



**HAL**  
open science

# User-Centered Analysis of Corpora using Semantic Features Redundancy

Thibault Roy, Pierre Beust, Stéphane Ferrari

► **To cite this version:**

Thibault Roy, Pierre Beust, Stéphane Ferrari. User-Centered Analysis of Corpora using Semantic Features Redundancy. 2008. hal-00203565

**HAL Id: hal-00203565**

**<https://hal.science/hal-00203565>**

Preprint submitted on 10 Jan 2008

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# User-centered Analysis of Corpora Using Semantic Features Redundancy

Thibault Roy, Pierre Beust and Stéphane Ferrari  
GREYC Computer Science Laboratory  
University of Caen (France)  
[troy,beust,ferrari@info.unicaen.fr](mailto:{troy,beust,ferrari}@info.unicaen.fr)

## 1. Introduction

Accessing textual information is still a complex activity when users have to browse through large corpus or long texts. In order to help users in such tasks, we propose a model dedicated to lexical representation of thematic domains as well as tools for personal corpora analysis.

The lexical model is a differential one, inspired by Saussure semiotics. It consists in structuring and describing lexical units by way of semantic features which are differences between terms meanings. Each thematic domain is represented by a set of terms characterized by many semantic features. These are built by the user using an interactive freeware tool developed by our team. Generally, domains include between 30 and 100 terms.

Lexical resources are identified in the corpus with the *ProxiDocs* tool. It returns interactive maps and reports built from the distribution of domains terms in the corpus. Maps reveal proximities and links between texts or sets of texts. The most often repeated semantic features in texts and in sets of texts are pointed out on the maps. According to the Interpretative Semantics of F. Rastier (Rastier, 1987), we call such a redundancy “intertextual isotopies”. These intertextual isotopies can represent redundancies of global domains which reveal topics of the considered texts, or can indicate a local semantic property, such as violence expression for instance, shared by some texts of the corpus.

In this paper, we present in a first section the lexical model we propose as well as the related tools for building personal lexical resources and interactively visualising them in a corpus. The second section deals notions linked with the semantic features and particularly with the intertextual isotopies. We also propose in this section methods to detect them in corpus. Section 3 presents two experiments in order to illustrate how such a redundancy can be usefully used for two kind of tasks: information retrieval in a Web pages corpus and semantic analysis of conceptual metaphors in a domain-specific corpus of newspapers. Finally, we conclude on the importance to take into account the intertextual isotopies, and more generally the global context established by the corpus, in tasks of access to information.

## 2. Interactions between Users and Texts

In our researches we are interested in electronic management of textual documents. For a lot of tasks of extraction and retrieval of information, the discovery of thematic domains in sets of documents is an important and often difficult analysis. We propose a model and several tools for this kind of analysis.

The tools *VisualLuciaBuilder* (<http://www.info.unicaen.fr/troy/lucia>) and *ProxiDocs* (<http://www.info.unicaen.fr/troy/proxidocs>) described here help their users allowing them to build and visualize lexical resources and graphical representations (called here “thematic maps”) of sets of documents. These maps allow users to discover thematic differences and similarities existing between each document of the analyzed set.

Our main propositions are that the model (and the tools that implement this model) we need for the management of textual documents, on the one hand, has to take into account more personal data because different subjects with different points of view can have different ways to understand a same text or set of texts and, on the other hand, has to allow more interactions between texts and the users because the understanding is an activity.

## 2. Building Personal Lexical Resources

Our model is called LUCIA (Perlerin, 2004) for Located User-Centered Interpretative Analysis. It is inspired by F. Rastier's works on Interpretative Semantics (Rastier, 1987). This model considers that the understanding of a text is a personal perception of meaning built within the redundancy of semantic features in the text, called Isotopies (Greimas, 66). These semantic features can be represented by the way of differential descriptions of the semantic content of used terms. LUCIA differential descriptions, called devices; are not supposed to be exhaustive, but they only reflect the author's opinion and vocabulary.

