxford: Oxford University Press.

Guillory, J. (1993). Cultural Capital: The Problem of Literary Canon Formation. Chicago: University of Chicago Press.

Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of 35th Meeting of the Association for Computational Linguistics (pp. 174-181). Madrid, Spain.

Herrnstein Smith, B. (1988). Contingencies of Value. Harvard: Harvard University Press.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004). Seattle, WA.

Kamps, J., Marx, M., Mokken, R.J. & de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (pp. 1115-1118). Lisbon, Portugal.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

Lecker, R. (Ed.). (1991). Canadian Canons: Essays in Literary Value. Toronto: University of Toronto Press.

Mann, W.C. (2005). RST Web Site, from http://www.sfu.ca/rst

Mann, W.C. & Thompson, S.A. (1988). Rhetorical

Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243-281.

Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization. Cambridge, Mass: MIT Press.

Martin, J.R. & White, P. (2005). The Language of Evaluation. New York: Palgrave.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using Machine Learning techniques. Proceedings of Conference on Empirical Methods in NLP (pp. 79-86).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal.

Remplin, C. (1995). Feminism and the Politics of Literary Reputation. Lawrence: University of Kansas Press.

Soricut, R. & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'03). Edmonton, Canada.

Stone, P.J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), Text Analysis for the Social Sciences. Mahwah, NJ: Lawrence Erlbaum.

Stone, P.J., Dunphy, D.C., Smith, M.S. & Ogilvie, D.M. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press.

Taboada, M. (2004a). Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish. Amsterdam and Philadelphia: John Benjamins.

Taboada, M. (2004b). The genre structure of bulletin board messages. Text Technology, 13(2), 55-82.

Taboada, M., Anthony, C. & Voll, K. (2006). Creating semantic orientation dictionaries. Proceedings of 5th International Conference on Language Resources and Evaluation (LREC) (to appear). Genoa, Italy.

Taboada, M. & Grieve, J. (2004). Analyzing appraisal automatically. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07) (pp. 158-161). Stanford University, CA.

Tuchman, G. & Fortin, N.E. (1989). Edging Women Out: Victorian Novelists, Publishers, and Social Change. New Haven: Yale University Press.

Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning (ECML-2001). Freiburg, Germany.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of 40th Meeting of the Association for Computational Linguistics (pp. 417-424).

Turney, P. & Littman, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus (No. ERB-1094, NRC #44929): National Research Council of Canada.

Turney, P. & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4), 315-346.

White, P.R.R. (2003). An Introductory Course in Appraisal Analysis, from http://www.grammatics.com/appraisal

Whitelaw, C., Garg, N. & Argamon, S. (2005). Using Appraisal groups for sentiment analysis. Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005) (pp. 625-631). Bremen, Germany.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning subjective language. Computational Linguistics, 30(3), 277-308.

Narrative Models: Narratology Meets Artificial Intelligence

Pablo Gervás[1], Birte Lönneker-Rodman[2], Jan Christoph Meister[3], Federico Peinado[1]

[1] Universidad Complutense de Madrid
Natural Interaction Based on Language
Facultad de Informática
28040 Madrid
Spain
fpeinado@fdi.ucm.es, pgervas@sip.ucm.es

[2] University of Hamburg
Narratology Research Group
Institut für Germanistik II
Von-Melle-Park 6
20146 Hamburg
Germany
birte.loenneker@uni-hamburg.de

[3] Ludwig-Maximilians-University
Institut für Deutsche Philologie
Schellingstraße 3
80539 München
Germany
jan.c.meister@germanistik.uni-muenchen.de

Abstract

This paper reports on research cooperation on narrative models in the framework of automated Story Generation. Within this framework, narrative models in Artificial Intelligence (AI) and the Humanities are considered both from the point of view of AI and the point of view of the Humanities. In order to provide other researchers, especially those interested in Computational Literary Analysis, with insights from literary narrative generation, existing Story Generation systems are reviewed and their underlying models of narrative are discussed. The existing gap between narrative models in different disciplines is analysed. We conclude that a methodological combination of description, analysis and generation holds the potential for a mutually beneficial qualitative breakthrough in research on narrative models.

1. Introduction

This paper reports on research cooperation on narrative models in the framework of automated Story Generation. The cooperation involves the *Story Generator Algorithms* project conducted at the Universities of Hamburg and Munich, in close association with the *Narratology Research Group* at the University of Hamburg, and the TSTL initiative (The Story Telling Laboratory) at the *Natural Interaction based on Language* research group of the Universidad Complutense de Madrid. The background of the authors ranges from literary studies over linguistics to computer science. The German group works on theoretical investigations of Story Generation and Computational Narratology. The central aim of this project is to evaluate the impact of automated systems on traditional research into narrative, also called Narratology. The Madrid group carries out research on the design and implementation of Story Generation applications, with a special emphasis on formal modelling of the knowledge that may be required.

On a small scale, the two teams started working towards an aim similar to that of the *Workshop Towards Computational Models of Literary Analysis*, and tried to bring together their today still relatively independent research enterprises. With respect to the topic of this workshop, we would therefore like to bring forward two points:

1. although narrative analysis and generation necessarily use different techniques in practice, they can share the abstract models underlying any theoretical and practical research on narrative;
2. cooperation across the borders of our respective scientific disciplines shows that the challenges and obstacles encountered in both computational generation and computational analysis of narratives are closely related to conceptual key problems discussed in Narratology.

The main purpose of this paper is to provide the emerging community of Computational Literary Analysis with insights from narrative generation and Narratology, and to identify potential common topics based on a discussion of narrative models. The potential offered by adopting this interdisciplinary perspective on narrative phenomena has been pointed out by Ryan (1991). Like Ryan, we are convinced that different communities dealing with narrative models can learn from each other, and that efforts can be joined, but we are also aware of the fact that communication problems might arise. This concern is reflected in the structure of the paper: Sections 2 and 3 deal with narrative models in Artificial Intelligence (AI) and in the Humanities, seen from a Humanities perspective. They are mirrored by Sections 4 to 5, where the same areas are discussed from an AI viewpoint. The paper is rounded up by a conclusion in Section 6.

2. Narrative Models in AI – Seen from the Humanities

2.1. Main Approaches and Inspirations

Artificial Intelligence uses two techniques in Story Generation: planning/problem solving, and production grammars. Specific rules used in their algorithms might be influenced by insights from literary studies or other fields (e.g. psychology of reading and writing).

Each Story Generator pretty much relies on only one work in narrative theory, if at all. For example, the idea of implementing the generator MINSTREL (Turner, 1994) ultimately goes back to Vladimir Propp's *Morphology of the Folktale* (1968), which was first published in Russian in 1927. Turner was intrigued by Propp's "morphology", i.e., "a description of the tale according to its component parts and the relationship of these components to each other and to the whole" (Propp, 1968:19). The general description of fairy tales, derived by Propp from a corpus of 100 Russian tales, can be interpreted as a story grammar (Turner, 1994:1–2). When Propp invented his fairy tale morphology, the core idea underlying his description was that of character functions that allow to abstract from concrete acts performed by individual characters. Incidentally, this primacy of action over characters had already been proposed in Aristotle's Poetics. Propp took this idea further and formulated his own findings as follows:

> "1. Functions of characters serve as stable, constant elements in a tale, independent of how and by whom they are fulfilled. They constitute the fundamental components of a tale.
> 2. The number of functions known to the fairy tale is limited." (Propp, 1968:21)

This allows Propp to define such generic character functions as AN INTERDICTION IS ADDRESSED TO THE HERO or THE VILLAIN RECEIVES INFORMATION ABOUT HIS VICTIM, which abstract from individual details of the actions they subsume (e.g., kind of interdiction, way of uttering it, name and nature of hero, etc.). Based on his corpus, Propp worked out a formula that describes all possible combinations and sequences of character functions in a fairy tale. Directly or indirectly, Propp's work inspired numerous Story Generators and interactive narrative systems (cf. Section 4 below). His ideas are easily accessible, but what his *Morphology* describes are really only some of the principles of story structure, without any relation to aesthetic values or effects, discourse organization, or surface representation in natural language. Therefore, his ideas are usually combined with other approaches in implemented Story Generators, or even considered only as a starting point, but without actual relevance for the implemented product, as in MINSTREL: "I did eventually write a computer program that tells stories. But [...] Propp's intriguing little grammar was nowhere to be seen." (Turner, 1994:2).

Propp is also mentioned as a precursor, or as the "primogenitor" of story grammars, by (Bringsjord & Ferrucci, 2000:154). These authors use story grammars in the Thorndyke (1977) notation for formalizing the knowledge of their Story Generator BRUTUS. A well-known story grammar similar to Thorndyke's is the one developed by Rumelhart (1975) with the aim of serving as a basis for a (cognitive) theory of story summarization (see. Figure 1).

(1) Story → Setting + Episode
(2) Setting → (State)*
(3) Episode → Event + Reaction
(4) Event → {Episode|Change-of-state|Action|Event + Event}
(5) Reaction → Internal Response + Overt Responsee
(6) Internal Response → {Emotion|Desire}
(7) Overt Response → {Action|(Attempt)*}
(8) Attempt → Plan + Application
(9) Application → (Preaction)* + Action + Consequence
(10) Preaction → Subgoal + (Attempt)*
(11) Consequence → {Reaction|Event}

Figure 1: Rules of Rumelhart's (1975) Story Grammar

Very few fully implemented stand-alone Story Generators take an approach completely different from that of story grammars. Instead of grammars, MINSTREL uses problem solving in the form of case-based reasoning and introduces the meta-level author goals *theme*, *drama*, *consistency*, and *presentation* (Turner, 1994). MEXICA (Pérez y Pérez & Sharples, 2001), on the other hand, is a Story Generator influenced by a psychological account of creative writing, the so-called cycle of cognitive engagement and reflection (Sharples, 1999). Instead of planning a story towards an explicit goal, MEXICA starts with an initial action around which it builds more and more actions, referring and comparing to a corpus of *previous stories*. The creation process switches between engagement and reflection: during engagement, actions are selected. The reflection stage checks the story (fragment) for coherence and, if necessary, introduces more actions to fulfill all preconditions of the previously retrieved actions. Also, the interestingness of a story is ensured by requiring it to display certain features: especially, it must show a given pattern of tension, which is also calculated based on the previous stories.

2.2. General Strengths and Weaknesses

This subsection is further subdivided into two parts, according to the classical narratological subdivision of narratives into two representational domains. These domains can be referred to by the French terms *histoire* ("story", "content", or "what is told") and *discours* ("text", "presentation", or "how it is told"). Subsection 2.2.1. presents strengths and weaknesses of AI systems concerning the *histoire* domain of narrative; AI system performance in the *discours* domain is then discussed in Subsection 2.2.2.

2.2.1. The *histoire* domain

From the examples introduced in Subsection 2.1 above, it becomes obvious that in Story Generation, much effort is spent on designing a model of the narrative world.

> "To a certain extent a story is a model of a tiny world, peopled with story characters, natural forces, locations, and inanimate objects. To understand these things and their interrelationships requires a tremendous amount of knowledge that humans take for granted." (Turner, 1994:4)

As far as the narrated world is concerned, MINSTREL has to have detailed knowledge about actions ("acts"), states and beliefs, as well as character goals, and the relationships between them. The set of relationships includes, for example, *motivation* (a state can motivate a goal) and *evidence* (a state can also be an evidence for a belief). Turner reports that MINSTREL's knowledge about the narrated world and its case base correspond to "one or two short stories about King Arthur" (Turner, 1994:8). Accordingly, MINSTREL's world comprises a class hierarchy of "Story Things" containing genre-typical classes such as LANCE, SWORD, HERMIT, and DRAGON (Turner, 1994:50). Whether actions, states and beliefs are equally hierarchically structured is not directly clear from Turner's book. They seem, however, to show a very flat organization: for example, all actions are direct children of the class ACT.

Given the present-day existence of large ontologies (e.g., SUMO[1]) and lexical semantic databases (e.g., WordNet[2]), the knowledge MINSTREL and other Story Generators dispose of seems very restricted. This can be partly explained by the fact that the Story Generation systems did not have any knowledge beforehand; the entire encoding was done from scratch, in a representation format that was suitable for the individual formalisms and procedures of the system.

