# User-Centered Analysis of Corpora
# Using Semantic Features Redundancy

Thibault Roy,[1] Pierre Beust[1] and Stéphane Ferrari[1]

## 1. Introduction

Accessing textual information is still a complex activity when users have to browse through large corpora 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's semiotics. It consists in structuring and describing lexical units by the 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 through an interactive tool developed by our team. Generally, domains include between 60 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, 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 an expression of violence for instance, shared by some texts of the corpus.

In this paper, first we present the lexical model as well as the related tools for building personal lexical resources and interactively visualising them in a corpus. The second section deals with notions linked to 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 useful for two kinds 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 generaly 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 many tasks of information extraction and retrieval, 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.

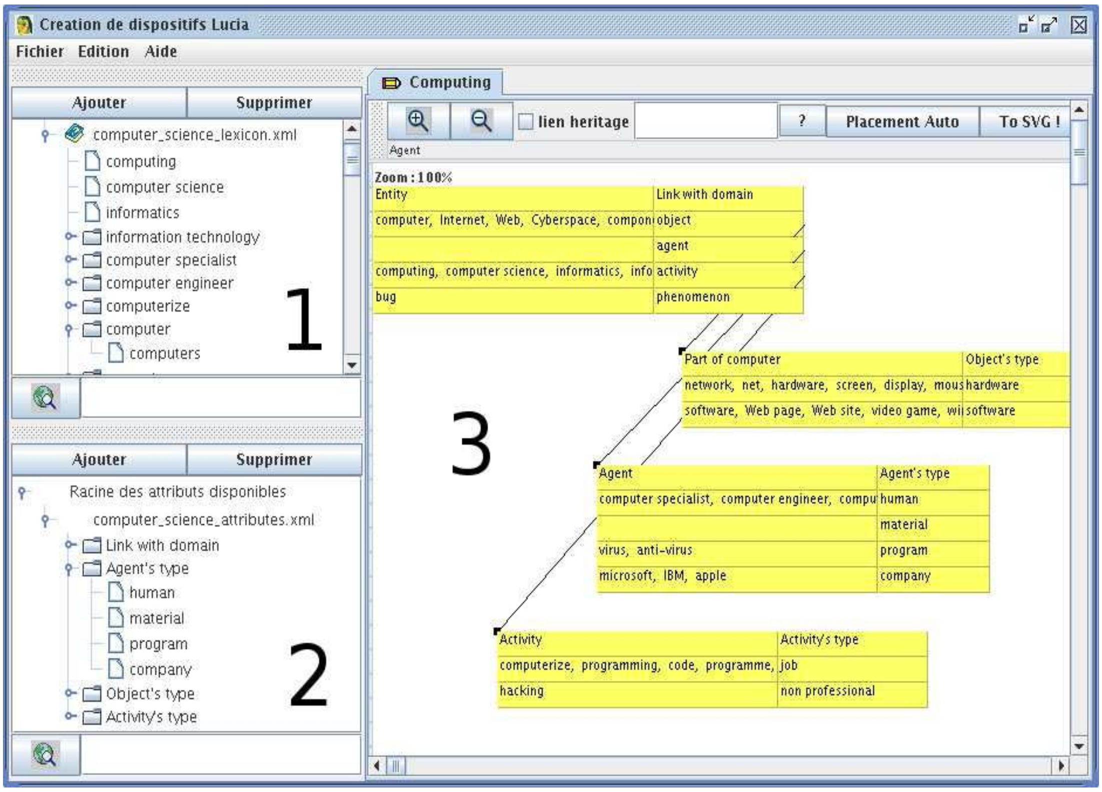
[1] GREYC Computer Science Laboratory, University of Caen Basse-Normandie (France)
*e-mail*: Thibault.Roy@info.unicaen.fr, Pierre.Beust@info.unicaen.fr, Stephane.Ferrari@info.unicaen.fr

The tools *VisualLuciaBuilder* (http://www.info.unicaen.fr/~troy/lucia) and *ProxiDocs* (http://www.info.unicaen.fr/~troy/proxidocs) help their users building and visualizing lexical resources and graphical representations (called "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 personal data because different users 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 interactions between texts and the users because the understanding is an activity.

