

Image and Text Mining Based on Contextual Exploration from Multiple Points of View

Florence Le Priol

LaLIC-STIH, EA 4509, Paris-Sorbonne University
Maison de la Recherche, 28 rue Serpente, 75006 Paris, France
florence.le_priol@paris-sorbonne.fr

Abstract

In this paper, we present an image and text mining tool named TNT. This tool is based on Contextual Exploration and work on different points of view. It can process a corpus of all sizes in French or in English. The web interface associated with this tool, offers a reorganization of the text guided by the images and annotated segments that are associated.

What is Contextual Exploration ?

Founding hypothesis

Contextual Exploration (CE) is a operational semantics method for corpus analysis according to different points of view. CE is independent of the corpus studied.

The founding hypothesis of CE, as explained J.-P. Desclés (Desclés 1997), is the following: texts contain specific linguistic units that are relevant indicators for completing a precise task. However, the identification of these indicators is not sufficient. The analysis of identified linguistic unit in a context necessarily involves other complementary linguistic clues that must be co-present in the context. These clues participate directly in solving the task.

Therefore, for any given task, we must identify linguistic units and provide a procedure which involves exploring the context in seeking certain relevant linguistic clues that are co-present in the context, so as to advance in solving the task. The task is to annotate textual segments (sentences or paragraphs) where an indicator and the contextual clues associated with the point of view to search (location, part-whole, definition, events ...) have been identified.

Give an example to illustrate this hypothesis.

The task is to identify segments of text containing a part-whole relationship. Take these two sentences:

- (1) Les avocats de la **partie** civile ont réclamé cinq milliard de francs cfa à titre de dommage et intérêts.

(Lawyers for the plaintiff claimed five billion CFA as damages and interest.)

- (2) La myologie est une partie de l'anatomie qui traite des muscles. (Myology is part of anatomy that concerns muscle.)

The term “partie” (part) is considered as an indicator for this point of view, but its presence does not suggest that it is a part-whole relationship. Indeed, in the sentence (1), there is no part-whole relationship. In the sentence (2), the presence of complementary clues in the context of the indicator, “est une” (is a) on the left and “de” (of) on the right, allows to conclude to a part-whole relationship.

The EXCOM platform

CE is operational and implemented in the EXCOM platform (Djioua & al. 2006, Alrahabi 2010). The heart of the EXCOM platform is the module of semantic annotation. It's a decision procedure presented in the form of a rule base associated with linguistic resources organized by points of view with a semantic map.

This module of semantic annotation is preceded by a module of segmentation into sections, paragraphs and sentences. According to each task, specific modules of pre-processing can be added before the module of segmentation, for example a module to transform html to txt, and some modules of post-processing can be added after the module of semantic annotation, for example a module to do summary sheet, a module to link images and text...

In its current version, the EXCOM platform (www.excom.fr) is implemented in Java. The linguistic resources are structured in XML files.

Points of view and semantic map

The semantic annotation module of the EXCOM platform uses, as we have explained, Contextual Exploration where linguistic resources are organized by points of view based on semantics maps.

Different points of view have been studied and are

implemented in the EXCOM platform:

- location relation between concepts;
- direct and indirect quotations;
- definition;
- argument (hypothesis, conclusion...);
- the notion of encounter (who met whom?);
- the notion of causality;
- bibiosemantic (which authors are cited and how?);
- links between images and associated textual comments.

These points of view are organized in a semantic map, a kind of general ontology. The semantic map structures indicator classes, clue classes and CE rules that lead to decision making.

As an example, consider the semantic map of location relations (see Figure 1).

In French, location relations are narrowly attached to the linguistic expression of primitives "*est*" and "*a*" (Desclés 87). Static relations are binary relations which make it possible to describe states (static situations) in the expert domain.

The function of the archirelator of location is to establish a bond of location between a located entity and a locating entity; it is a general statement of relation. Identification, localization, whole-and-part, attribution, and ruption are its specifications.

The semantics of each location relation corresponds to intrinsic properties: functional type (standard semantics of relation arguments); algebraic properties (reflexivity, symmetry, transitivity, ...); properties of fitting in combination with the other relations in the same context (i.e. in a given static situation).