Probably the most remarkable achievement of knowledge encoding in MINSTREL and MEXICA is the

[1] http://ontology.teknowledge.com/ [10 April, 2006]
[2] http://wordnet.princeton.edu/ [10 April, 2006]

representation of preconditions and consequences of action classes, or actions in the case-base (corpus). Their existence makes it possible to flexibly combine states and actions, instantiated by characters and possibly other "Story-Things" filling their slots, and to control whether the invented combination is possible according to the laws of the narrated world.

On the other hand, this achievement also shows why large-scale resources would represent a semantic "overkill" for these systems, or – to put it the other way round – a relational scarcity. It is currently simply impossible to create a case-base or a corpus of previous stories in the abstract representation format needed by the systems, which would illustrate *general* knowledge about concepts such as LOVE, REVENGE, and ANGER, not restricted to a small domain (for example, the narrative world King Arthur lives in, the presumed world of the old inhabitants of Mexico, etc.). If Computational Literary Analysis ever arrives at the stage of advanced Story Understanding, capable of turning natural language stories into abstract semantic representation, also Story Generation might become a more prosperous field. Still, even the Story Generation community alone is far from having a common representation format, currently preventing such knowledge exchange.

As far as the *histoire* domain is concerned, we believe that the Humanities can learn something from Story Generation. The necessarily clear-cut definitions in the systems allow their designers to "grasp" such notions as EVENT vs. EXISTENT, or the causal relations *effect* and *motivation* (cf. Chatman, 1978), which refer to phenomena within the *histoire* domain. If the Story Generation system works, the designer has achieved – among other things – one of the possible consistent representations of this narrative domain.

2.2.2. The *discours* domain

Narrative models used in Story Generation are much less developed where the *discours* domain, the way of telling, is concerned. Usually, when a Story Generator has created an abstract representation of a story, it sends it directly to a front-end that generates natural language text. In BRUTUS, on the other hand, story grammars and natural language grammars are closely intertwined; actually, a high-level story grammar is broken down into paragraph grammars, which are broken down into sentences types or "literary augmented sentence grammars" (Bringsjord & Ferrucci, 2000:194). In other words, the language generation process used in BRUTUS is "all about choosing words" (and nothing more) because "the story outline is a map identifying the sequence of sentence types that will be used to tell the story" (Bringsjord & Ferrucci, 2000:196). Words are grouped into classes according to certain features, including their function in literary narrative, and the sentence grammar indicates from which group of words the Natural Language Generation (NLG) module is allowed to choose. For example, the nouns *brick*, *greens*, and *youth* are classified as ICONIC FEATURES of *university*, allowing for the production of Example (1).

> (1) Dave Striver loved the university. He loved its ivy-covered clocktowers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth.

Another example is the linking of modifiers to nouns by the relation BIZARRO-MODIFIER. Using this relation, *bleeding* has been encoded as a modifier of the nouns *sun*, *plants*, *clothes*, *tombs*, and *eyes*. This literary-linguistic knowledge is used in the production of Example (2).

> (2) Hart's eyes were like big bleeding suns.

Whereas the technique used in BRUTUS illustrates some stylistic devices such as analogy or "the bizarre", used especially in literary narrative (as opposed to factual narrative), the actual NLG process reminds much more of template filling than of full-fledged NLG (Reiter & Dale, 2000) with its document structuring, microplanning, and surface realization phases.

The use of inflexible techniques for Natural language rendering of automatically generated narratives might as well be due to the fact that very few attempts exist to make Natural Language Generators fit for (literary) narrative input. The only Natural Language Generator that explicitly aims at this goal is STORYBOOK (Callaway & Lester, 2002). However, STORYBOOK uses a proprietary input representation, the so-called *narrative stream* format, and, to our knowledge, there are no interfaces to the output of implemented Story Generators. The input to STORYBOOK, then, is mainly encoded by hand.

Narrative discourse techniques such as large-scale ellipsis, flashback, repetition, summary, or changes in perspective are not used explicitly or purposefully in Story Generation. In our research, we have not yet encountered any system that would include a narrative discourse middleware able to produce variation at this stage. The architecture of a narratologically enhanced generator, which would be more aware of *discours* phenomena of narrative, is sketched in (Lönneker, 2005). The same paper also contains an example of how the discourse structurer might handle the phenomenon of "embedded" narratives.

3. Narrative Models in the Humanities – A Self-Assessment

Within the Humanities, different disciplines have developed narrative models. Our review concentrates on those proposed by linguists and literary scholars because in the history of Narratology as a discipline, linguistic and literary approaches have been more or less intertwined.

3.1. Linguistics

In linguistics, text-grammar-based models of narrative were popular in the 1970s, under the influence of Artifical Intelligence (cf. e.g. van Dijk, 1972; van Dijk, 1980). Whether or not they might be an adequate model of the cognitive processing of narratives has been discussed at length in (Wilensky, 1983). In our opinion, well-defined discourse relations that co-ordinate or sub-ordinate text segments, which have been proposed by Longacre (1983) and Mann & Thompson (1988), among others, might be a better choice. Discourse relations can be used to build a text representation either top down or bottom up, and are currently popular in computational linguistic approaches to discourse.

3.2. Literary Studies

In literary studies, encompassing theories of the general structure of narrative have been proposed only by a few scholars. In the early days of Narratology, which was born under the influence of structuralism and formalism, aspects of both *histoire* and *discours* domain of narrative

were treated. Sometimes, the textual form – a part of the *discours* domain, or the surface representation of "narrative units" (Todorov, 1969:16) – is even considered irrelevant, as already in Propp's work. For example, (Todorov, 1969) develops an inventory of relations between *histoire* units, parts of which are reproduced in Figure 2.

(I) Temporal relations	
1. Emphasis:	a(X) + ... + a(X)
2. Inversion:	a(X) + ... + ¬a(X)
[...]	
(II) Causal relations	
1. Modification	[...]
2. Desire	[...]
3. Motivation	a(X) ⇒ PROP
4. Result	PROP ⇒ a(X)
5. Punishment	b(X) ⇒ c(Y,X)

Figure 2: Relations between two propositions (PROP), based on Todorov (1969)

Later, and especially with the broad reception of the – still structuralist – publications by Gérard Genette, the focus shifted towards *discours*, including many aspects of surface (re-)presentation in natural language. In this approach, an investigation of the underlying *histoire* is undertaken only in view of a better analysis and explanation of *discours* phenomena. Since the 1990s, cognitive approaches to Narratology have been proposed, but – so far – none of the sketched models is as encompassing and as widely recognized as those developed by the structuralist generation.

3.3. Problems for Story Generation

Working out which of the above mentioned models of narrative could be useful in practical Story Generation, we noticed a number of problems. Some of them are presented in the remainder of this Subsection. After an introductory general remark (3.3.1.), two clusters of points are mentioned that hinder a fast and clear-cut formalization of narrative models, as discussed in literary theory: Terminological issues (3.3.2.) and (upwards or downwards) scalability (3.3.3.).

3.3.1. General

In general, most Humanities models of narrative contain formalizations only at very abstract levels, if at all. By formalizations, we mean here a representation in some logic language (e.g., predicate calculus) or other structured representation, including tables, graphs, etc. Indeed, most works dealing with narrative and not going back directly to the structuralist tradition are composed in "plain prose". Especially, there seems to be a tendency to apply formal notions to the abstract *histoire* level only. Phenomena at *discours* level that apply to the *structure* of discourse (e.g., discourse relations) are sometimes formalized in linguistics and are usually described in words only – sometimes accompanied by tables – by literary scholars (Genette, 1980). Where models are based on the *discours* (text) layer of a narrative or include it, genuine Humanities models usually lack formality, though their descriptions might offer a variety of authentic examples.

Perhaps one of the most vexing problems confronted by Humanists in their attempts at modelling narrative is the fact that the empirical notion of narrative per se is, at least in part, of a historical (rather than a universal) nature. A natural reader's processing of a symbolic representation as a 'narrative' is influenced by idiosyncratic choices, particularly where it comes to the decoding of semantic markers. For the natural reader, choices for 'making sense' of something as a narrative abound already at the fundamental level of *histoire*. The highly combinatory and constructive nature of processing information as *events* and *actions* has been demonstrated in (Meister, 2003). It remains to be seen whether the methodological restrictions inherent in computational approaches will allow for the design of models that can truly capture this level of empirical complexity, or whether an idealized notion of 'narrative' has to be used as a frame of reference.

3.3.2. Terminology and Granularity

Many critics have commented on parts of the literary narrative models mentioned in Section 3.2. above, and similar ones. The criticism led to further developments of the models as well as to proposals of contrasting terminology. For example, regarding narrative *discours* techniques, we still witness undecided discussions about terminology, but we also encounter extremely fine-grained terminological subdistinctions. The latter case can be observed with the numerous subtypes of anachronies (such as flashback) introduced by Genette (1980) and further developed by Ireland (2001).

Nevertheless, there is also still a certain degree of remaining fuzziness in the models. Often, this is detected only when formalising them. An example is the question whether an inner narration ("embedded" narration) can be told in indirect speech, or whether it is necessarily presented in direct speech (cf. Lönneker, 2005).

Even terminological gaps do still exist. For example, presenting one and the same event several times, but in different ways (for example, from different points of view), is called repeating narrative by Genette (1980). But the technique of repeatedly presenting *variants* of an event does not seem to have a name, although it can be an important trait of certain narratives, including films.

3.3.3. Scalability

Linguistic models of discourse are usually applied to texts containing a couple of sentences, up to several pages. Their appropriateness can be directly tested on such relatively short texts. Whether they scale to long works, for example to a novel, is in general not known because they have not been applied to these texts, which would involve a huge amount of work.

Models of literary texts, on the other hand, deal with texts that span over hundreds or thousands of pages; their explanations resort either to mini-scale examples (snippets of real works) or to summaries. Today, it is nearly impossible to empirically test models of literary texts because it is difficult to map them onto the text of actual works, or to "down-scale" them. Generation and analysis could work together here in order to see what representation of the texts is necessary to meet the models, or vice versa.

In order to illustrate the difference between linguistic and literary approaches, consider the example of a flashback. Linguistically seen, it might seem mandatory that this anachrony be indicated by surface markers such as conjunctions, adverbs, or a tense shift. However, in a large-scale work, it is possible that one or more entire chapters constitute a flashback to the surrounding text, and

that this relationship is indicated exclusively by contextual markers, including semantic knowledge and world knowledge (e.g., a character who was at first an old man reappears as a student in the flashback).

4. Narrative Models in the Humanities – Seen from AI

Most current efforts on narrative modelling in the Humanities have yet to permeate to AI research. Two possible lines of thought may be followed to study the situation: one is to consider which narrative models from the Humanities are being considered in AI, and another is to try to identify the reasons why AI researchers lack motivation for extending their exploration of narrative models in the Humanities.

4.1. Models under Consideration

Of the many theories of narrative developed in the Humanities, only a few have bridged the gap to become tools in the hands of AI researchers. Of these, Propp's *Morphology of the Folktale* (1968), mentioned in Section 2.1. above, is the most extended one, having been applied in several AI systems for Story Generation, including Automatic Novel Writing (Klein et al., 1973), Geist (Grasbon & Braun, 2001), OPIATE (Fairclough, 2004) or ProtoPropp (Díaz-Agudo, Gervás & Peinado, 2004). Propp's system of character functions as narrative modules has been exploited by AI researchers well beyond its intended purpose, both in terms of its use as a kind of grammar for generating new stories and in terms of its applicability to domains wildly different from the Russian folk tales from which it arose.