A device is a set of tables bringing together lexical units of a same semantic category, according to the user's point of view. In each table (for instance the table *agent* in the area 3 of the figure 1), the user has to make explicit differences between lexical units with sets of attributes (for instance *agent's type* used in the table *agent*) and values (for instance, *human, material, program* and *company* of the attribute *agent's type*). A table can be linked to a specific line of another table in order to represent semantic associations between the lexical units of the two tables. All the units of the second table inherit of the attributes and related values describing the row it is linked to.

The tool *VisualLuciabuilder* is an interactive user-centered application that allows its user to build LUCIA devices for the representation of the lexical domains of his choice, according to his own point of view. It allows a user for the step-by-step creation and revision of lexical resources through a graphical interface. This GUI (see figure 1) contains three distinct zones.

- *Zone 1* contains one or many lists of lexical units selected by the user. They can be automatically built in interaction with a corpus. The user can add, modify or delete lexical units.
- *Zone 2* represents one or many lists of attributes and values of attributes as defined by the user.
- *Zone 3* is the area where the user "draws" his LUCIA devices. He can create and name new tables, drags and drops lexical units from zone 1 into the tables, attributes and values from zone 2, etc. He can also associate a colour to each table and device.

The LUCIA device showed by the Figure 1 is made of 4 tables using 4 attributes. It provides a semantic knowledge representation including almost 30 terms (as *computer, internet, bug, web site* or *IBM* for instance). From this LUCIA device, descriptions of the semantic content of terms can be extracted. For instance the term *hacking* is represented by the following semantic features : *Activity's type : non professional, link with domain : activity*. This LUCIA device is a small example. Generally the devices we use contains between 60 and 100 lexical units. The LUCIA lexical resources are therefore not very large in comparison to a lexical data base. Aims are not the same because a lexical data base indicates a shared representation of meaning for a large community of speakers. In opposition a LUCIA device indicates a very specific point of view of one person or a small group of persons on a quite small set of words. This is why this kind of lexical resource can be revised easily step-by-step as long as the user (which has not to be specialist in lexicology) use it.