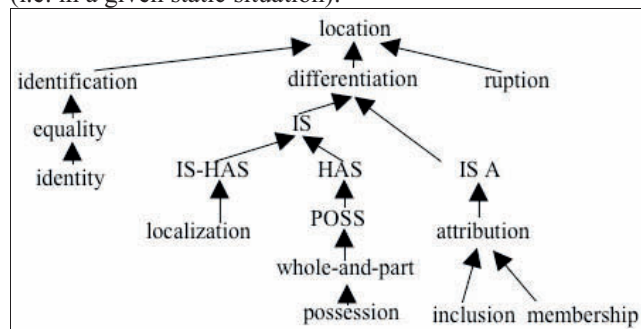


Figure 1. Semantic Map of location relations (Le Priol & al. 2006)

TNT: image and data mining tool based on Contextual Exploration

TNT (Textual – Non Textual) is a tool based on Contextual Exploration and the EXCOM platform, bacteria (white dots) induced polymerization of actin (green) at the entry site (arrow).used to annotate the information concerning non textual data and make the link between annotated text and corresponding non textual data (fixed or moving

images, photographs, video, sound...). Indeed, from texts of the web, the interface presents, after a fully automatic process, the images of the text with their caption and text segments, located at any position in the text, referring to image as shown in Figure 2.

Figure 2. Example of TNT treatment on french corpus¹

This screenshot presents a extract of result obtained for a french scientific paper in biology. It shows the image followed by its caption and a text segment commenting on this picture. The caption ('Légende') is a paragraph, the comment ('Commentaire direct') is a sentence. It is to be noted that the segment with "direct comment" annotation, corresponding to segment No 99, is far enough away from the image and its caption (segment No. 34).

In fact, the TNT tool proposes a complete reorganization of the text based on images, captions and associated

¹ Figure 6. Chlamydia cause the reorganization of the actin cytoskeleton. Bacteria (white dots) induce polymerization of actin (green) at the entry site (arrow). Cells fixed 10 min after infection. Caption, segment No 34 In fact, the entry of Chlamydia is more like a phenomenon of phagocytosis, although it takes place in epithelial cells (Figure 6). Direct Comment, segment No 99

comments as shown in Figure 3. In this representation, we have the original text on the left and the result of TNT treatment on the right. We can see that comments on the first image (in yellow) come only from the top of the text. However, comments for the second (in pink) and the third (in blue) images are distributed throughout the text. Furthermore, we see that comments related to the second image interspersed in the original text, with comments associated with the first and the third images. So the result proposes a reorganization of the original text: each image with its caption and all associated comments.

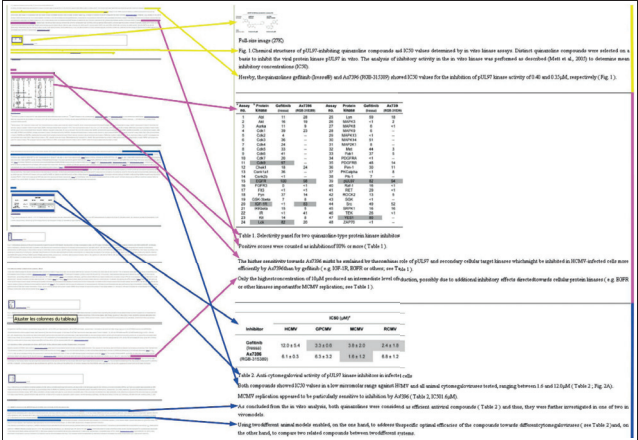


Figure 3. Reorganization of the text proposed by TNT

The process

The process that we propose leads from a text in html to an interface for image and text mining.

Most often, the texts that we treat are HTML files but the module of segmentation accepts, as input, only TXT files. So, the first step is to convert HTML to TXT while keeping memory of the location of the images for further processing.

The second step is segmentation into sections, paragraphs and sentences. The semantic annotations identified during the third step, will, for this task (link between images and text), on the paragraphs or sentences.

The annotated segments are either caption (annotation: Legende) or direct comment (annotation: Commentaire direct).

Captions are generally identified by markers such as 'Figure 1', 'Table 3'... In the example of the Figure 4, the sentence is annotated 'Legende' because of the presence of 'Figure 1.'

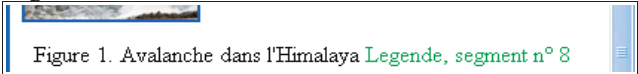


Figure 4. Example of sentence annotated as a caption²

Comments are identified either by the presence of a reference to an image (see Figure 5), or with markers like 'la photo ci-dessus montre' ('the photo below shows'), 'This is an overview' (see Figure 6).