At a different level of detail, another favourite is the three-act restorative structure. This model, derived from Joseph Campbell's analysis of the structure of myths, is a dominant formula for structuring narrative in commercial cinema (Vogler, 1998). It has had a great impact on another branch of computer research which is related with narrative: the design of story-based video games. This type of game provides a skeletal plot for the player to follow. At an abstract level, the software of such a video game acts as a Story Generator: when the user has run the software to its completion – i.e. to the end of the game – a "story" has been generated, usually as a result of collaborative work between the software and the player. For various reasons arising from the interactive nature of these applications, the model proposed by Campbell for a *heroic quest* has been widely used. Heroic quests are well-suited to videogames because in them the player can discover the obstacles at the same time as the hero, and they tend to map progress through the quest onto progress through physical space, which is easy to model. Software of this kind usually has all the ingredients of a three-act restorative structure:

1. a central protagonist,
2. a conflict introduced in the first act,
3. a second act propelled by the hero's false resolution of this dilemma,
4. and a third act in which the dilemma is resolved once and for all.

More complex narrative models are considered in recent research efforts in interactive storytelling. For instance, some effort has been made to model works oriented to the film industry such as McKee (1997), which is used as inspiration underlying the technology of the interactive drama Façade (Mateas & Stern, 2003). McKee's model focuses on the interplay between characters and events – story events impact the characters, and the characters impact events. This interplay leads to a meaningful emotional experience for the audience. He considers a structure of narrative based on acts, sequences, scenes, and beats. Beats are seen as the atoms of this structure, and they consist of a *pair of action and reaction* by characters in the story. This provides a reference model for the way in which the characters in Façade interact with one another and the player, how they react emotionally, and how their actions affect the player.

Another source that is also being considered in AI is the work of Chatman (1978). This model constitutes a step up from the models of Propp or Campbell in the sense it considers a wider range of media, from literature to film. From the point of view of the AI researcher in search for a model, the greatest advantage of Chatman's approach is his effort to identify a common core of elementary artefacts involved in several approaches to narrative theory. Chatman studies the distinction between story and discourse (in the sense of *histoire* and *discours*, cf. Section 2.2. above), and he proposes ways of decomposing each of these domains into elementary units. His idea of structuring story and discourse in terms of *nuclei* and attached *satellites* provides a very good way of organising internally the knowledge entities that computational systems rely on for conceptual representation.

4.2. The Obstacles in Considering More Complex Models

Given the number of alternative theories that have not received this degree of attention, it may be interesting to consider what makes AI researchers opt for what are in context very simplistic models. Often, the most significant reason behind a particular choice of theory is that AI researchers find it easier to work with models that a previous researcher has already translated to AI jargon and applied in some previous computer program. Another important factor is that AI research in complex topics such as Story Generation usually applies a method of successive approximations, starting from the simplest possible model and exploring it until all its possibilities have been exhausted.

However, the most important obstacles can be found in the differences in purpose – when trying to model narrative – between Humanities and AI. In order to be applicable for AI research, a model of narrative must be capable of accounting for the elementary communication issues behind narrative. It must identify clearly the simplest basic elements with which the narrator operates. Definitions must be clear cut, and susceptible of computational treatment. They must not allow various possible interpretations.

In an effort to cover as much as possible of the infinite range of human expression, narrative theories in the Humanities move at a level of abstraction which is prohibitely expensive to represent in computational terms, and which precludes all possibilities of pragmatically efficient computation. Narrative models in the Humanities usually arise within a given school of thought. There is little consensus across different schools on what the basic elements

of narrative are – events?, motifs?, ... – and how they are defined. Different theories that agree on the importance of a given element may provide definitions for it that imply radically different design decisions from a computational point of view.

For the AI researcher looking for a model of narrative to use when developing a computational system, this situation presents various problems. On one hand, there are a large number of different theories. On the other hand, the set of ontological commitments on which each theory is based may not be explicit in the formulation, but implicit in the particular school of thought in which the theory arises. This information is crucial when deciding on the appropriateness of a given theory for a specific purpose, but it is usually unavailable to researchers without a complete narratological background.

Faced with this panorama, AI researchers gravitate towards narrative models which seem closest to their needs. These usually happen to be either very early attempts – such as Propp's – or models that focus on one very specific type of narrative – like Campbell's work on the hero's journey. This type of model may fulfill the requirements for developing computational solutions.

5. Narrative Models in AI – A Self-Assessment

In AI there is a long standing tradition in terms of research efforts in Story Generation. It started in the early days, with the same optimism and ingenuity that characterised early efforts at natural language processing and other simulations of human behaviour. Subsequent realization of the difficulties involved led to periods during which no research was undertaken on this area. But periodically the topic recovers strength. There was a big boom in the 1990s, around the concept of Narrative Intelligence (Mateas & Sengers, 1999). And there is a more recent effort concerned with the role of interaction and storytelling in the field of virtual environments (Mateas & Stern, 2003; Grasbon & Braun, 2001; Fairclough, 2004), applying these results to videogames, pedagogical applications, etc.

In order to treat computationally – or simply attempt to reproduce – a given phenomenon, the elements involved in it must be represented in some manner susceptible of computational treatment and a certain process or algorithm must be applied to it. In virtue of this, *every implemented Story Generator carries an implicit model of narrative*, irrespective of whether it is explicitly based on a given theoretical model of narrative. Such implicit models cover two different aspects. On one hand, they must provide some representation of stories, which can be interpreted as a particular model of what a story is. On the other hand, they must define a specific process for generating the story, which can be interpreted as a model of the actual process of Story Generation.

Bailey (1999) distinguishes between three different approaches to automated Story Generation:

1. **Author models.** Here, an attempt is made to model the way a human author goes about the task of creating a story. MINSTREL and MEXICA would be classed as examples of this approach.
2. **Story models.** They are based on an abstract representation of the story as a structural (or linguistic) artefact. Systems based on story grammars fall under this category.
3. **World models.** In these models, generating a story is seen as constructing a world governed by realistic rules and peopled with characters with individual goals. The story arises from recording how the characters go about achieving their goals. Tale-Spin (Meehan, 1977), the classic Story Generator inspired on Aesop's fables, operated in this way.

To this initial classification, Bailey adds his own (the fourth) approach, based on modelling the response that a story draws from a given reader.

Additionally, the fact that many of the new storytelling systems are based on interactive environments adds another dimension to the narrative. In these systems, the members of the audience become themselves characters in the story, so the role of authorship is progressively becoming distributed between the interactors and the designers. This may be considered as the fifth possible approach to model Story Generation.

Each type of system focuses on a different aspect of Story Generation, but *they must all provide implicit solutions to all other aspects* – however simple those solutions may be.

5.1. Representations of Stories in AI Systems

The narrative models currently in use in AI approaches to narrative generally present a simplistic approach to the representation of stories in several senses.

5.1.1. Linear versus Branching Stories

On one hand, they tend to consider a story as a linear sequence. This is true of the rendition of a story as text, but, conceptually, stories beyond the simplest joke have several branches whenever more than one character is doing something relevant to the story in different places at the same time. Additionally, the chronological order of events in the story (*histoire*) may be transgressed when a particular *discours* is generated for it. More elaborate models of narrative need to be contemplated to account for this complex nature of a branching and partially-ordered *histoire*, and the processes involved in converting it into a linear – possibly anachronical – discours.

5.1.2. The Role of Causality

World models concentrate on a concept of story that gives a central role to the causality relations between the events that make it. They tend to rely on planning methods to construct the stories from an initial description of the world and a set of goals to be achieved. Tale-Spin is an early example, but there is a flourishing school of Story Generation research (Mateas & Stern, 2003; Cavazza, Charles & Mead, 2002) still following this approach. However, the planning paradigm is biased towards producing plans in the shape of an inverted tree: a number of branches (causes) all converge towards a final goal (the result). This somehow improves on the linear conception of the fabula, but it is still very restricted. Real stories also include forward branching (each cause may lead to different effects on different characters), and they rarely have one single end point where the goal of the story can be said to be achieved. Most stories, in fact, have no single identifiable goal. The generation approach based on causality representation in the world model also restricts the

composition process to backward chaining from a desired goal towards a set of plausible causes. Whereas this may be the way in which some writers work, it is clear that the option of working forward from causes in search of their possible effects – usually applying principles of human nature to guide the way and explore possibilities – should also be considered as a possible model.

Some author models also consider causality as a fundamental aspect of story telling. MINSTREL, for instance, depends on planning techniques, but from the point of view of how an author plans the story that he is constructing. To account for the causality relations in the story, the system includes specific consistency goals.

5.1.3. Modelling the Reader

Bailey's (1999) approach is based on the idea that something is a story if and only if some reader identifies it as such when being exposed to it. This defines a story only in terms of a particular reader, but Bailey tries to abstract a general description of what makes all readers recognise something as a story. This requires having some way of modelling and/or measuring the reader's reaction to a story. As Bailey himself confesses in his paper, there is still a gap between existing work on this topic from the point of view of AI and the Humanities, in the sense that there is a large body of literature on the influence of narrative on the reader that has not been applied to AI research.

5.1.4. Representing Mental Images

Certain AI efforts at Story Generation – such as BRUTUS (Bringsjord & Ferrucci, 2000) – consider the important role of modelling the mental images being processed by all the participants in a story. This involves the mental image that the reader forms of the story – which it is feasible to model in a story telling system –, but also the mental images that characters have of one another, and the situations in which they find themselves.

Modelling mental images is subject to a problem of recursion – each character may have a mental image of the mental image that another character has of the mental image that… – which would need to be cut short at the very earliest approximation possible. The alternative of not modelling mental images at all is the simplest solution available, but it runs the risk of producing stories where characters exhibit autistic behaviours.

MEXICA (Pérez y Pérez & Sharples, 2001) takes a small step forward in this direction by considering the emotions of characters and the way in which their oscillations affect the perception of tension in the story.

5.2. Models of the Story Generation Process Implicit in AI Systems

Most existing work on Story Generation tends to focus on composing conceptual representations of stories, based on a given world – with a specific set of locations, characters and objects – which is to be told by some particular simple solution for rendering the concepts as text. However, a human author creating a new story actually works at least three different levels: he *creates a world* in which the story occurs (5.2.1.), he *imagines a story* in that world (5.2.2.), and he selects a particular *way of telling* the story that best presents it to the reader (5.2.3.).

5.2.1. Creating Worlds

The role of the setting in which a story occurs, and the nature and description of the characters that take part in it, is undoubtedly fundamental in human storytelling, and yet it has not been addressed by AI research in the field. Recent attempts have been made to provide story telling systems based on planning with a certain ability for modifying the initial story world that they operate upon (Riedl & Young, 2006), but their ability is restricted to changing the starting position of objects whose initial location has not been explicitly assigned by the user. Although this does improve the kind of narrative that can be generated, it is clearly still very far from the freedom that a human author exercises in making up his settings and characters.

5.2.2. Creating Stories

Most existing Story Generation systems focus on the task of building a story, taking as input a given description of the world and relying on simple natural language transcription modules to convert the story into text. In a sense, this would be consistent with considering Story Generators as content determination modules in a classic Natural Language Generation pipeline (Reiter & Dale, 2000), with subsequent stages accounting for discourse and sentence planning and surface realization.

5.2.3. Telling Stories

STORYBOOK (Callaway, 2002) is an exception in the sense that it focuses on the task of telling the story given a narrative stream already complete (see also Section 2.2.2. above). Callaway actually proposes a generic architecture for a storytelling system in which there is a prior module – a narrative planner – that generates the input (the narrative stream) for STORYBOOK. This narrative planner seems to be concerned with the task of creating a story, but nothing is said about the truly creative task of inventing a specific world in which the story is to take place.

6. Conclusion

Neither the under defined nor the over specific concepts developed in literary theory and Narratology seem good choices for AI formalizations. In the same vein the limited scope of predominantly descriptive linguistic models renders these unsatisfactory. Conversely, Artificial Intelligence approaches in Story Generation are generally based on a highly reductionist concept of 'story' which ignores the Humanities' disciplines insights into the complexity and dynamics of narrative. The reasons for the respective shortcomings of the proposed models, and the problems to adapt most of the models originating outside one's own field have been outlined in this paper.

In our view, a methodological combination of description, analysis and generation – in other words: an inter-disciplinary approach – holds the potential for a mutually beneficial qualitative breakthrough in research on Story Generation, and on narrative models in general. This inter-disciplinary approach might start by identifying those existing narrative models in the Humanities whose set of ontological commitments is better suited for the Story Generation task, and by searching for (or producing) computationally oriented implementations of these models.