## 2.1 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). We consider with this model that the understanding of a text is a personal perception of meaning built within the redundancy of semantic features, called *isotopies* (Greimas, 66). These semantic features can be represented by the mean of differential descriptions of the semantic content of used terms. LUCIA differential descriptions, called *devices*, are not supposed to be exhaustive, but they reflect the author's opinion and vocabulary only.



| Entity | Link with domain |
|---|---|
| computer, Internet, Web, Cyberspace, compon | object |
| | agent |
| computing, computer science, informatics, info | activity |
| bug | phenomenon |

| Part of computer | Object's type |
|---|---|
| network, net, hardware, screen, display, mous | hardware |
| software, Web page, Web site, video game, wi | software |

| Agent | Agent's type |
|---|---|
| computer specialist, computer engineer, compu | human |
| | material |
| virus, anti-virus | program |
| microsoft, IBM, apple | company |

| Activity | Activity's type |
|---|---|
| computerize, programming, code, programme, | job |
| hacking | non professional |

**Figure 1**: VisualLuciaBuilder's interface

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 allowing its user to build LUCIA devices for the representation of the lexical domains of his choice, according to his own point of view. It provides a user with a graphical interface for the step-by-step creation and revision of lexical resources. 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 figure 1 is made of 4 tables. It provides a semantic knowledge representation including almost 30 terms (as *computer*, *internet*, *bug*, *web site* or *IBM* for instance). Descriptions of the semantic content of terms can be extracted from this LUCIA device. 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 compared to a lexical database. Aims are not the same because a lexical database 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 these lexical resources can be revised easily. Moreover, the user has not to be a specialist in lexicology.

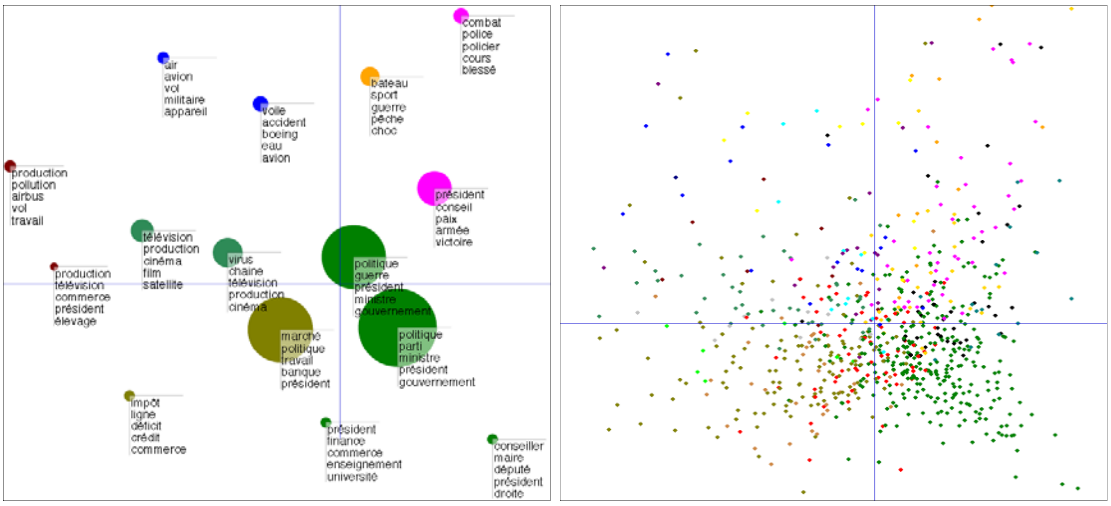
## 2.2 Corpora visualisation using personal lexical resources

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 in projecting 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 visualized on maps. At last, 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.

Maps of figure 2 are resulted after an experimentation of *ProxiDocs.* It was realized on a corpus of around 800 articles from 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 the user many global visualisations of his corpus according to his lexical resources. These maps are:

- *Maps of documents of the corpus* in 2 or 3 dimensions (left map of figure 2) or 3 dimensions. Each point on this kind of map represents a document of the analyzed corpus and its color is related to the user's domain the most frequent in the document, by the way of the legend reminding the user's devices.
- *Maps of sets of documents* also available in 2 or 3 dimensions (right map of figure 2) where discs represent clusters of semantically close documents. The disc's size is proportional to the number of documents contained in the cluster. Its colour is the one of the most represented device in the cluster.
- *Clouds of the lexical units* appearing in the corpus inspired by the web site *TagCloud* (http://www.tagcloud.com). A cloud (as figure 3 shows) reveals which lexical units from the selected devices have been found in the documents of the corpus. They are sorted out in alphabetical order and their size is proportional to their number of occurrences in the corpus.