² Figure 1. Avalanche in the Himalayas Caption, segment N°8

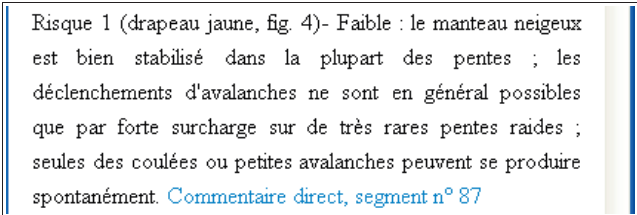


Figure 5. Example of sentence annotated 'Commentaire direct' ('Comment') by a reference³

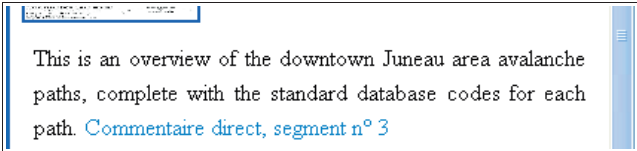


Figure 6. Example of sentence annotated 'Commentaire direct' ('Comment') by markers

It is not enough to annotate text segment as a caption or a direct comment, it is also important to link the commentary with the proper images or non textual object. It is the fourth step. The first situation occurs when the annotated sentence with reference (annotation: 'Legende') is tied to a comment resumed this reference as the example of the Figure 2. The second situation is such that the comment refers to the image with markers; for example, a sentence specifying the image location such as 'above', 'below'... In this case, the comment is related to the image with the significance of the marker. We also have markers that do not indicate the position of the non textual data. In this case, we propose a hypothesis. For example a sentence annotated 'Commentaire direct' by markers as 'this map details' (see Figure 12) is linked to the previous images.

Multiple points of view

Identifying the images and their captions as well as the associated text segment is sufficient for text mining guided by the images as we have just shown. Indeed, we can display

- images, captions and comments to a text;
- the result of a multi-text search by keyword as in the captions or comments;
- the captions and the comments for a selected image in the database of images.

Nevertheless, it is possible to complete this research by offering a image and text mining based on the annotation of several points of view. Thus, we propose now a more advanced search. We can see, besides the image, caption and associated comment, text segments annotated by other points of view. These new segments that could be called 'indirect comment' but which are annotated by the annotation or their points of view, have obviously semantic links with caption or direct comments.

³ Risk 1 (yellow flag, fig. 4) - Low: Snow is generally very stable; Avalanches are unlikely except when heavy loads are applied on a very few extreme steep slopes. Any spontaneous avalanches will be minor (sluffs). In general, safe conditions. Direct comment, segment No 87

A complete example of treatment

The following example can illustrate the way TNT works and image and text mining with multiple points of view. This example is based on the text named 'Juneau urban avalanche maps' taken from the Southeast Alaska Avalanche Center website.

As we explained in the description of the process, treatment starts with a cleaning up of the HTML file (Figure 7) to obtain a TXT file keeping the memory of images as shown in line 5 and at the end of the line 6 in Figure 8.

```

34 content="" csheight="48" xpos="176" height="80"
valign="top" width="464"><font size="6"><b>Juneau
Urban Avalanche Maps</b></font></td>
35 <td height="80" width="1"><spacer
type="block" height="80" width="1"></td>
36 <tr height="540">
37 <td rowspan="8" height="2835"
width="16"><spacer type="block" height="2835"
width="16"></td>
38 <td colspan="3" xpos="16" align=
"left" height="540" valign="top" width="734"><img
src=
"alaska_en_fichiers/DowntownJuneauMapBIGWEB.jpg"
39 alt="" border="0" height="540" width="734"></td>
40 <td height="540" width="1">
<spacer type="block" height="540" width="1"></td>
41 </tr>
42 <tr height="68">
43 <td class="StandardTextBox"
content="" csheight="48" colspan="3" xpos="16"
height="68" valign="top" width="734">This
is an overview of the downtown Juneau area
avalanche paths, complete

```

Figure 7. Extract of HTML code of the original text

```

4 Juneau Urban Avalanche Map
5 &lt;img
src="alaska_en_fichiers/DowntownJuneauMapBIGWEB.j
pg"&gt;
6 This is an overview of the downtown Juneau area
avalanche paths, complete with the standard
database codes for each path. We also have a
larger scale version (542 KB), twice this size.
&lt;img src
="alaska_en_fichiers/SAACBehrendsAvenuePathMapWEB
.jpg"&gt;
7 This map details the runout zone of the Behrends
Avenue avalanche path. The area marked A is the
severe hazard zone. The area marked B is the
special engineering zone. Specially engineering