7. Acknowledgements

This research is partly supported by DFG (German Research Foundation) grant ME 1546/2-1, and by the TIN2005-09382-C02-01 project funded by the Spanish Ministry of Education and Science.

8. References

Bailey, P. (1999). Searching for Storiness: Story-Generation from a Reader's Perspective. In *Working Notes of the Narrative Intelligence Symposium, AAAI Fall Symposium Series. (= Technical Report FS-99-01.)* Menlo Park, CA: AAAI Press, pp. 157–164.

Bringsjord, S. & Ferrucci, D. (2000). *Artificial Intelligence and Literary Creativity: Inside the mind of Brutus, a StoryTelling Machine.* Hillsdale, NJ: Lawrence Erlbaum.

Callaway, C. B. & Lester, J. C. (2002). Narrative Prose Generation. *Artificial Intelligence*, 139(2), pp. 213–252.

Campbell, J. (1968). *The Hero with a Thousand Faces.* 2nd edition. Princeton, NJ: Princeton University Press.

Cavazza, M., Charles, F. & Mead, S. J. (2002). Character-Based Interactive Storytelling. *IEEE Intelligent Systems*, 17 (Special Issue on AI in Interactive Entertainment), pp. 17–24.

Chatman, S. B. (1978). *Story and Discourse: Narrative Structure in Fiction and Film.* Ithaca, NY & London: Cornell U. P.

Díaz-Agudo, B, Gervás, P. & Peinado, F. (2004). A Case Based Reasoning Approach to Story Plot Generation. In P. Funk, P. A. González Calero (Eds.), *Advances in Case-Based Reasoning. Proceedings of the 7th European Conference on Case Based Reasoning. LNAI 3155.* Berlin & Heidelberg: Springer, pp. 142–156.

van Dijk, T. (1972). *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics.* The Hague & Paris: Mouton.

van Dijk, T. (1980). *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Hillsdale, NJ: Lawrence Erlbaum.

Fairclough, C. R. (2004). Story Games and the OPIATE System. Doctoral Thesis. University of Dublin, Trinity College.

Genette, G. (1980). *Narrative Discourse. An Essay in Method.* Ithaca, NY: Cornell U. P.

Grasbon, D. & Braun, N. (2001). A Morphological Approach to Interactive Storytelling. In M. Fleischmann, W. Strauss (Eds.), *Proceedings of Cast'01, Living in Mixed Realities. Special Issue of Netzspannung.org/Journal, The Magazine for Media Production and Intermedia Research*. Sankt Augustin: FhG-IMK, pp. 337–340.

Ireland, K. (2001). *The Sequential Dynamics of Narrative. Energies at the Margins of Fiction.* Madison, NJ: Fairleigh Dickinson U. P. et al.

Klein, S. et al. (1973). Automatic Novel Writing: A Status Report. Tech. Report 186. University of Wisconsin, Computer Science Department.

Longacre, R. E. (1983): *The Grammar of Discourse.* New York & London: Plenum Press.

Lönneker, B. (2005). Narratological Knowledge for Natural Language Generation. In *Proceedings of the 10th European Workshop on Natural Language Generation*, pp. 91–100.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical Structure Theory. Toward a Functional Theory of Text Organisation. *Text, an Interdiscipinary Journal for the Study of Discourse*, 8(3), pp. 243–281.

Mateas, M. & Sengers, P. (1999). Narrative Intelligence: An Introduction to the NI Symposium. In *Working Notes of the Narrative Intelligence Symposium, AAAI Fall Symposium Series. (= Technical Report FS-99-01.)* Menlo Park, CA: AAAI Press, pp. 1–10.

Mateas, M. & Stern, A. (2003): Façade: An Experiment in Building a Fully-Realized Interactive Drama. In *Game Developers Conference, Game Design track*, San Jose, CA. Available http://www.lcc.gatech.edu/~mateas/publications/MateasSternGDC03.pdf [11 April, 2006]

McKee, R. (1997). *Story. Substance, Structure, Style, and the Principles of Screenwriting.* New York: Regan Books/Harper Collins.

Meehan, J. R. (1977). TALE-SPIN, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence.* Cambridge, Mass., Morgan Kaufmann, pp. 91–98.

Meister, J. C. (2003). *Computing Action. A Narratological Approach.* Berlin & New York: de Gruyter.

Pérez y Pérez, R. & Sharples, M. (2001). MEXICA: a computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(2), pp. 119–139.

Prince, G. (1982). *Narratology. The Form and Functioning of Narrative.* Berlin et al.: Mouton.

Propp, V. (1968). *Morphology of the Folktale.* 2nd edition. Austin, TX: University of Texas Press.

Reiter, E. & Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge: Cambridge U. P.

Riedl, M. & Young, R. M. (2006) Story Planning as Exploratory Creativity. *New Generation Computing*, 24(3–4), Special Issue on Computational Creativity, in press.

Rumelhart, D. (1975). Notes on a Schema for Stories. In D. G. Bobrow (Ed.), *Representation and Understanding. Studies in Cognitive Science.* New York et al.: Academic Press, pp. 211–236.

Ryan, M.-L. (1991). *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington & Indianapolis: Indiana U. P.

Sharples, M. (1999). *How we write. Writing as creative design.* London: Routledge.

Thorndyke, P. W. (1977). Cognitive Structures in Comprehension and Memory of Narrative Discourse. *Cognitive Psycholoy*, 9(1), pp. 77–110.

Todorov, T. (1969). *Grammaire du Décaméron.* The Hague & Paris: Mouton.

Turner, S. R. (1994). *The Creative Process: A Computer Model of Storytelling.* Hillsdale, NJ: Lawrence Erlbaum.

Vogler, C. (1998). *The Writer's Journey: Mythic Structure for Storytellers and Screenwritiers.* 2nd edition. London: Pan Books.

Wilensky, R. (1983). Story Grammar versus Story Points. *The behavioral and brain sciences*, 6, pp. 579–623.

Cognitive Emotion Theories, Emotional Agents, and Narratology

Gesine Lenore Schiewer

University of Berne
German Studies – Unitobler – Länggassstr. 49 – CH 3000 Berne 9
gesine.schiewer@germ.unibe.ch

Abstract

Since Rosalind W. Picard in 1997 published her studies on „Affective Computing", emotion theories have become an important subject to artificial intelligence, information technology, and robotics. In this context cognitive emotion theories are not only appreciated but there are also some of them which have recently been newly developed from it. In my paper I shall balance reasons for adopting cognitive emotion theories in literary analysis. I refer on the fact that only cognitive emotions result in *individual*, *variable* and *relative* emotion-steered actions, whereas non-cognitive emotions result in unshirkable consequences. That is because cognitive emotions can be defined as radically individual opening a deliberate choice. Therefore the dimension of cognitive emotions is of special interest for the study of literature. After all literary texts show a preference for implications caused by deliberate actions even though the consequences are often not foreseen.

1. Emotional agents and emotion theories

In information technology and robotics, 'agent' is a current notion indicating so-called intelligent programs. Normally these are programs with a certain autonomy in the accomplishment of jobs; programs carry out an order in a similar manner to a human agent. Hence artificial intelligence is concerned with 'agents'.

Some agents feature a so-called BDI-architecture (believe, desire, intention). In this case human actions including the regarding decision-making processes is of specific significance as a template. An agent with a BDI-architecture is intended to imitate human decision-making. This means that decision-making is not the computation of an ideal and optimal plan once and for ever. In fact, decision-making is a permanent and optimal adaptation to changing circumstances. Thus the agent checks its aims under dynamic circumstances frequently whereas under static circumstances this is not necessary.

The significance of emotions in decision-making processes has been recognized increasingly in humanities and in cognitive sciences; therefore artificial intelligence adopts this strategy too. Therefore, 'emotional agents' are characterized by emotional behavior. This means that emotions are implemented in agent-architectures as a component of decision-making processes. Emotions are supposed to improve the man-machine-interaction in the first place and to steer behavior in the second place.

Further on important efforts are attempted in the development of 'affective computing' at present. In 1997 Rosalind W. Picard has published a fundamental study on this subject. Affective computing means computers "understanding" emotions (analysis) as well as computers "expressing" emotions (synthesis) and even "having" emotions (Ruebenstrunk, 1998). In this context emotion theories and computer models of emotions play a decisive role.

First, emotions comprehend many different aspects including physiological aspects, individual and subjective feelings, cognitive facets, the expression of emotions, social and cultural aspects and so on. German emotion psychologist Dieter Ulrich accentuated already in the eighties: "The decision for or against a specific model of emotions depends on what we are aiming at. Nobody is able to study emotions generally. In any case the regarding interest of exploration has to be specified." (Ulich, 1989, p. 125). Therefore no authoritative definition and no exclusive notion of emotions exists but a great number of models.

With regard to affective computing the so-called 'cognitive emotion theories' and 'appraisal theories' are discussed predominantly. The elicitation of emotions is understood as a consequence of specific cognitions. Cognitive appraisal is considered to be central to emotions. Hence the analysis of cognitive triggers of emotions and their consequences in expression, planning and acting as well represent an important aspect of these theories.

There are many different approaches of this kind; attention should be paid for example to those of Andrew Ortony, Gerald L. Clore and Allan Collins, Ira J. Roseman, Klaus R. Scherer, Nico H. Frijda, Keith Oatley and P.N. Johnson-Laird and some other. Lets take a look at them.

'Cognitive' theories of emotion have started (according to contemprary interpretations) with Aristotle's remarks on emotions in his 'Rhetoric', accompanied by some remarks in his *De anima* and *Nikomachian Ethics*. Some other philosophical classics, whose analytic treatment could serve as a source for modern cognitive theories of emotion, may count as predecessors of ‚cognitive' theories of emotion as well: René Descartes *Les passions de l'ame,* Baruch de Spinoza *Die Ethik und Kurze Abhandlung von Gott, dem Menschen und seinem Glück,* David Hume *A Treatise on Human Nature,* Immanuel Kant *Anthropologie in pragmatischer Hinsicht,* Adam Smith *Theory of moral sentiments.* Among early classics in psychology are Wilhelm Wundt *Grundzüge der Physiologischen Psychologie* and *Völkerpsychologie,* William James *The Emotions* and Carl Stumpf *Gefühl und Gefühlsempfindung*.

The ‚cognitive turn' in linguistics, anthropology and psychology which dates back to the sixties of the last century, found its application in emotion theory in the analysis of the role of appraisals of everyday situations in the elicitation of emotions. According to the latest variants of appraisal theory, persons use a fixed number of dimensions or criteria in the evaluation of situations:

1. Intrinsic characteristics of objects or events, such as novelty or agreeableness

2. The significance of the event for the individual's needs or goals.

3. The individual's ability to influence or cope with the consequences of the event, including the evaluation of 'agency'.

4. The compatibility of the event with social or personal standards, norms, or values.

The concept 'appraisal' has been used first in Magda Arnold's (1960) *Emotion and Personality* and has been deepened and detailed in the work of Richard Lazarus and his coworkers (Cf. Lazarus, Averill & Opton (1970) *Toward a Cognitive Theory of Emotion*, Lazarus (1977) *Ansätze zu einer kognitiven Gefühlstheorie*, Lazarus (1984) *On the Primacy of Cognition*, Smith & Lazarus (1993) *Appraisal Components, Core Relation Themes, and the Emotions*, Lazarus & Lazarus (1994) *Passion and Reason. Making Sense of our Emotions*, in a historical perspective Lazarus (1999) *The Cognition-Emotion Debate: a Bit of History*). The state of the art in cognitive emotion theory is recapitulated in Reisenzein (2000), Reisenzein, Müller & Schützwohl (2003), and in Scherer (1999) in historical perspective. The total state of devolpments concerning all aspects and dimensions connected with cognitive theories of emotion, together with methodological questions and empirical research, is given in Scherer, Schorr & Johnstone (Eds.) (2001). Another now very prominent ‚structural' access to emotions is Ortony, Clore & Collins (1988) *The Cognitive Structure of Emotions* (abbreviated: OCC-theory) with their '*Emotions-as-valenced-reactions Claim'* based on the situational concern of individuals, thereby referring to reactions to events, to actors and to objects respectively. Gordon (1987) discusses in *The Structure of Emotions. Investigations in Cognitive Philosophy* the difference between *Factive* und *Epistemic emotions* with reference to further *pivotal distinctions* which are usable to generate types of emotion.