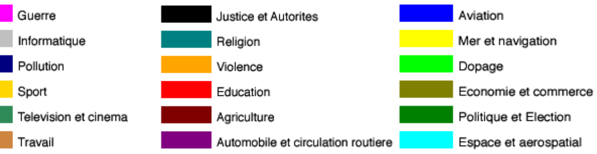


**Figure 2**: Examples of corpus' maps

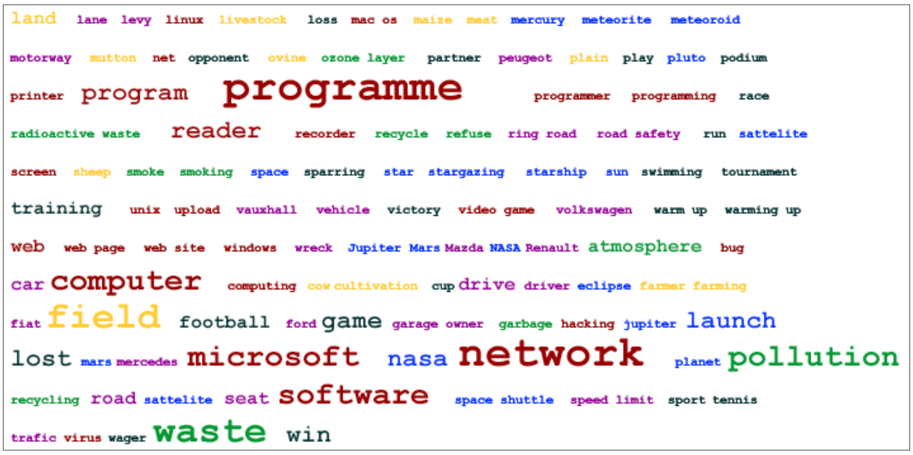

land lane levy linux livestock loss mac os maize meat mercury meteorite meteoroid
motorway mutton net opponent ovine ozone layer partner peugeot plain play pluto podium
printer program programme programmer programming race
radioactive waste reader recorder recycle refuse ring road road safety run sattelite
screen sheep smoke smoking space sparring star stargazing starship sun swimming tournament
training unix upload vauxhall vehicle victory video game volkswagen warm up warming up
web web page web site windows wreck Jupiter Mars Mazda NASA Renault atmosphere bug
car computer computing cow cultivation cup drive driver eclipse farmer farming
fiat field football ford game garage owner garbage hacking jupiter launch
lost mars mercedes microsoft nasa network planet pollution
recycling road sattelite seat software space shuttle speed limit sport tennis
trafic virus wager waste win

**Figure 3**: An example of a corpus' cloud

The *Proxidocs*' maps are built to propose the user a set of interactions with his corpus and his personal lexical resources. It is by the way of these interactions that the user is able to build his own understanding of texts. The allowed interactions are:

- Lexical labels : on a cluster of document map, a user can ask for each cluster to be labeled with the five most frequent lexical units. These labels are interactive. When the mouse is over a lexical unit, this term is colored in red any place it appears on the map. This can quickly reveal the main lexical redundancy between clusters.
- Interactive legend : when a user points a part of the legend, all the related documents are emphasized. This allows the user to have a visualization of a semantic proximity between texts all over the map according to a specific sub-set of his lexical resources.
- Chronologic maps : when the documents of the user's corpus are timestamped (e.g. actuality reports), *ProxiDocs* can build dynamic maps of clusters showing their chronological evolution. The periodicity is a parameter. Some groups can appear or disappear.

- Hyperlinks : on a document map each point is a hyperlink pointing to the corresponding document. Words from the LUCIA devices are colored in the text according to the legend. On a clusters map, each disc is also a hyperlink pointing to a report on the cluster. It indicates the lexical units sorted by frequency and the main redundancies of semantic features found in the texts of the cluster. We call these redundancies *intertextual isotopies*, we detail this notion next.