```

Figure 8. Extract of text after conversion HTML to TXT

The step of segmentation provides an XML file based on the following DTD:

```

<!ELEMENT article (titrePrincipal,
articleInfo, section)>
<!ELEMENT titrePrincipal (#PCDATA)>
<!ELEMENT articleInfo (developpeur, auteur,
source, motsCles)>
<!ELEMENT developpeur (#PCDATA)>
<!ELEMENT auteur (#PCDATA)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT motsCles (motCle)+>

```

```

<!ELEMENT motCle (#PCDATA)>
<!ELEMENT section (titre, (paragraphe)+)>
<!ELEMENT titre (#PCDATA)>
<!ELEMENT paragraphe (phrase)+>
<!ELEMENT phrase (#PCDATA)>
<!ATTLIST section id CDATA #REQUIRED>
<!ATTLIST paragraphe id CDATA #REQUIRED>
<!ATTLIST phrase id CDATA #REQUIRED>

```

It is divided into two parts:

- one part contains all meta data (tag titrePrincipal and articleInfo)
- the other part (see an extract in Figure 9) is the segmented text where each tag has an id attribute to identify it: section, paragraph (tag 'paragraphe') or sentence (tag 'phrase').

```

16 <paragraphe id="1">
17 <phrase id="1">Juneau Urban Avalanche Map</
phrase>
18 <phrase id="2">&lt;img
src="alaska_en_fichiers/DowntownJuneauMapBIGWEB.j
pg"&gt;</phrase>
19 <phrase id="3">This is an overview of the
downtown Juneau area avalanche paths, complete
with the standard database codes for each path.</
phrase>
20 <phrase id="4">We also have a larger scale
version (542 KB), twice this size.</phrase>
21 <phrase id="5">&lt;img
src="alaska_en_fichiers/SAACBehrendsAvenuePathMap
WEB.jpg"&gt;</phrase>
22 <phrase id="6">This map details the runout
zone of the Behrends Avenue avalanche path.</
phrase>
23 <phrase id="7">The area marked A is the
severe hazard zone.</phrase>
24 <phrase id="8">The area marked B is the
special engineering zone.</phrase>
25 <phrase id="9">Specially engineering
residential structures for avalanche impact is a

```

Figure 9. Extract of text after segmentation

For this example, we have annotated the text from only two points of view, obviously the point of view of text-images and the point of view of location relations of which we have given the semantic map above. The result is a XML file whose DTD is identical to the DTD for the segmented file except for paragraph (tag 'paragraphe') and sentence (tag 'phrase') elements that are defined as follows:

```

<!ELEMENT paragraphe (phrase)+>
<!ELEMENT phrase (annotation)*>
<!ELEMENT annotation (avantIndicateur,
Indicateur, apresIndicateur)>
<!ELEMENT avantIndicateur (avantIndice,
Indice, apresIndice)*>
<!ELEMENT Indicateur (#PCDATA)>
<!ELEMENT apresIndicateur (avantIndice,
Indice, apresIndice)*>
<!ELEMENT avantIndice (#PCDATA)>
<!ELEMENT Indice (#PCDATA)>
<!ELEMENT apresIndice (#PCDATA)>
<!ATTLIST paragraphe
id CDATA #REQUIRED
statut (annotate) #IMPLIED
IdAnnotation CDATA #IMPLIED >
<!ATTLIST phrase

```



```

id CDATA #REQUIRED
statut (annotate) #IMPLIED
texteOrigine CDATA #IMPLIED >
<!--ATTLIST annotation
  IdAnnotation CDATA #REQUIRED
  title CDATA #REQUIRED
  idRegle CDATA #REQUIRED
  typeRegle CDATA #REQUIRED
  prioriteRegle CDATA #REQUIRED
  porteeRegle (segment/paragraphe)
#REQUIRED
  placeIndicateur CDATA #REQUIRED
  longueurIndicateur CDATA #REQUIRED
  listeIndicateur CDATA #REQUIRED
  placeIndiceAvant CDATA #IMPLIED
  longueurIndiceAvant CDATA #IMPLIED
  listeIndiceAvant CDATA #IMPLIED
  placeIndiceApres CDATA #IMPLIED
  longueurIndiceApres CDATA #IMPLIED
  listeIndiceApres CDATA #IMPLIED >