The specific approach of Ortony, Clore and Collins is currently considered to be one of the most elaborated and systematic ones (cf. Reisenzein, Meyer & Schützwohl, 2003, p. 171). It is characterized by the general intention "to lay the foundation for a computationally tractable model of emotion. In other words, we would like an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotions." (Ortony, Core & Collins 1988, p. 2).

The authors attempt to give the outlines of an account of how cognitive appraisals as cognitive antecedents of emotions are made. The working characterization views emotions as valenced reactions to events, agents, or objects, their particular nature being determined by the way in which the eliciting situation is construed (cf. Ortony, Clore & Collins 1988, p. 13). The most general issue concerns the question of emotional differentiation, that is, the question of what distinguishes one emotion from another.

A basic distinction between reactions to events, agents, and objects gives rise to three basic classes of emotions: being pleased vs. displeased (reaction to events), approving vs. disapproving (reactions to agents), and liking vs. disliking (reactions to objects). The authors make clear that these three basic emotion classes can in turn be differentiated into a number of distinct groups of emotion types. Reactions to events breaks into three groups: one, the Fortunes-of-others group, focuses on the consequences for oneself of events that affect other people, whereas the other two, the Prospect-based and Well-being groups, focus only on the consequences for oneself. Reactions to agents are differentiated into four emotions comprising the Attribution group, and reactions to objects lead to an undifferentiated group called the Attraction group. Finally there is assumed to be also a compound group of emotions, the Well-being/Attribution compounds, involving reactions to both the event and the agent simultaneously (cf. Ortony, Clore & Collins, 1988, p. 33).

A further step takes into account that one of the most salient aspects of the experience of emotions is that their intensity varies both within and between people. Therefore the theory of emotion of Ortony, Clore and Collins addresses the question of what determines intensity. Their general view is that the intensity of emotions is influenced by a number of variables, all of which are present in the construal of the situation that gives rise to the emotion in the first place. Thus, in order to address the question of intensity, they consider the mechanism whereby emotion-inducing stimuli are appraised.

Ortony, Clore and Collins state that a person's appraisal of an emotion-inducing situation is based on three central variables: desirability, praiseworthiness, and appealingness, which apply to Event-based emotions, Agent-based emotions, and Object-based emotions, respectively. Desirability is evaluated in terms of a complex goal structure, where there is a focal goal that governs the interpretation of any event. The desirability of the event is appraised in terms of how it facilitates or interferes with this focal goal and the subgoals that support it. Similarly, the praiseworthiness of an agent's actions is evaluated with respect to a hierarchy of standards, and the appealingness of an Object is evaluated with respect to a person`s attitudes.

Following this approach goals are distinguished from standards in terms of what one wants vs. what one thinks ought to be. Three kinds of goals are distinguished: 1) Active-pursuit goals are goals that a person tries to obtain, such as becoming a concert pianist, 2) Interest goals are goals that are usually not pursued, because one has little control over their realization, as with preserving one's health or seeing one's friends succeed, and 3) Replenishment goals are goals that wax and wane, such as hunger and getting gas for one's car. The question of whether a goal is partially fulfillable, like making a million dollars, of fulfillable only in all-or-none terms, like winning a Nobel Prize, is considered to be orthogonal to these goal types. Ortony, Clore and Collins are convinced that these distinctions all play a role in determining the intensity with which people experience different emotions (cf. Ortony, Clore and Collins, 1988, p. 58). Among the variables that affect the intensity of different emotions are global variables, which affect all emotions, and local variables, which affect particular groups of emotions, too (cf. Ortony, Clore and Collins, 1988, p. 83).

Beyond this rather static and structural explanation of emotions the authors don't exclude the fact that there are emotion sequences, which is an important step forward to the dynamical flow of emotional processes.

They say that to the extent that one explains an action on the basis of internal traits or dispositions of the actor, these may mediate still other affective reactions. Because of a pervasive impetus to make causal attributions for significant events, experiencing one of the Event-based emotions will often be the occasion for experiencing one of the Attribution emotions. It can be inferred that the fact that people tend to seek causes for the significant events and actions that they experience means that there is a tendency for a movement from Event-focused to Agent-focused to Object-focused emotions (cf. Ortony, Clore & Collins, 1988, p. 169 seq.). In other words there may be a cycle in which emotion-inducing situations lead not only to emotions themselves, but also to a need to cope with the emotions to which they give rise. The extent to which a person does cope, or thinks he can cope, in some cases creates new, additional emotions, along with new demands on the coping mechanisms (cf. Ortony, Clore & Collins, 1988, p. 181).

In general, they accentuate that in many cases, the function of emotion is to lead the organism to cope with the emotion-inducing situation and the emotion itself.

They even make an attempt to explain why and under what conditions human beings are not able to cope with the emotion-inducing situation or the emotion itself. The unexpectedness of an event is considered to be a criterion of utmost relevance: "The result may be that there is a great deal of cognitive disorganization. This is true both for positive and negative emotions." (Ortony, Clore & Collins, 1988, p. 178 seq.).

This admittedly convincing explanation does not cover any specific circumstances of the mentioned disorganization, however. It was cognitive psychologist Jerome Bruner who accentuated the concerned question already in 1990 in his book *Acts of Meaning*: "This reciprocal relation between perceived states of the world and one's desires, each affecting the other, creates a subtle dramatism about human action which also informs the narrative structure of folk psychology. When anybody is seen to believe or desire or act in a way that fails to take the state of the world into account, to commit a truly gratuitous act, he is judged to be folk-psychologically insane unless he as an agent can be narratively reconstrued as being in the grip of a mitigating quandary or of crushing circumstances. It may take a searching judicial trial in real life or a whole novel in fiction (as with André Gide's *Lafcadio's Adventure*) to effect such a reconstrual." (Bruner, 1990, p. 40). Bruner makes clear that folk psychology is invested in canonicality and focuses upon the expectable and/or the usual in the human condition. Naturally, this includes the expectable and/or usual in the affective life. But nonetheless a culture must contain a set of norms, following Bruner it must also contain a set of interpretative procedures for rendering departures from those norms meaningful in terms of established pattern of belief. Bruner says that it is narrative and narrative interpretation upon which folk psychology depends for achieving this kind of meaning. "Stories achieve their meanings by explicating deviations from the ordinary in a comprehensible form – by providing the 'impossible logic [...]." (Bruner, 1990, p. 46). In Bruners view the function of a story is to find an intentional state that mitigates or at least makes comprehensible a deviation from a canonical cultural pattern (cf. Bruner, 1990, p. 45 seq.).

Thus, Bruner suggests to establish a connection between cognitive science, emotion theories, and narratology. Therefore, it is essential to take a look at current discussion of emotions in literary studies.

2. On reception of current emotion theories in literary studies

Rhetorics as well as literature as long as it was based on rhetorics, has always been concerned with emotions. Surprisingly enough, the modern studies of literature seemed to ignore a systematic discussion of the whole complex of affective aspects in literature and of literature, though there are a lot of substantial and important studies of the history of the complex of affects and literature. But quite recently an incresing interest in this topic is remarkable (i.e. Alfes 1995, Winko 2003). With respect to literature Simone Winko refers on an essential differentiation with respect to emotion theories: it concerns a so-called wider conception of cognition integrating thinking and feeling on the one hand and on the other hand a narrow conception of cognition exclusively comprising cognition and therefore excluding emotion as an independent aspect of the human mind. Winko argues for adopting the narrow conception of cognition in literary studies: this means that according to Winko a broad conception of emotion is favourable, which takes into account that emotions can be regarded as a mental phenomenon als well as a physiologic, psychologic, social or cultural occurrence.

This sounds good and seems to tend to a so-called syndrome theory. But looking closely, there appear some problems. Winkos belief is that cognitive conceptions of emotions are not apt to literary studies because in her mind they don't allow to grasp the perceptional aspects of emotions. In consequence of this she suggests to adopt a notion of emotion taking emotions as emergent characteristics of the physical system of human beings.

In spite of Winkos definite position the topic has to be discussed furthermore. Therefore, I shall balance reasons for adopting cognitive emotion theories – albeit not exclusively – in the studies of literature although and respectively even because they are based on the wider conception of cognition integrating thinking and feeling.

3. Perspectives of cognitive emotion theories for literary analysis

As aforementioned, Ortony, Clore and Collins take into account the differentiation between positive and negative consequences of emotions, that is, that emotions may result in a great deal of cognitive disorganization. Furthermore, they accentuate the fact that emotions can cause dramatic disruptions in judgment and performance especially is recognized by creators of literature, which thrives on the imagined emotions of its characters. Ortony, Clore and Collins think that the basic recipe is very simple. "The writer describes a situation that readers recognize as being *important* to a character in the sense that it has important implications with respect to the goals, standards, or attitudes that the character is known or assumed to have. Then, the character is portrayed as correctly or incorrectly construing the situation as good or bad relative to these goals or standards or attitudes, and typically is described as having, or is assumed to have, a valenced (i.e., a positive or negative) *reaction* to the

situation. Finally, the construal together with the reaction usually results in some sort of change in the character's judgment or *behavior*." (Ortony, Clore & Collins 1988, p. 3).

Further on the authors make clear that in their mind the description of the actual situation as undesirable – from the point of view of the experiencing individual encoding the relevant situation in a particular way – is often sufficient to produce in readers an awareness of a character's affective states. They assume that the described situations are sufficient to produce individual emotions. It has even not to be stated what emotions a character is experiencing because if the described situation contains the eliciting conditions for a particular emotion, the experience of that emotion can be inferred. Following Ortony, Clore and Collins this assumption is proved by the fact that millions of readers, often over decades or even centuries, all infer similar emotions from the described situations. Therefore they assume that this view cannot be too wrong (cf. Ortony, Clore and Collins 1988, p. 3).

This rather general, but nonetheless convincing prospect of the analysis of literary descriptions of emotions and their eventual consequences has to be elaborated.

First, is becomes clear that against Simone Winkos position cognitive emotion theories are – albeit not exclusively – extremely interesting in literary analysis considering the fact that only cognitive emotions result in *individual*, *variable* and *relative* emotion-steered actions, whereas non-cognitive emotions result in unshirkable consequences. Wolfgang Gessner has discussed this aspect in his recently published book on his "radical cognitive emotion theory" titled *Die kognitive Emergenz von Emotionen*. Following Gessner that is because cognitive emotions can be defined as radically individual opening a deliberate choice. In contrast, non-cognitive emotions like for example disgust are not individual but constitutive for human beings. Therefore the dimension of cognitive emotions is of special interest for the study of literature. After all literary texts show a preference for implications caused by deliberate actions even though the consequences are often not foreseen.

Second, Gessner elaborates the analytical instruments regarding the inner perspective of human beings and their interpretation of emotion-inducing situations. He focuses in his theory on the analysis of the individual interpretation mechanisms regarding an emotion-inducing situation, whereas usually appraisal theories usually focus on standard situations of emotion elicitation. Gessner says that a complete emotion theory has to explain the specific kind of individual dispositions, that is the individual cognitive triggering, causing a specific interpretation of an emotion-inducing situation (cf. Gessner 2004, p. 127).

Without being able to present Gessners approach in detail in this paper, it becomes clear immediately that this theory focuses especially on individual and subjective factors vis-à-vis a given situation including wrong and erroneous interpretations of it. It is this aspect that is promising from the point of view of literary analysis.

Perhaps it is useful to remind a literary instance. I choose a best-known text, namely Goethe's *Faust*. This text is one of the most striking examples of a complex discussion of the problem of emotions. It shows not only the failure of the individual decisions but also the dimensions of emotions in the much broader context of social and human aspects. Already in the Prologue in Heaven Lord and Mephistopheles talk about human error:

Mephistopheles:
"What will you wager that you do not lose him,
Supposing always you will not demur
About my guiding him in paths I choose him?"
The Lord:
"You shall have leave to do as you prefer.
So long as earth remains his mortal dwelling;
For man must strive, and striving he must err."