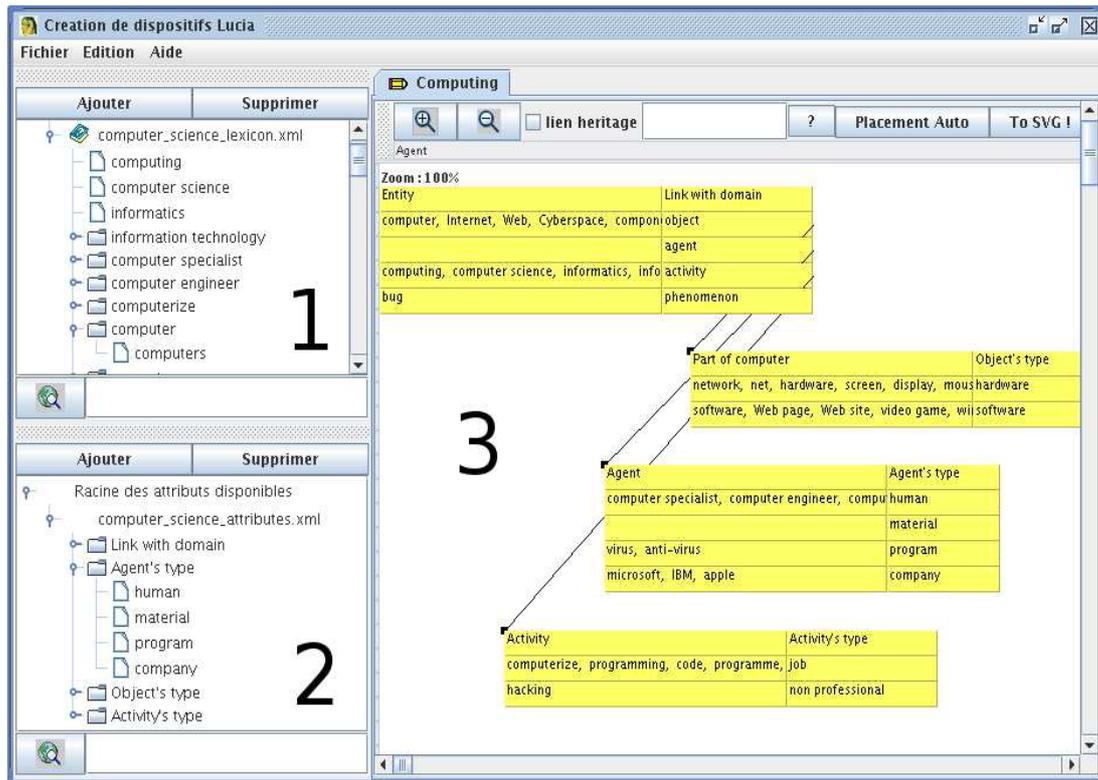


Figure 1: VisualLuciaBuilder's interface

## 2.2. Corpora visualisation using personal lexical resources

The *ProxiDocs* (Roy and Beust, 2004) tool builds global representations from LUCIA devices and a collection of texts. The first stage in the *ProxiDocs* process consists in counting terms of each LUCIA device in each document. We associate a list of numbers with each document, and, each list constitutes a N-Dimensional vector (N is the number of devices specified by the user in his lexical resource). The next stage consists to project these N-Dimensional vectors on a 2 or 3 dimensional space using statistical methods (such as the Principal Components Analysis (PCA) method or the Sammon method). Each document is then represented by a point that maps can show. At least, in order to underline subsets of documents with similar semantics, we use a clustering method called Ascendant Hierarchical Clustering. Following this process *ProxiDocs* maps show the distribution of the lexicon of the user's LUCIA devices in the corpus. This is useful to reveal proximities and links between texts or between sets of texts.

For instance, an experimentation of *ProxiDocs* from a corpus of around 800 articles of the French newspaper "Le Monde" of 1989 (around 700,000 words) and a generalist set of lexical resources can reveal the two kinds of maps presented in figure 2.

*ProxiDocs* can build several kind of maps (see <http://www.info.unicaen.fr/troy/lucia> for examples of maps) in order to suggest to the user many global visualisations of his corpus according to his lexical resources. These maps are :

- *Maps of documents of the corpus* in 2 dimensions (left map of figure 2) or 3 dimensions. Each point on kind of map represent a document of the analyzed corpus and the color of each point is related (by the way of the map's legend on the bottom which indicate the several LUCIA devices of the user's lexical resource) to the main thematic domain of the document.



- Lexical labels : on a cluster of document map, user can demand that each cluster can be labeled with the five most frequent lexical unit. These labels are interactive because when a lexical unit is pointed with the mouse, this term is colored in red in all the labels of the map. This can quickly reveal the main lexical redundancy between clusters.
- Interactive legend : when user points a part of the legend a in a document map, all documents related to the pointed thematics are put in a prominent position. This allow the user to have a visualization of a semantic proximity according to specific sub-set of his lexical ressources between texts all over the map.
- Chronologic maps : when the documents of the user's corpus are timed stamped (as for instance the documents in a corpus of actuality reports), ProxiDocs can build interactive maps of clusters of documents which can automatically change regarding to the main periods in the temporality of the corpus. This can show thematic groups of texts that appear or disappear at several dates in the duration of the corpus.
- Hyperlinks : on a document map Each point is an hyperlink to the represented document where words of the LUCIA devices found in the text are colored with the apporiate color according to the legend. On a cluster of document map, each disc is also an hyperlink to a web page that provides a report on the cluster. This report indicates the lexical units sorted by frequency and the main redundancy of semantic features founded in the texts of the cluster. Theses extract what we call *intertextual isotopies* and this is what we detail in the following part of this paper.

### 3. Semantic Features Redundancy

Previously, we presented the LUCIA model and tools for building and exploiting lexical ressources in this model. Cartographic interactive views are detailed. Such views reveal many elements including reports on texts and sets of texts. These reports contain a lot of informations, including the intertextual isotopies which are, for us, elements of meaning in the texts. In this section, we first present the notions linked to such isotopies, then, we explain how we detect them in texts.

#### 3. Notions used

A **lexical unit** is a functional unit composed of many morphemes and which can correspond to one or more words. Two types of lexical units can be distinguish : simple lexical units and complex lexical units, according they describe on graphical word (for instance, as *water* or *pretty*) or many graphical words (for instance, as *remote control* or *french fries*). In (Rastier *et al.*, 1994), a lexical unit is defined like a "meaning unit". The authors consider it as a base to each semantic analysis of texts.