```

In Figure 10, we can see a sentence annotated 'commentaire direct' (direct comment) by point of view text-image (TNT).

```

43 </phrase>
44 <phrase id="6" statut="annotate" texteOrigine=
  "This map details the runout zone of the Behrends
  Avenue avalanche path.">
45 <annotation IdAnnotation="4" title=
  "tnt_en.commentaire_direct" idRegle=
  "tnt_en.commentaire_direct-3" typeRegle="1"
  prioriteRegle="1" porteeAnnotation="segment"
  placeIndicateur="6" longueurIndicateur="3"
  listeIndicateurs="elementmontextuel_en.xml"
  placeIndiceApres="10" longueurIndiceApres="7">
46 <avantIndicateur>This</avantIndicateur>
47 <indicateur>map</indicateur>
48 <apresIndicateur>
49 <avantIndice />
50 <indice>details</indice>
51 <apresIndice>the runout zone of the
  Behrends Avenue avalanche path.</apresIndice>
52 </apresIndicateur>
53 </annotation>
54 </phrase>
55 <phrase id="7" statut="annotate" texteOrigine=

```

Figure 10. Extract of text after annotation

After annotation, we proceed to build links between images and associated annotated segments as explained above to obtain a new XML file (Figure 11). The DTD of this new file is:

```

<!--ELEMENT article (titrePrincipal,
  articleInfo, original, (segment_image)+)>
<!--ELEMENT titrePrincipal (#PCDATA)>
<!--ELEMENT articleInfo (developpeur, auteur,
  source, motsCles)>
<!--ELEMENT developpeur (#PCDATA)>
<!--ELEMENT auteur (#PCDATA)>
<!--ELEMENT source (#PCDATA)>
<!--ELEMENT motsCles (motCle)+>
<!--ELEMENT motCle (#PCDATA)>
<!--ELEMENT original (dossier_original,
  fichier_original)>
<!--ELEMENT dossier_original (#PCDATA)>
<!--ELEMENT fichier_original (#PCDATA)>
<!--ELEMENT segment_image (image, phrase,
  (commentaire_direct,
  commentaire_annotation)*>

```

```

<!--ELEMENT image (#PCDATA)>
<!--ELEMENT commentaire_direct (phrase)>
<!--ELEMENT commentaire_annotation (phrase)>
<!--ELEMENT phrase (#PCDATA)>
<!--ATTLIST segment_image
  idsegment CDATA #REQUIRED
  idpara CDATA #REQUIRED
  balisesegment (phrase/paragraphe)
  indicateur CDATA #REQUIRED
  indiceavant CDATA #IMPLIED
  indiceapres CDATA #IMPLIED >
<!--ATTLIST commentaire_direct
  idsegment CDATA #REQUIRED
  idpara CDATA #REQUIRED
  balisesegment (phrase/paragraphe)
  indicateur CDATA #REQUIRED
  indiceavant CDATA #IMPLIED
  indiceapres CDATA #IMPLIED >
<!--ATTLIST commentaire_direct
  idsegment CDATA #REQUIRED
  idpara CDATA #REQUIRED
  balisesegment (phrase/paragraphe)
  indicateur CDATA #REQUIRED
  indiceavant CDATA #IMPLIED
  indiceapres CDATA #IMPLIED
  title CDATA #REQUIRED >

```

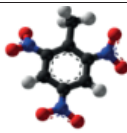
```

24 </segment_image>
25 <segment_image idsegment="4" idpara="1"
  balisesegment="phrase" indicateur="img"
  indiceavant="" indiceapres="">
26 <image>
  alaska_en_fichiers/SAACBehrendsAvenuePathMapWEB.jpg</image>
27 <phrase>img
  src="alaska_en_fichiers/SAACBehrendsAvenuePathMap
  WEB.jpg"&gt;</phrase>
28 <commentaire_direct idsegment="6" idpara="1"
  balisesegment="phrase" indicateur="map"
  indiceavant="" indiceapres="details">
29 <phrase>This map details the runout zone
  of the Behrends Avenue avalanche path.</phrase>
30 </commentaire_direct>
31 <commentaire_annotation idsegment="7" idpara=
  "1" balisesegment="phrase" indicateur="area"
  indiceavant="this" indiceapres="marked A" title=
  "relations_reperage.identification">
32 <phrase>The area marked A is the severe
  hazard zone.</phrase>
33 </commentaire_annotation>
34 <commentaire_annotation idsegment="8" idpara=
  "1" balisesegment="phrase" indicateur="area"
  indiceavant="this" indiceapres="marked B" title=
  "relations_reperage.identification">
35 <phrase>The area marked B is the special
  engineering zone.</phrase>
36 </commentaire_annotation>
37 </segment_image>
38 <segment_image idsegment="20" idpara="4"

```

Figure 11. Extract of result after post-processing

This result is formatted with a CSS stylesheet and proposed in a web interface (Figure 12) that allows to search by keyword (par mot-clé) or directly by image (par image).