Indeed, Faust is erring. Let us take for example the *Gretchen*-episode. Even Mephistopheles tries to keep him back in the beginning:

Faust:
"But, none the less, she must be turned fourteen."
Mephistopheles:
"There speaks the lad who plays the libertine,
And thinks he has the right to every flower,
Knowing no grace or honourable name
Beyond his reach, to pluck it and devour;
It often can't be done, Sir, all the same.
[…]
Pray hear me now, Sir, pleasantry apart,
I tell you once for all, that lovely girl
Is never to be taken in a whirl.
We stand to lose by forcing of the pace,
When gentle subterfuge would meet our case."

Corresponding the cited dialogue in the *Prologue in Heaven*, Goethe writes in a letter from 15th September 1804 to his confidant Eichstätt that was is justly called a wrong striving is an unavoidable detour to reach the end. This is because every return from an error forms the human being as an individual and as a whole as well. Therefore it seems clear to him that God prefers a penitent sinner to ninety-nine one believers.

Faust's behavior driven by passion and strong emotions has to be judged in the perspective of Goethe's letter to Eichstätt. The whole problem of emotions and their theoretical description becomes even more complex than the mentioned theories of Ortony, Clore and Collins and of Gessner as well take into account. Nonetheless, the are able to describe the individual facets of the elicitation of emotions in a convincing manner.

4. Summary and conclusions

No serious attempts have been made to integrate the many-faceted attempts of explanations and theories governing the manifold aspects of emotions, which altogether contribute to human behaviour as a whole.

Emotions are not solely based in anatomy, biology, psychology or culture. Thus, the study of emotions cannot be confined to only one discipline like psychology, philosophy, ethnology, sociology or linguistics. Emotions in human interaction seen as appraisal with respect to the given situation or the communicated information content cannot be addressed by reducing the complex and isolating individual aspects which are analysed from a specific, for instance, psychological perspective (cf. WEIGAND 2004). Only in interdisciplinary approaches researchers may hope to arrive at really innovative results which can cover the area of emotions including the requirements of full-fledged and efficient communication tools and action devices. Therefore a broad scientific

horizon of psychology, linguistics, information science, image and signal processing as well as philosophy has to contribute to an integrative approach avoiding the reproach of psychologism (cf. Metzger [6]2001 and Scherer 2000).

Further on there are not only systematic approaches to be taken into account but historical ones, too. For example, Karl Bühler presents in the subtitle of his book on *Ausdruckstheorie* from 1933 an outstanding methodological principle: a system must be presented by writing the history of the thoughts concerned. It is even Bühler's assumption that a systematic concept has to be highlighed and completed by its historical development. He gives several reasons for this scientific principle, one of them is very simple but nonetheless fundamental: familiar knowledge is in danger to be forgotten by a systematic approach. In fact, structures, incidents and continuities have to be correlated in a constructive manner.

Bühler's assumptions takes an interesting rebound in his latest writings published in *Das Gestaltprinzip im Leben des Menschen und der Tiere* (1960). Here, Bühler is dealing with cybernetics. He makes a comparision between the steering of computers or machines on the one hand and human thinking on the other hand. At that time cybernetic storages have to be depleeted completely before the memory can be used once more. Entirely different is the situation when human beings think in a creative manner: human thinking is inventiv or „gestaltistisch" exclusively when it is based on already acquired knowledge, even if this knowledge not systematized. Therefore, human thinking has to be regarded as systematic and historical at the same time, not just as systematic.

5. References

Alfes, Henrike F. (1995): *Literatur und Gefühl. Emotionale Aspekte literarischen Schreibens und Lesens*. Opladen: Westdeutscher Verlag.

Arnold, M. B. (1960). *Emotion and Personality*. Bd I. Psychological Aspects, Bd. II. Neurological and Physiological Aspects. New York: Columbia University Press.

Bruner, Jerome (1983): *In Search of Mind*. New York: Harper & Row Publishers.

Bruner, Jerome (1990): *Acts of Meaning*. Cambridge: Harvard University Press.

Gessner, W. (2004). *Die kognitive Emergenz von Emotionen*. Paderborn: Mentis.

Lazarus, R. S. e. a. (1977). *Ansätze zu einer kognitiven Gefühlstheorie*, BIRBAUMER 1977, 182-207.

Lazarus, R. S. (1984). On the Primacy of Cognition. *American Psychologist*, 39, 1984, 124-129.

Lazarus, R. S., & Lazarus, B. N. (1994). *Passion and Reason. Making Sense of Our Emotions*. New York: Oxford University Press.

Lazarus, R. S. (1999). *The Cognition-Emotion Debate: a Bit of History*, DALGLEISH & POWER 1999, 3-20.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge U.P.

Picard, R. W. (1997). *Affective computing.* Cambridge (Mass.): MIT Press.

Picard, R. W. (2001). *Affective medicine: Technology with emotional intelligence*, MIT Media Lab Report 537.

Picard, R. W. (2002). *What Does It Mean for a Computer to "Have" Emotions*? In P. P. TRAPPL, 213-235 (Ed.). Cambridge, Mass.: The MIT Press.

Reisenzein, R. (2000). Exploring the Strength of Association between the Competents of Emotion Syndromes. The Case of Surprise. Cognition and Emotion 2000, 14 (1), 1-38.

Reisenzein, R., Meyer, W.-U., & Schützwohl, A. (2003). *Einführung in die Emotionspsychologie* Bd. III: Kognitive Emotionstheorien. Bern: Hans Huber.

Ruebenstrunk, G. (1998). *Emotionale Computer. Computermodelle von Emotionen und ihre Bedeutung für die emotionspsychologische Forschung*.Unpublished manuscript, http://www.ruebenstrunk.de/.

Ruebenstrunk, G. (2004). *Emotional Machines*. Paper presented at the V2_Lab workshop, Rotterdam.

Scherer, K.R. (1990). *Psychologie der Emotion*. Göttingen/Toronto/Zürich: Hogrefe.

Scherer, K. R. (1999). *Appraisal Theory*, DALGLEISH & POWER 1999, 637-664.

Scherer, K.R., Schorr, A. & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotions. Theory, Methods, Research*. Oxford: Oxford University Press.

Scherer, K.R. & Wallbott, H.G. (1990). Ausdruck von Emotionen, in: K.R. Scherer, 345-422.

Weigand, E. (Ed.) (2004). *Emotion in dialogic interaction. Advances in the complex*. Amsterdam: Benjamins.

Winko, Simone (2003): *Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Berlin: Erich Schmidt.

Abstract

This paper presents a method on locating proverbs in literary texts. The basis for our work was the electronic dictionary of Modern Greek proverbs compiled from 2500 Modern Greek canonical proverbial forms. In contrast with other multi-word expressions proverbs do not have free elements. They have a fixed form but they can also exhibit variation. In order to create an exhaustive linguistic resource we had to represent all proverbial variants. The dictionary we built up relies on the finite state technology. We applied the FSTs library in a literary text, we isolated the proverbs and we collected the concordances. We discuss the advantages of FSTs as it concerns the recognition of frozen sentences like proverbs. Canonical proverbial forms can be recognized in a very high percentage. Finite-state transducers also recognize proverbs with discontinuous constituents because of insertions or half-truncated proverbs.

LOCATING PROVERBS WITH FINITE-STATE TRANSDUCERS IN LITERARY TEXTS

Tsaknaki Olympia

Laboratoire d'informatique, IGM, Université de Marne-la-Vallée

77454 Marne-la-Vallée Cedex 2 France

tsaknaki@univ-mlv.fr

INTRODUCTION

In the present paper we concentrate on issues in automatic recognition of proverbs in texts. Proverbs have been considered for many years as a marginal linguistic phenomenon which did not merit the interest of linguists. Recent publications of researchers [1] prove that the initiation of a linguistic study on the paremiological heritage is necessary and very useful. The studies which deal with proverbs underline their presence and use in our everyday life. Apart from the oral use, proverbs can be detected in advertisements, political speeches and songs and especially in journalistic texts[2]. It is important to notice that proverbs are also merged in literary texts[3] and paraliterature.

Proverbs are frozen sentences. They are non-compositional, they do not accept insertions among their constituents, they cannot be actualized, the order of their constituents cannot be changed and they have lexical and syntactic restrictions. Additionally, they can be defrozen (G. Gross 1996).

In order to acquire proverbs in texts we built up an electronic dictionary compiled from 2500 Modern Greek proverbs. The proverbs were attested in general language dictionaries as well as in proverb collections. We also used as sources native speakers and Press. Journalistic texts express the dynamics of the Modern Greek language which evolves continuously and adopts new forms we must take into consideration.

The electronic dictionary of proverbs was created to analyse any kind of text corpora (Kyriacopoulou & Tsaknaki 2002). In this paper we apply it to a literary text.

Our approach is based on a thorough linguistic analysis. The framework of the lexicon-grammar theory introduced by M. Gross (1975) on the basis of a transformational theory (Harris 1968) was adopted. Within the above-mentioned framework, linguistic resources are organized in three forms: electronic dictionaries, finite-state automata and lexicon-grammar matrices.

Proverbs do not always appear in texts in their canonical proverbial form. Punctuation marks or lexical elements can be inserted among the proverbial constituents. They can also appear half-truncated or defrozen. Defrozen proverbs are not subject to be studied at the framework of this paper.

As a starting point, we describe the different types of variants in proverbs. We then present finite-state transducers representing proverbs. We also point out how important is the correct sentence detection in the case of proverbs consisted of two or more sentences. Moreover, we deal with discontinuous constituents and half-truncated proverbs once they are inserted in discourse. We finally apply finite-state transducers we created to a literary corpus.

[1] The Journal *Langages* devoted an entire issue to proverbs. This issue also contains a bibliography of important references (See *Langages* 139 (2000) «La parole proverbiale»). Linguistic studies on proverbs are among others the following: Lakoff & Turner (1989), Hasan-Rokem (1992), Conenna (1994), Kleiber (1994), Anscombre J. C. (1994, 1997), Anastassiadis-Symeonidis (1998), Conenna (1998a, 1998b), Kleiber (1999a, 1999b), Michaux (1999), Schapira (1999), Conenna (2000), Conenna & Kleiber (2002), Anscombre J. C. (2003), Conenna (2004).

[2] Mieder (1983), Gavriilidou (2002), Risto Järv : http://haldjas.folklore.ee/folklore/vol10/toughjob.htm.

[3] Numerous publications concern proverbs in literature. E. g.: Anstensen (1966), Bryan & Mieder (1994), Mieder & Bryan (1996), Bryan & Mieder (1997).

VARIANTS

Variants in proverbs can be in different levels: graphic, orthographic, morphological, lexical and morphosyntactic. As an example we give some types of variants:

- Graphic level

*Όποιος σπέρνει τεμπελιά**,** θερίζει πείνα*
Who seeds laziness**,** reaps hunger
Όποιος σπέρνει τεμπελιά θερίζει πείνα[4]
Who seeds laziness reaps hunger

- Orthographic level

This category includes the pure orthographic variants and the phonological variants which cause changes in the orthographic level.

***Επάνω** στη βράση κολλάει το σίδερο*
***Απάνω** στη βράση κολλάει το σίδερο*
***Πάνω** στη βράση κολλάει το σίδερο*
On the boil the iron sticks[5]

- Morphological level

Certain parts of speech can be said or written in Modern Greek in two or more equivalent ways. As it concerns verbs, contrary to other languages, this phenomenon is very frequent[6] (Kyriacopoulou 1990). In the following example two verbs sharing the same radical have lexically different suffixes:

*Όποιος **ανακατώνεται** με τα πίτουρα, τον τρώνε οι κότες*
*Όποιος **ανακατεύεται** με τα πίτουρα, τον τρώνε οι κότες*
Who **blends** with bran, is eaten by hens

- Lexical level

Proverbs, as frozen sentences, present many constraints in the lexical level. Despite these constraints, variants in the lexical level are not impossible. The verbs *βαφτίζω* (baptize) and *βγάζω* (call) have a different meaning but they can be interchangeable in the following proverb:

*Ακόμη δεν τον είδαμε, Γιάννη τον **βαφτίσαμε***
We haven't seen him yet but we **baptized** him Giannis
*Ακόμη δεν τον είδαμε, Γιάννη τον **εβγάλαμε***
We haven't seen him yet but we **called** him Giannis

- Morphosyntactic level

Variants can also affect the morphosyntactic level. The preposition *από* (by) which indicates the reason can be substituted either by the preposition *σε* (in) which indicates the place or the preposition *με* (with) which indicates the means:

*Όποιος καεί **απ' το** χυλό, φυσάει και το γιαούρτι*
Who is burnt **by the** porridge, he even blows the yogurt
*Όποιος καεί **στο** χυλό, φυσάει και το γιαούρτι*
Who is burnt **in the** porridge, he even blows the yogurt
*Όποιος καεί **με το** χυλό, φυσάει και το γιαούρτι*
Who is burnt **with the** porridge, he even blows the yogurt

Certainly, there are cases where different kinds of variants co-exist.