A **seme** or a **semantic feature** is considered as the smallest meaning unit (Rastier, 1987), a seme is often formulate between two lines. For instance, the lexical unit *dog* could be characterised by the seme */mammal/*, */bark/*, */domestic animal/*, etc.

An **isotopie** is by definition (Rastier, 1987) a redundancy effect of the same seme in a text. This redundancy effect allows to characterize and to identify the intensity of some semes in a sentence, a text. For instance, in the sentence *the postman brings a letter to me*, the seme */mail/* is associated to the lexical unit *letter* because it is repeated in the lexical unit *postman*, building by this way an isotopie. Such a redundancy enables the relevant interpretation of *letter* in the sentence. It is not retained, for example, the meaning of letter as a letter of the alphabet.

An isotopie is by default intra-textual. The redundancy effect is considered in the context of a single text. A larger redundancy effect, “traveling” in all the texts of a set, can be considered too. Such a redundancy reveals global meaning informations on the set and could be very important in tasks of access to information in many texts. This seme redundancy in many texts is that we called a **intertextual isotopie**.

In the LUCIA model, we consider as semes the attributes defined by users to characterize the lexical units describing the domains of their choice. Redundancies of attributes in texts or in sets of texts of the corpus are respectively for us intratextual and intertextual isotopies. Such isotopies bring meaning informations and particularly, intertextual isotopies which “carry” global meaning informations on the corpus locally shared by texts or sets of texts.

### 3. *Intertextual Isotopies Computation*

In order to detect such isotopies in a corpus or in a subset of texts of the corpus, we propose to project users' LUCIA devices in the texts of the corpus. This projection brings to the fore in each text the lexical units of the devices as well semes they carry. The first step consists the determination of the intratextual isotopies in each text. This determination consists in counting each seme associated to each lexical unit. Thus, a list of the repeated semes is built for each texts. The most repeated semes corresponding to the intratextual isotopies of the text.

The next step consists in compute the intertextual isotopies shared by texts of a same set. In this computation, we take into account the global context of the corpus and the generic isotopies it contains. By this way, we try to minimize global and generic informations carrying by the corpus in order to realise a sharp and discriminating analysis of texts and subsets of texts of the corpus.

We first propose to mesure the part of an intertextual isotopie in a set of texts with the following formula :

$$part(isotopie, set\ of\ texts) = \frac{100 \times \text{number of redundancies of the seme associated to the considering isotopie in the set of texts}}{\text{number of redundancies of all semes defined by the users in their LUCIA devices}}$$

Then, to take into account the real weight of each intertextual isotopie in a set of texts according to the corpus, we define the formula below :

$$weight(isotopie, set\ of\ texts) = part(isotopie, set\ of\ texts) - part(isotopie, corpus)$$

According to the sign of the obtained value, an “excess” or a “deficit” of the considered isotopie is observed in the sets of texts in comparison to the global level of the corpus. A null value indicates that the isotopie is present in a same way in the set of texts and in the corpus.

Such informations are significant to characterize the local level of the set of texts in respect to the global level of the corpus. Thus, it is possible to present to the users the most excessed isotopies associated to each set of texts. Such isotopies bring to the fore the contribution of the set of sets to the corpus, describing like this the set of texts in a relevant way.

The next section of this paper illustrates such computation and its significant participation to two tasks different tasks of access to information in texts.

## 4. Experiments

### 4.1 Experiment 1 : a Study of Metaphorical Expressions

The aim of the IsoMeta experiment is to observe how conceptual metaphors (Lakoff and Johnson 1980) are used in a domain-specific corpus. The corpus used is constituted of about 600 articles from the French newspaper “Le Monde”, all about Stock Market. Three conceptual metaphors have previously been observed in this domain: the War in Finance, the Health of Economics and the Meteorology of Stock Market. One device was built for each source domain: War (in red in figure 4), Health (blue) and Meteorology (green).