TNT v.2 : Textes - données
Non Textuelles

Contact

Traitement

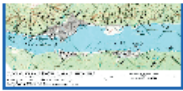
Fouille de Donn

par texte (fr)

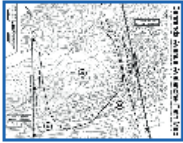
par texte (en)

par image

par mot-clé



This is an overview of the downtown Juneau area avalanche paths, complete with the standard database codes for each path. [Commentaire direct, segment n° 3](#)



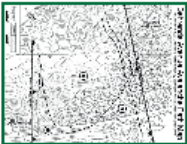
This map details the runout zone of the Behrends Avenue avalanche path. [Commentaire direct, segment n° 6](#)

The area marked A is the severe hazard zone. [relations_reperage.identification, segment n° 7](#)

The area marked B is the special engineering zone. [relations_reperage.identification, segment n° 8](#)

Figure 12. The result in the web interface

The picture below shows the result of a search by image.



This map details the runout zone of the Behrends Avenue avalanche path. [Commentaire direct, segment n° 7](#)

The area marked A is the severe hazard zone. [relations_reperage.identification, segment n° 8](#)

The area marked B is the special engineering zone. [relations_reperage.identification, segment n° 9](#)

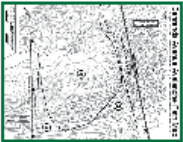
[texte original : alaska_en.html](#)

[texte annoté : alaska_en.html.txt.xml](#)

Figure 13. Search by image

By clicking on the picture capitalized in the database, the image is displayed as well as all associated annotated segments. In this example, no caption is attached to the image but a direct comment ('Commentaire direct') and two annotations of relation of identification.

Figure 14 displays a search by the keyword 'zone'. All pictures, capitalized in the database, which contains a associated segment with the keyword are displayed regardless of the original text. In our example, we have several pictures including one from the text about avalanche in Juneau and one from a thesis in computational linguistics. This search will bring up the polysemy.

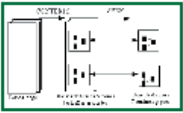


[texte original : alaska_en.html](#)

[texte annoté : alaska_en.html.txt.xml](#)

The area marked A is the severe hazard **zone**. [relations_reperage.identification, segment n° 8](#)

The area marked B is the special engineering **zone**. [relations_reperage.identification, segment n° 9](#)



[texte original : Chapitre_1.html](#)

[texte annoté : Chapitre_1.html.txt.xml](#)

Il s'agit alors d'utiliser SEEK(version1) sur ces **zones** textuelles afin de déterminer les relations sémantiques que les termes de chaque classe entretiennent entre eux (Figure 7). [Commentaire direct, segment n° 192](#)

Figure 14. Search by the keyword 'zone'⁴

References

- Alrahabi M. 2010. *EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe*. PhD thesis. Paris-Sorbonne University.
- Desclés J-P. 1987. *Réseaux sémantiques: la nature logique et linguistique des relateurs*, in Langages, Sémantiques et Intelligence Artificielle, 55-78
- Desclés J-P. 1997. *Systèmes d'exploration contextuelle in Co-texte et calcul du sens* (C. Guimier), PUC. 215-232.
- Djioua B. & al. 2006. *EXCOM: an automatic annotation engine for semantic information*. FLAIRS-19, Melbourne, Florida, 11-13 may. 285-290.
- Le Priol F. 2008. *Automatic annotation of images, pictures or videos comments for text mining guided by non textual data*. FLAIRS-21, Miami, Florida, 15-17 may. 494-499.
- Le Priol F. & al. 2006. *Automatic annotation of localization and identification relations in platform EXCOM*, FLAIRS-19, Melbourne, Florida, 11-13 may. 307-312.

⁴ It involves using SEEK (version 1) on these text fields to determine the semantic relationships that the terms of each class with one another (Figure 7). Direct comment, segment No192