All variants cannot be exhaustively listed in the general or specialized dictionaries for many reasons. The aim of these resources is not a detailed description nor their size does permit it. Furthermore, proverbs are orally transmitted and they can undergo changes even if the latter develop very slowly. The resources cannot always be updated.

From the brief description in this section the conclusion that can be reached is that locating proverbs yields good results to the extent that variants are exhaustively and systematically represented.

Given that proverbs are completely frozen sentences, one could make the hypothesis that their constituents are extremely fixed and consequently it would be sufficient for the representation of proverbs to create a list whose number of entries would be equal to the number of proverbs collected. However, according to the linguistic analysis we have undertaken and briefly presented, this solution seems to be inadequate. The variability proverbs can exhibit and the different types of variants they present do not allow the use of this kind of formal representation. In order to acquire proverbs we opted for the representation of all variants of a proverb in a finite-state transducer, i.e. finite-state automaton with input/output labels.

The above-named devices offer readability, precision and efficiency. They are attractive as they compress data and they allow the addition of new variants. Redundancies are avoided. We created a library of finite-state transducers of 2500 Modern Greek proverbs.

We will show how finite-state transducers can provide solution in the recognition of canonical proverbial forms as well as in the case of proverbs with half part omitted or discontinuous elements. Defrozen proverbs must be studied separately because they can be modified in different and unpredictable ways (Tsaknaki 2005).

The transducer presented in Figure 1 represents all variants of the proverb: *Ο Θεός αγαπάει τον κλέφτη, αγαπάει και το νοικοκύρη* (God loves the thief, he loves also the host[7]):

[4] Every proverb is followed by its literal translation in English. In case an equivalent exists we add it as a footnote.
[5] Strike while the iron is hot.
[6] Examples in English are: *burnt/burned, dreamt/dreamed* and in French *paye/paie, j'assois/j'assieds.*
[7] Literal translation.

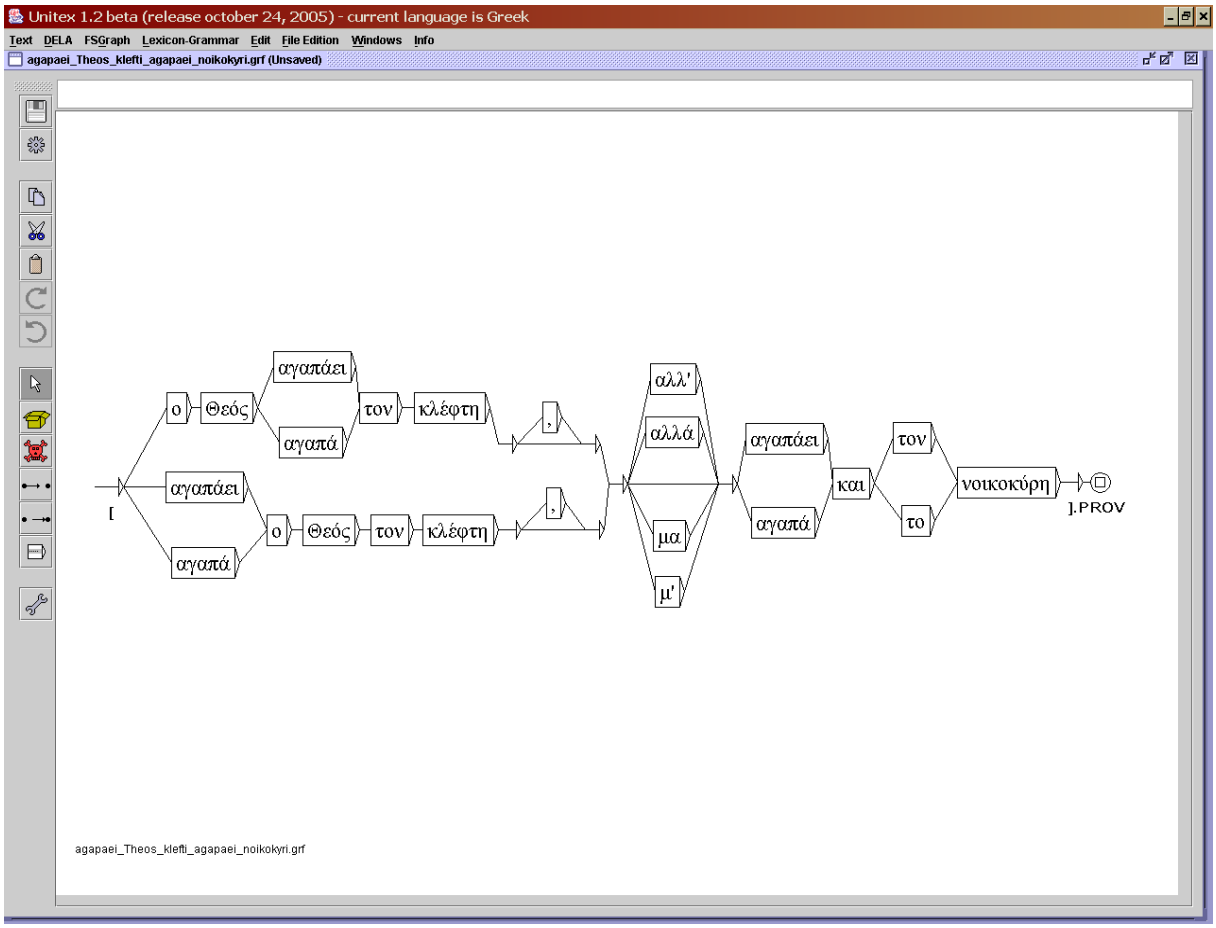

Figure 1. Finite-state transducer representing the proverb *Ο Θεός αγαπάει τον κλέφτη, αγαπάει και το νοικοκύρη* and its variants

SENTENCE DETECTION

Sentence boundaries in a corpus must be always detected. In the field of Natural Languages Processing, reliable sentence detection is the first important step. If a text is not well segmented into its constituent sentences, a thorough linguistic analysis is not possible. One of the major problems is that the punctuation marks do not necessarily classify periods.

The notion of sentence must be also very clear in the case of proverbs. In Modern Greek exist proverbs consisted of two or more sentences[8] followed either of a full stop (.) or a question mark (;). They are not numerous or very frequent. Nevertheless, they must be represented and recognized. One important issue is that the sentences must not be recognized separately but as one unit. For this reason we created the finite-state transducer PROVS1 (Figure 2) which represents all proverbs in Modern Greek composed of two or more sentences. The transducer permits the recognition of the proverb even if the sentences of the latter are conjoined into a single separated by a comma. The graph in Figure 3 segments a text into sentences. We added the sub-graph PROVS1.

INSERTIONS & HALF-TRUNCATED PROVERBS

Another difficulty which can arise frequently and we have to face is the use of insertions stuck in among the constituent elements. The presence of insertions such as *λένε* (they say), *λέει η παροιμία* (says the proverb), *έλεγαν οι παππούδες μας* (said our grandfathers), *είναι γνωστό ότι* (it is known that) should be predicted as well as any kind of word or phrase that can be inserted, e.g. *Η σιωπή στο ποδόσφαιρο είναι χρυσός* (Silence in football is gold[9]), *Περασμένα αλλά ποτέ ξεχασμένα* (Passed, but never forgotten[10]). Additionally, punctuation marks, e.g. the comma or the ellipsis, must be taken into account. In an opposite case situation, the system will not continue the process.

To arrive at our aim we created a finite-state transducer named INS (Figure 4) where we represent possible insertions that could impede locating all the constituent elements and discard the sequence. <MOT> stands for any word. INS can recognize any sequence of words inserted in the proverb even in the case that it is preceded or followed by punctuation marks such as parentheses, commas or ellipsis.

[8] The phenomenon was also observed in other languages (Anscombre 2000 : 13).

[9] TA NEA, 06-08-01, Canonical proverbial form = Silence is gold.

[10] TA NEA, 05-07-00, Canonical proverbial form = Passed, forgotten.

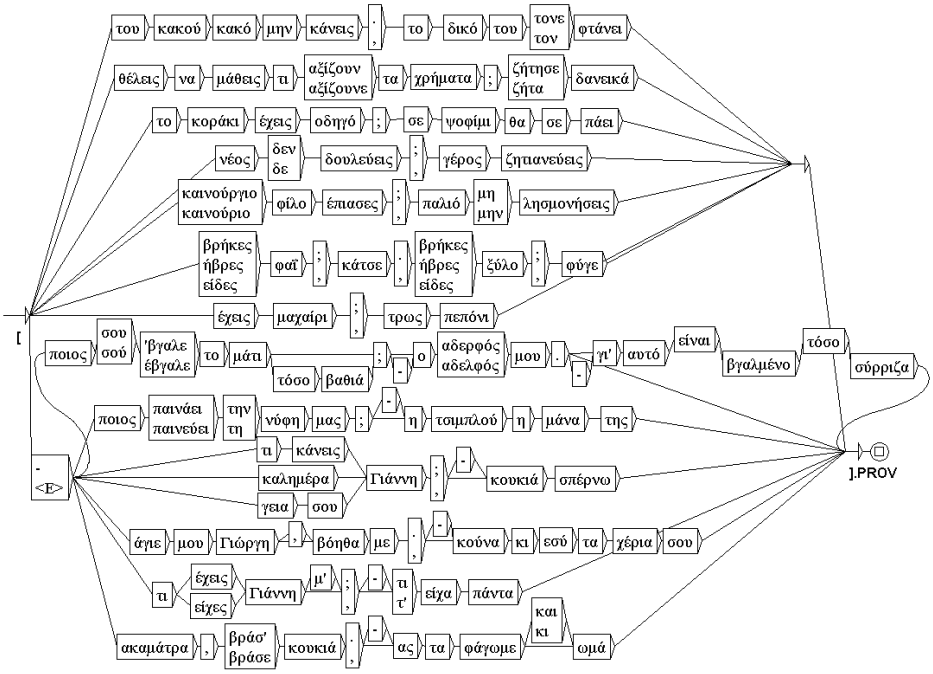

Figure 2. Transducer representing the proverbs consisted of two or more sentences