We showed in (Roy and Ferrari 2007) the resulting map of the corpus reveals what we called the *metaphoricity* of the documents. In the bottom of the map, for instance, group 14 in figure 4, the lexicon related to these source domains is mostly used in a metaphorical way.

We propose to illustrate the effect of weight computing in this context of metaphorical interpretation.

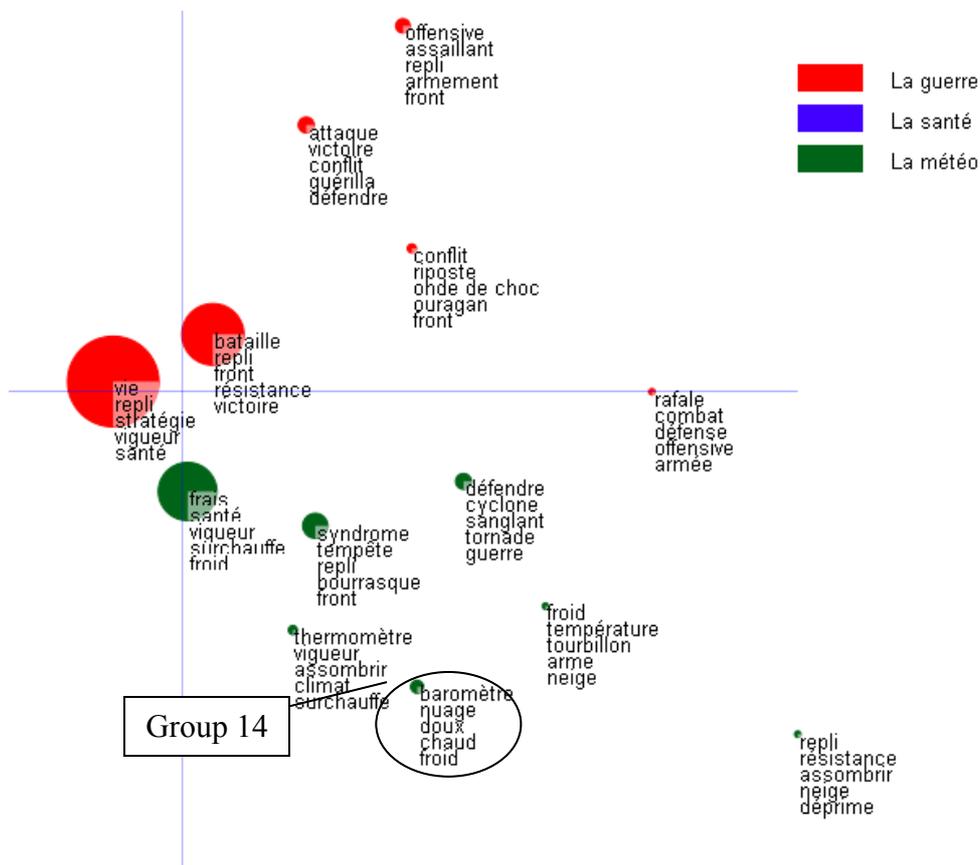


Figure 4: IsoMeta Map of the Corpus

This group is constituted of 4 documents. In these documents, vocabulary from the meteorology domain is used mostly in a metaphorical way to talk about Stock Market phenomena.

Attribute	Score without weight	Score with weight
Axe ( <i>axis</i> )	25.8% (position 3)	17% (new position 1)
Evaluation ( <i>opinion</i> )	26.7% (position 2)	5.7% (new position 2)
Rapport au domaine ( <i>role in the domain</i> )	33.9% (position 1)	0.1% (new position 3)
Direction ( <i>direction</i> )	4.4% (position 5)	-0.4% (new position 4)
Fonction ( <i>function</i> )	8.9% (position 4)	-5.8% (new position 5)

Table 1: Distribution of scores with and without weights in the study of metaphorical expressions

In table 1, we can observe two main differences between the positions before and after weight computing. The first one is the order of the three first attributes characterizing the documents of this group. The “role in the domain” attribute was the most important in the group. This is mostly due to its position in the devices: it is a generic attribute shared by almost all the lexical entries of our devices. When interested by interpreting the documents of our corpus, its importance has to be reduced: all the documents of our corpus share this attribute. The weight computing sets to it an almost null value (0.1%), which means that “role is the domain” is not relevant to characterize what the documents of this specific group 14 are about. On the opposite, the “axis” and “opinion” attributes are bringing to the fore after weight computing. They are the attributes which locally characterize in the best way the documents in the group. The second main difference concerns the two last attributes, which are characterized, after weighting, by a negative score. They can therefore be considered as not relevant in this group.

