Daniela Alderuccio

"A hypothesis of formalization of literary data for text analysis: a case study on Karl Kraus' writings"

Introduction

The growing availability on the Web of literary heritage is going to make easier humanistic researches, on the one hand facilitating access to information sources and documents and on the other hand providing a knowledge representation of texts, enabling its sharing and reuse. One of the major problems to face in knowledge representation is the formalization of literary data. The main difficulty is to capture the richness of word meanings into an established form, which allows automatic data treatment, preserving the essence of a thing anyway.

This challenge is related to the different nature of Computer Science and of the Humanities. The former has its foundation in establishing a formal representation of what exists (formal languages and modeling of reality); the latter is based on interpretation, whose subjectivity escapes from classification or rules. It is recognized that accuracy in literary analysis is related to cultural background and literary sensibility, but the underlying ambiguity of natural languages poses to researchers further difficulties: a specific term may have different or contradictory meanings and interpretations; authors frequently use different words or expressions to refer to the same meaning.

By developing common formalisms, Computer Science tools aim at reaching a sharable agreement on world representation. Similarly, in order to give an objective basis to concepts (starting point of the analysis), an application of this formal approach in the literary domain may allow experts to define and share a common vocabulary, to reach an agreement on word senses, thus reducing ambiguity.

In the hypothesis proposed in this paper, the use of a reference tool (such as an ontology) seems to offer a means to face this challenging task with success: by keeping from misunderstanding in reading texts and by limiting subjectivity in their analysis, the first expected result is a better comprehension of literary phenomena; by improving knowledge representation of a literary text, the second effect of formalization is the
retrieval of more relevant texts for research purposes.

### Application and Results

In the analysis of a literary phenomenon, some of the aspects to be considered are:

- the ambiguity of natural languages, that poses to experts problems in order to limit subjectivity in interpreting texts;

- and the heterogeneity of information sources to select (historical, cultural, geopolitical), that determines the need of retrieving relevant documents for the analysis.

Identifying criteria able to deepen the study of a literary phenomenon and to extract interesting documents on that subject, would be of great utility. The adoption of a linguistic resources (namely the ontology of WordNet [11]) as reference tool, seems to be a viable idea in order to reach both goals.

In order to test this approach in humanistic research, the "Dualism Truth vs. Propaganda" [2] in Karl Kraus has been investigated, using WordNet, the on-line reference system designed at the Cognitive Science Laboratory of the University of Princeton, to model lexical memory. Kraus was an Austrian intellectual and one of the bitterest satirists of fin-de-siècle Vienna, to be compared with Jonathan Swift for his satiric vision and command of language. He was a critic, a playwright, a poet, a journalist and the editor of the magazine "The Torch" - Die Packel [8]) - for about 36 years. Strongly believing in a language as a medium to express the truth, one of his major concerns was the German language and its misuse by the press. As a journalist he believed in informing the public rather than overwhelming it with propaganda: his main goal was to report facts, instead of interpreting them. Referring to this informative function of journalism, he wrote: "My duty is to say the Truth to Mankind"2.

Basing on Kraus' writings, the literary phenomenon under analysis has been synthesized into four keywords: "Language", "Truth", "Journalism", "Propaganda". The meanings of these selected terms have been defined using WordNet concept disambiguation. Because in this lexical database English nouns, verbs, adjectives and adverbs are organized into synonym sets called synsets (each representing one underlying lexical concept), disambiguation is based on lexical and semantic relations with other concepts.

Examination of WordNet definitions has led to: the exploration of keywords meanings; the delimitation of their semantic fields; and the finding of other related couples of opposing concepts such as: Truth vs. Verisimilitude, Language vs. Paralanguage, Journalism vs. Propaganda. The application of this ontology-based approach has been able to improve the comprehension of the "Dualism Truth vs. Propaganda" in Karl Kraus (1874-1936). As main consequence, by using WordNet it has been possible to study the literary phenomenon under analysis, confirming the validity of Kraus' position towards information problems and finding the core of the antagonism between "Propaganda and Truth".

As far as the second goal of this research is concerned (that is to find more relevant text for analysis), in order to apply the proposed approach, two sets of Kraus’ aphorisms
(Kraus, 1955) - »Writing and Reading« and »By Night«[4] - have been digitized. Then, by a human indexing operation performed using the ontology contained in WordNet, it has been assigned to each aphorism a category, based on semantic fields. The above selected keywords (»Language«, »Truth«, »Journalism«) have been adopted as indicator of semantic fields. Each aphorism has been labelled by the presence/absence of these fields. Despite the fact that »By Night« has no occurrences of the keyword »Journalism«, human analysis shows that it contains two relevant aphorisms5 for the comprehension of the »Dualism Truth vs. Propaganda« in Karl Kraus. In »By Night« the keyword »Journalism« is absent, but it is present the word »Zeitung« = newspaper, an implicit form, but semantically related to the keyword »Journalism«. If the goal of the search were to find all sets of aphorisms where Language and Truth and Journalism occur, probably this set of aphorisms would have been ignored, because not pertinent with the query. By defining semantic fields and categorizing aphorisms using them, the proposed approach has made possible to select »By Night« as a relevant document.

Conclusions

The achieved results show that literary data formalization based on ontologies is able to improve the accuracy of literary research. By including definitions of basic concepts in the domain (also in a machine-interpretable form), by identifying relations among them and by defining semantic fields, WordNet allows experts to share information in a domain, to provide critical notes and comments on texts, and to interpret them.

Furthermore, from this study emerges that defining the semantic field of words (by applying definitions provided by an ontology) and indexing documents by adopting a semantic categorization is an effective way of representing the content of a text: the faculty to bring to light word meanings, hidden in texts in an implicit form, improves the retrieval of more relevant documents, matching humanistic research needs.

References


**Notes**

1 »An ontology is a specification of a conceptualization (...)That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents« in T. Gruber. »What is an ontology?« URLhttp://ksl-web.stanford.edu/kst/what-is-an-ontology.html (T. R. Gruber. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993)

2 "Mein Pflicht ist es, den Menschen die Wahrheit zu sagen", Kraus K.: Die Fackel, Band 11, no. 852-856 (May 1931), p. 95


4 "Writing and Reading" und "By Night" have been extracted from "Dicta and Contradicta" (Spruche und Widerspruch), a selection of aphorisms appeared in "The Torch" and published in 1909.