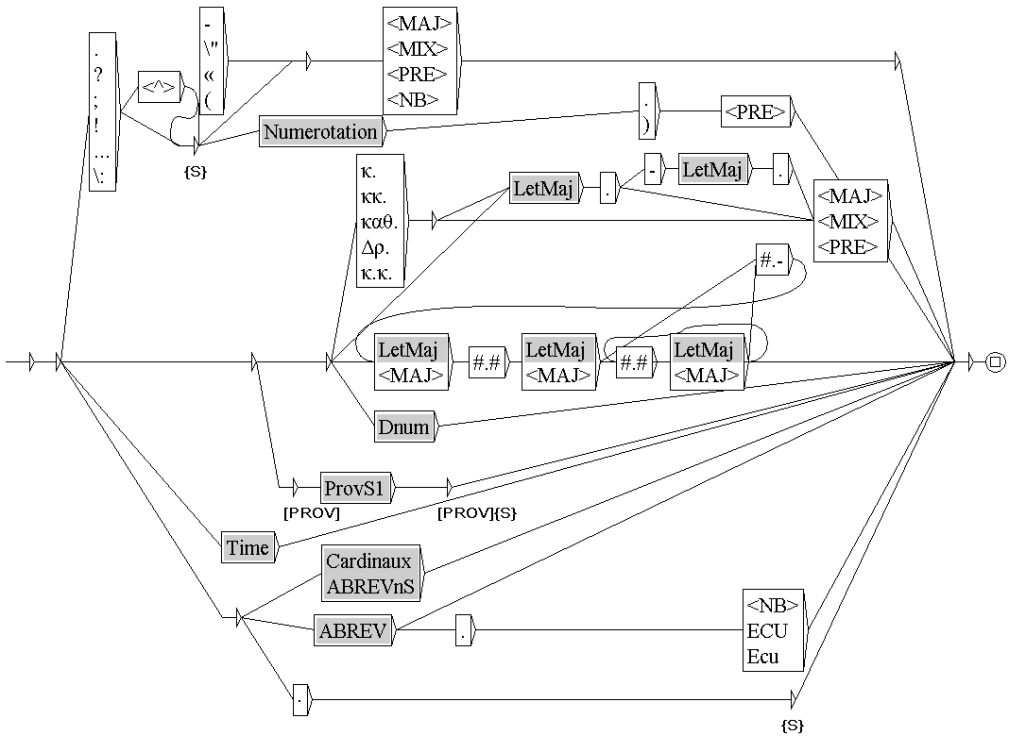

Figure 3. Graph for the sentence segmentation (Unitex, 23-04-2005)

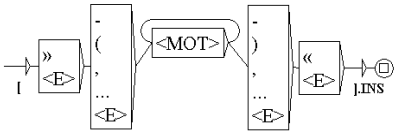

Figure 4. Finite-state transducer able to locate insertions among the proverbial constituents

Half-truncated proverbs, e.g. *Των φρονίμων τα παιδιά...* (The children of the prudent...[11]), *...πριν πεινάσουν μαγειρεύουν* (...cook before they get hungry) must be also recognized. The omission of the half part of the proverb can be easily represented by means of a transducer. The following finite-state transducer can locate the entire proverb *Των φρονίμων τα παιδιά πριν πεινάσουν μαγειρεύουν* (The children of the prudent cook before they get hungry) and also the half parts when they are separately presented:

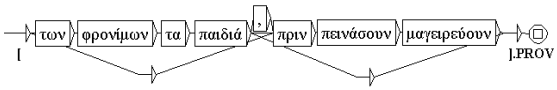

Figure 5. Finite-state transducer able to recognize half part of the represented proverb

APPLICATION TO TEXTS

In order to perform our method and test its efficiency we chose the well-known novel *Το τρίτο στεφάνι* (The third wedding)[12] written by Costas Taxtsis in 1962. This novel includes many proverbs and our method could be tested.

The first processing step was to apply the transducers we created to our corpus.

To treat our data we used Unitex, a corpus processing system, based on automata-oriented technology developed in the Institut Gaspard Monge, Université de Marne-la-Vallée, France (Paumier 2003). This system has been successfully used for the automatic treatment of many languages.

The system recognised the proverbs of the corpus and the output PROV was inserted at the left and right side of each located proverb. The transducer imposes the limits of the sequence, so a morphological and syntactic analysis cannot be performed.

Finally, concordances were collected. It is known that concordances are of great use for linguistic and stylistic studies. They are essential tools for a corpus treatment. In the case of literary studies they can show the different ways a word is used in classic and modern texts or the way an author or different authors use a word in their works. A series of concordances provided by the system is below presented:

αφεία με το δίσκο – [PROV] [από δήμαρχος κλητήρας].PROV{S} Ήλπιζαν ότι με τα κέρδη του καφεν

από κάθε άλλον – [PROV] [η καλή μέρα απ' το πρωί φαίνεται].[PROV] {S} Πήγα για να κάνω το χατίρι του Α

είπα με το νου μου, [PROV] [λαγός την πτέρην έσειε κακό της κεφαλής του].[PROV] {S} Δε θα διορθώσω εγώ

τα 'κανες όλ' αυτά, [PROV] [λαγός την πτέρην έσειε κακό της κεφαλής του].[PROV], αλλά δε μπορούσες του

λέει η παροιμία – [PROV] [μάθε τέχνη κι άστηνε, κι όταν πεινάσεις πιάστηνε].[PROV] {S} Δεύτερον σκέφτηκε ότι

του φαραώ. {S} Γιατί [PROV] [ο Θεός αγαπάει τον κλέφτη, αγαπάει και το νοικοκύρη].[PROV] {S} Σ' αφήν

μου' λεγε, "πως [PROV] [τα ρούχα δεν κάνουν τον άνθρωπο].[PROV] {S} Εδώ στην Ελλάδα ο κόσμος σε κρ

δεν είχε πάθει τίποτα, [PROV] [το κακό σκυλί ψόφο δεν έχει].[PROV] {S} Την ώρα του τορπιλισμού σουλατσάρι

The proverb:

-Ποιος σού 'βγαλε το μάτι τόσο βαθιά; (Who took you out the eye so deeply?)
- Ο αδερφός μου. (My brother.[13])

was also recognized in the text:

όπως το λέει κι η παροιμία: [PROV] [ποιος σού 'βγαλε το μάτι τόσο βαθιά; - ο αδερφός μου]. [PROV] {S}

Finally, the transducer INS helped us recognize the proverb: *Του παιδιού μου το παιδί είναι δυο φορές παιδί μου* (The child of my child is twice my child[14]) which is interrupted in the text by the verb *έλεγε* (said).

> Εγώ βεβαίως καταλαβαίνω το πνεύμα του Άκη. Αυτός τη θυμάται περισσότερο απ' όλους μας. Τον μεγάλωσε δα, του στάθηκε καλύτερ' από μάνα. Πώς τον αγαπούσε! «**Του παιδιού μου το παιδί**», έλεγε, «**είναι δυο φορές παιδί μου.**» (p. 306)
> (I certainly understand Akis' way of thinking. He remembers her more than everyone of us. She brought him up, she stood by him more than a mother. How much she loved him! «**The child of my child**», έλεγε, «**is twice my child.**»)

CONCLUSION

In this study our work concentrates on the automatic recognition of Modern Greek proverbs. We have shown that finite-state technology can produce satisfying results. The method we used can also be expanded to other languages. There has been already realized an important study as regards French and Italian proverbs (Conenna 1998a, 1998b, 2000). We can proceed to a comparative

[11] Literal translation.
[12] The book has been translated several times into English under the titles *The Third Wedding* (London : Alan Ross, 1967, New York : Red Dust, 1971) or *The Third Wedding Wreath* (Athens : Hermes, 1985).
[13] Literal translation.
[14] Literal translation.

study between libraries of finite-state transducers representing proverbs in order to study crosslingual similarities and differences.

Besides the fact that essential elements are covered, open problems such as attested proverbial forms subjected to changes have not been discussed but it must be noticed that they should be the object of an extended research.

To conclude we comment on the great utility of finite-state transducers in handling proverbs. They:

- allow a rigorous description and a systematic analysis of our data. Their application to a literary corpus can allow the system to localize all proverbial forms

- can find a proverb in a text as much times as it is repeated.

REFERENCES

Anastassiadis-Symeonidis A. (1998). Le proverbe en grec moderne. In S. Mejri, A. Clas, G. Gross, T. Baccouche (Eds), *Le figement lexical: Préactes des 1ères Rencontres Linguistiques Méditerranéennes (17-19/9/1998)*. Tunis : CERES, pp. 77-85.

Anscombre J. C. (1994). Proverbes et formes proverbiales : valeur évidentielle et argumentative. *Langue française* 102, pp. 95-107.

Anscombre J. C. (1997). Reflexiones críticas sobre la naturaleza y el funcionamiento de las paremias. *Paremia* 6, pp. 43-54.

Anscombre J. C. (2000). Parole proverbiale et structures métriques. *Langages,* 139, pp. 6-26.

Anscombre J. C. (2003). Les proverbes sont-ils des expressions figées?. *Cahiers de Lexicologie,* 82 (1), pp. 159-173.

Anstensen A. (1966). *The proverb in Ibsen,* New York : AMS Press, Inc.

Bryan G. B. & W. Mieder (1994). *The proverbial Bernard Shaw : an index to proverbs in the works of George Bernard Shaw*, Westport, CT: Greenwood Press.

Bryan G. B. & W. Mieder (1997). *The proverbial Charles Dickens : an index to the proverbs in the works of Charles Dickens.* New York : Peter Lang.

Conenna M. (1994). Considerazioni traduttologiche sul lessico-grammatica. *Lingua Franca,* 1, pp. 19-35.

Conenna M. (1998a). Sur un lexique comparé de proverbes. *Langages,* 23 (90), pp. 99-116.

Conenna M. (1998b). Le proverbe, degré ultime de figement?. In Mejri S., G. Gross, A. Clas, T. Baccouche (eds), *Le figement lexical.* Tunis : Actes des 1ères Rencontres Linguistiques Méditerranéennes, pp. 361-371.

Conenna M. (2000). Classement et traitement automatique des proverbes français et italiens. *BULAG,* Numéro spécial, Mélanges Gaston Gross, pp. 285-294.

Conenna M. (2004). Principes d'analyse automatique des proverbes. *Lexique, Syntaxe et Lexique-grammaire: Papers in honour of Maurice Gross, Linguisticae Investigationes: Supplementa* 24, pp. 91-104.

Conenna M. & G. Kleiber (2002). De la métaphore dans les proverbes. *Langue française,* 134, pp. 58-77.

Gavriilidou Z. (2002). Proverb in Greek press. In *Proceedings of the 5th International Congress in Greek Linguistics* (Sorbonne 3-15/9/2001), Paris : L'Harmattan, pp. 207-210.

Gross G. (1996) *Les expressions figées en français : Noms composés et autres locutions,* Paris : Ophrys.

Gross M. (1976). *Méthodes en syntaxe,* Paris : Harmattan.

Harris Z.S. (1968). *Mathematical Structures of Language,* New York : J. Wiley & Sons.

Hasan-Rokem G. (1992). The Pragmatics of Proverbs: How the Proverb gets its meaning. In: L. K. Obler, L. Menn (eds), *Exceptional Language and Linguistics,* New York : Academic Press.

Kleiber G. (1994). Sur la définition du proverbe. *Nominales,* Paris : A. Colin, pp. 207-224.

Kleiber G. (1999a). Les proverbes antinomiques: Une grosse pierre "logique" dans le jardin toujours "universel" des proverbes. *Bulletin de la Société de Linguistique*, XCIV/1, pp. 185-208.

Kleiber G. (1999b). Proverbe: sens et dénomination. Le proverbe, un pro…nom?. *Nouveaux Cahiers d'Allemand,* 3, pp. 515-531.

Kyriacopoulou P. (1990). Les dictionnaires électroniques. La flexion verbale en grec moderne. PhD Thesis, Université Paris VIII, Paris.

Kyriacopoulou T. & O. Tsaknaki. (2002) Representation of proverbs by finite-state automata, *Studies in Greek Linguistics,* 23, pp. 860-871.

Lakoff G. & M. Turner (1989). *More than Cool Reason. A Field Guide to Poetic Metaphor.* Chicago : Chicago University Press.

Michaux C. (1999). Proverbes et structures stéréotypées, *Langue française,* 123, pp. 85-104.

Mieder W. (1983). Verwendungsmöglichkeiten und Funktionswerte des Sprichwortes in der Wochenzeitung, *Deutsche Sprichwörter in Literatur, Politik, Presse und Werbung.* Hambourg, pp. 11-41.

Mieder W. & G. B. Bryan (1996). *Proverbs in world literature: a bibliography.* New York : Peter Lang.

Paumier S. (2003). De la reconnaissance de formes linguistiques à l'analyse syntaxique. Thèse de Doctorat, U.F.R. d'Informatique. Université de Marne-la-Vallée.

Schapira C. (1999). *Les Stéréotypes en français: proverbes et autres formules*, Paris : Ophrys.

Tsaknaki O. (2005). The Proverb in Translation: Usage in Modern Greek and Automatic Treatment. PhD Thesis, Aristotle University of Thessaloniki, Thessaloniki.