## 3. Semantic Features Redundancy

Previously, we presented the LUCIA model and tools for building and exploiting lexical resources in this model. Cartographic interactive views are detailled. 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.

### 3.1 Notions used

A **lexical unit** is a functional unit composed of many morphemes and which corresponds to one or more words. Two types of lexical units can be distinguished : simple ones, composed by a single graphical words (e.g. *water*, *pretty*) and complex ones, composed by more than one graphical words (e.g. *remote control*, *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 conventionally written between two slashes. 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 a same seme in a text. This redundancy effect allows the identification of pertinent 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 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. According to our opinion, redundancies of attributes in texts or in sets of texts are respectively intratextual and intertextual isotopies. Such isotopies bring meaning informations and particularly, intertextual isotopies which “carry” global meaning informations on the corpus localy shared by texts or sets of texts.

### 3.2 Intertextual Isotopies Computation

In order to detect such isotopies, 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 the computation of 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 present in the corpus in order to realise a precise 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 :

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

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 :

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

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 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 Meteoroly 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.

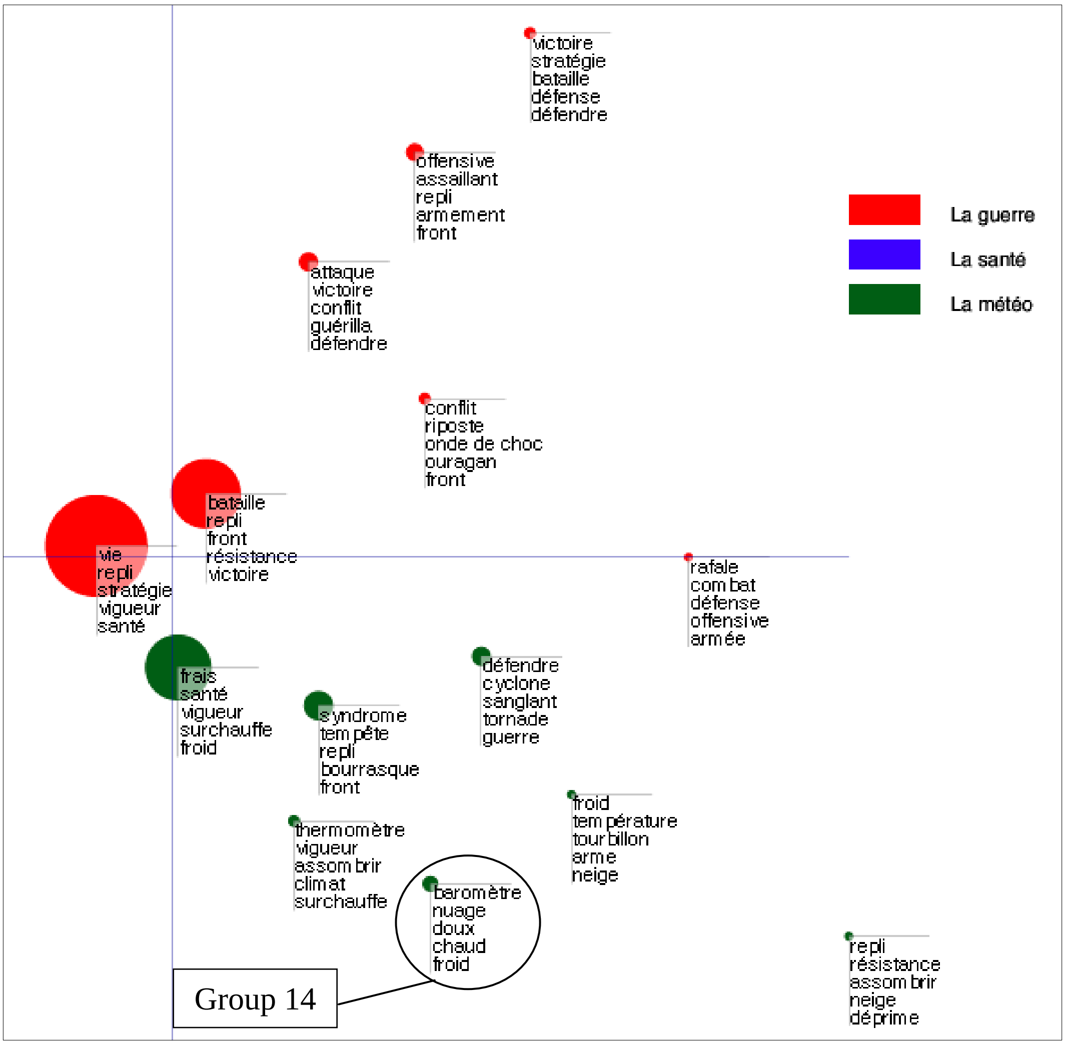


**Figure 4**: IsoMeta Map of the Corpus

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 (*axis*) | 25.8% (position 3) | 17% (new position 1) |
| Evaluation (*opinion*) | 26.7% (position 2) | 5.7% (new position 2) |
| Rapport au domaine (*role in the domain*) | 33.9% (position 1) | 0.1% (new position 3) |
| Direction (*direction*) | 4.4% (position 5) | -0.4% (new position 4) |
| Fonction (*function*) | 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 detailled presentation of this map, see e.g. (Roy and Ferrari, 2007). The intertextual isotopies detected in the marked group 3 on the map are detailled in Table 2.

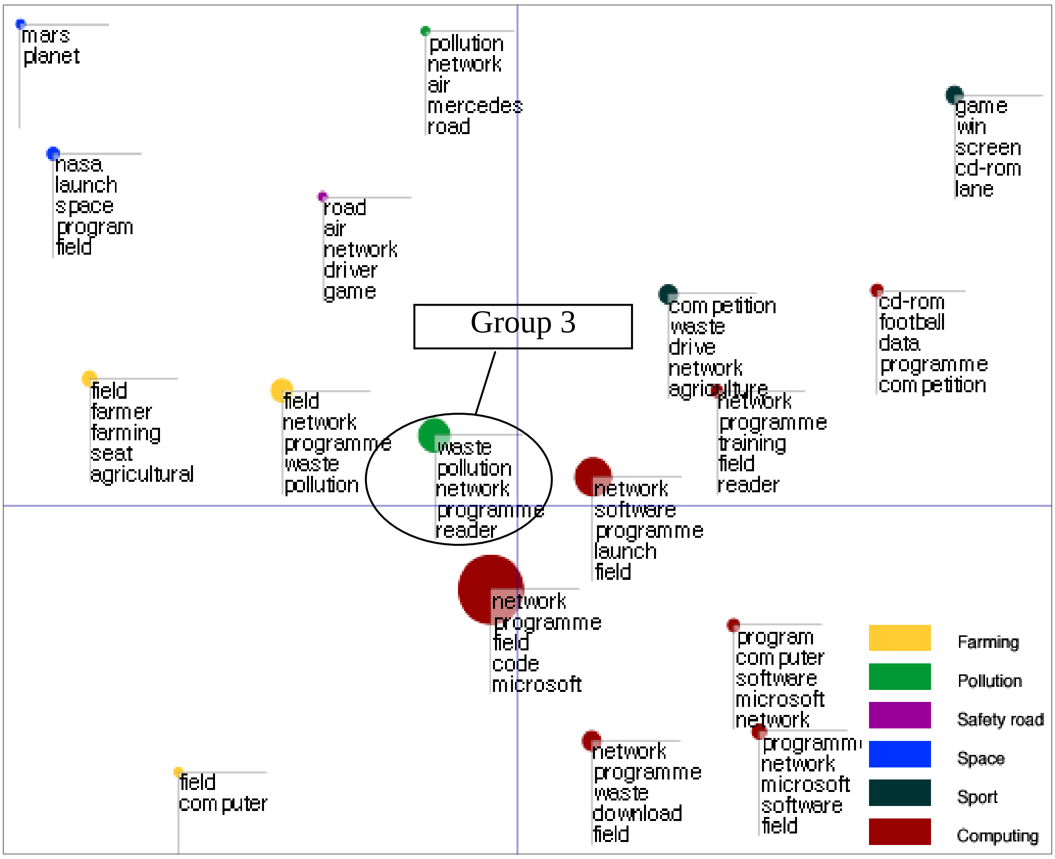


**Figure 5**: Map of the Web pages collection used during the second experiment

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

## 5. 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 illustrated 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 7th International Conference on Textual Data Analysis*, 978–87.

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).