When interpreting the documents of the group, we observe that these attributes are linked to lexicon from the meteorology device. The related words are used in a metaphorical way to describe changes of Stock Market values, as well as to give an opinion on these tendencies. The “axis” attribute is related to the physical dimension of the source domain: temperature, wind strength, and so on, for their measurable aspects. The metaphorical use of the words should lead to an interpretation in terms of measurable phenomena of the Stock Market. The current model does not allow for the substitution the metaphor implies, but the user who built the source devices is aware of this possible constraint: the values of some attributes may be substituted for a more accurate interpretation. The three remaining attributes are not relevant to characterize the documents of this group 14. The “role”, “function” and “direction” attributes are generic attributes shared by all the documents in the corpus.

#### 4.2 Experiment 2 : a Task of Information Retrieval on the Web

The first experiment illustrates our propositions for the study in corpus of three conceptual metaphors. This second experiment concerns information retrieval on the Web. The objective is to perform a search for information on the Web in a broad context: the “European decisions”. This search is realized with regards to the domains interesting the user. The domains representing the user's point of view are *agriculture*, *pollution*, *road safety*, *space*, *sport* and *computer science*. The corresponding LUCIA devices contain from 3 to 5 tables and from 30 to 60 lexical units.

In order to constitute the collection of texts, the key words “European decision” have been searched using the Yahoo search engine (<http://www.yahoo.com>) for texts in English. The returned first 150 links were automatically collected. The textual part of these documents, which were in three formats, HTML, PDF and DOC, were automatically isolated in order to constitute a corpus of text documents, each one between 1,000 and 50,000 tokens. As the previous experiment, the *ProxiDocs* tool is used in order to project the devices in the corpus, building a map of texts (Figure 5).

For a detailed presentation of this map, see e.g. (Roy and Ferrari, 2007). The intertextual isotopies detected in the marked group 3 on the map are detailed in Table 2.

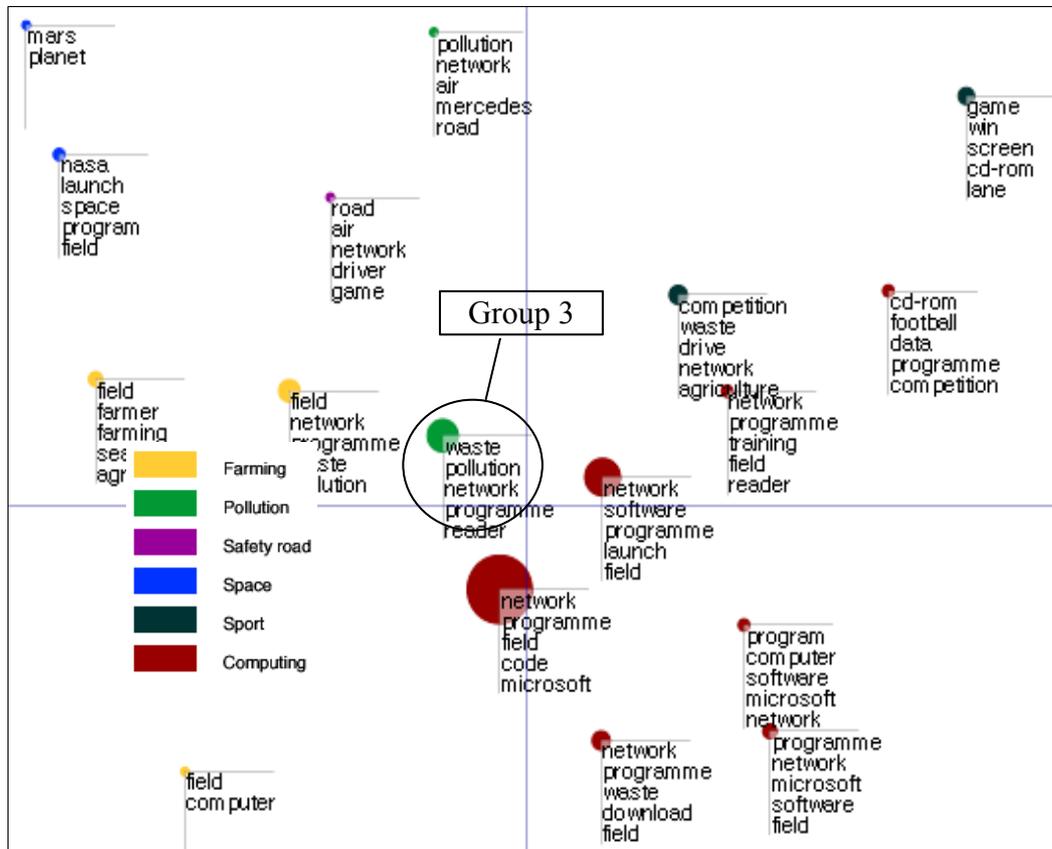


Figure 5: Map of the Web pages collection used during the second experiment

<i>Attribute (most repeated value)</i>	<i>Score without weight</i>	<i>Score with weight</i>
<i>Link with domain (object)</i>	70.5 % (pos: 1)	14.9% (pos: 1)
<i>Evaluation (bad)</i>	2.5% (pos: 3)	1.0% (pos: 2)
<i>State (gas)</i>	2.0% (pos: 4)	0.7% (pos: 3)
<i>Agent's type (organization)</i>	1.4% (pos: 5)	-0.02% (pos: 4)
<i>Object's type (hardware)</i>	16.0% (pos: 2)	-0.06% (pos: 5)
<i>Activity's type (job)</i>	0.02% (pos: 6)	-0.08% (pos: 6)

Table 2: Distribution of scores with and without weights of the group 3 in the second experiment

Briefly, the distribution of the intertextual isotopies scores of the group depreciates the attribute “Object's type” from the second position to fifth position. The attributes “Evaluation”, “State” and “Agent's type” win a place in the ranking. The attribute “Link with domain” and the attribute “Acivity's type” stay respectively in first and last position.

An analysis of the documents of the group reveals that it deals with problems related to the pollution, and more particularly with European decisions on the sustainable development. Thus, the weighting of the attributes of the group is quite relevant. The attributes “Evaluation” with the value “bad” and “State” with the value “gas” are brought to the fore which is really agreed with the main theme of the documents. On the opposite, the generic attribute “Object's type” is depreciated, it is not quite related to the content of the documents of the group.

## Conclusion

In this paper, we presented a centred-user approach for accessing textual information in corpora. Based on a model for semantic representation of domains, a set of interactive tools have been developed to help the user to specify his own point of view on a domain and using this knowledge to browse through a collection of texts. The notion of intertextual isotopie is then used in order to bring to the fore relevant semantic information on the analyzed texts. Two very different experiments illustrate their use and interest.

Such results raise interesting questions about the role of the tools in tasks of access to textual information. It is very useful to have a semantic representation of users' domains of interests and to use such a representation for textual analysis. By this way, we showed that basic functions of semantic features scoring can reveal interesting contents and, thus, can help users in tasks of access to textual information.

## References

- Greimas, A.J. (1966) *Sémantique structurale* (Paris: Larousse).
- Perlerin, V. (2004) *Sémantique légère pour le document* (PHD Thesis of the University of Caen).
- Rastier, F. (1987) *Sémantique interprétative* (Paris: Presses Universitaires de France)
- Rastier, F., Cavazza, M. and Abeillé A. (1994) *Sémantique pour l'analyse* (Paris: Masson).
- Roy, T. and Beust, P. (2004) ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus in *Proceedings of the 7<sup>th</sup> International Conference on Textual Data Analysis*, 978-987.
- Roy, T. and Ferrari, S. (2007) User Preferences for Access to Textual Information: Model, Tools and Experiments in *Advances in Semantic Media Adaptation and Personalization*, Studies in Computational Intelligence, Springer Verlag (to appear).