AnT&CoW, a tool supporting collective interpretation of documents through annotation and indexation

Gaëlle Lortal[1], Myriam Lewkowicz[1], Amalia Todirascu-Courtier[2]
[1]Université de technologie de Troyes
ISTIT Laboratory, Tech-CICO
12, rue Marie Curie BP 2060 10010 Troyes Cedex
{lortal, lewkowicz}@utt.fr
[2]Université Marc Bloch de Strasbourg
22, rue René Descartes 67084 Strasbourg
amalia.todirascu@umb.u-strasbg.fr

Abstract

This paper describes an Annotation Tool supporting Collaborative Work (AnT&CoW) and particularly collective interpretation of documents using annotation. In the first part, we present our methodology to design such a groupware based on a theoretical activity analysis, understanding discourse production activity as a complex writing/reading activity. Following a rhetorical discourse production theory (section 4), we model a discourse production activity and its mediatization by way of a tool (section 5). After existing annotations standards and tools have been detailed (section 6), we present our tool's requirements (section 7). AnT&CoW is following Annotea W3C standards and allows document annotating and then multi-dimensional indexing. Multi-dimensional indexing is based on a semiotic ontology represented in Topic Maps where three dimensions occur: argument, role and domain. Dimensions, mainly the domain specific dimension, are based on Natural Language Processing (NLP) techniques fitting the text up. In the last part, we present our Web-based application, its client/servers architecture and its visualization's features. Our prospects are then proposed.

1 Introduction

Nowadays, documents are a central point of interest in our organizations as many works in research show. For instance, in France, a multidisciplinary network from CNRS (National Centre for Scientific Research) works on "documents and contents: creation, indexation, navigation" (RTP-DOC). Three orientations around documents are then told apart; analyzing documents as a shape (studying structure of documents for its manipulation), as a sign (studying author's intentions when creating documents, document's intentionality), and as a medium (studying document's status in social relations) [Pédauque, R.T., 2003]. Following the second orientation, this paper intends to consider a document as a meaning-holder, that cannot be dissociated from a subject who is building or re-building it and who gives sense to it. Seeing document as a sign means that we are more interested in the creation process of a document, its interpretation, in other words in signs constituting it.

These questions are tackled here from the critical reading point of view, contrary to a reading which would not aim at producing knowledge or another text. A critical reading creates an interpretation enlightening not only the text that is read but also other texts. It can produce another text, a comment, a review, a criticism. We focus particularly on collective critical reading, which allows the building of a shared interpretation of an initial document between several participants. The drawing-up of a shared interpretation within a group takes part, according to us, of a collective sense making process [Weick, 1979]. Actually, Weick defines collective sense making in organizations as a process of collective reduction of the perceived ambiguity of a situation. By exchanging, discussing ideas, members of an organization will clarify and then share their understanding of a situation (transcribed in documents), gradually making sense.

Collective sense making in organizations from real lived situations is a theme that has been studied since the beginning of 80's. Weick's work emphasizes the sense making process, its creation and its evolution, and not the collective representation of sense. The collective sense is then not necessarily a common sense.

According to us, the collective interpretation of documents, which are the marks of the actions in the organization, will allow collective sense making. This cooperative interpretation process thus permits to take advantage of documents while letting able to overstep the setting in which the documents have been created. This process is also supporting individual identity since each participant puts his identity to the critical test, making it evolve through his/her interactions.

We propose to support this collective interpretation of documents by developing strategies for mediatized interac-

tions around numerical documents, mostly textual. Texts' interpretation is traditionally accompanied by gloss, note, commentary, and various kinds of annotations anchored to the text itself or linking several texts or fragments of text.
We then propose to support this discursive collaboration around documents by a system allowing documents' annotation for interpretation and appropriation, objective which is not yet supported by existing annotation-based software. Actually, these tools only allow isolated annotation as textual comments, with weak indexation (date, author), hardly usable as interactions' support in a group. In fact, in a situation where we want to support a methodical texts' interpretation, textual body of comments is promoted to discourse, its context is built up by the role of the author, the semantic content, the place of the annotation into the discussion's thread. Giving this context is essential to find the design rationale of an interpretation.
Studies have been conducted at the KMI (Knowledge Media Institute) on functions of discursive comments of a document. They gave rise to the "Digital Document Discourse Environment" (D3E) [Sumner *et al.*, 2000], a web tool in which exchanging messages on a document are allowed. But, as the design of this tool has not been bound to any study of the activity of document analysis, the collaborative process of interpretation is not treated .Moreover, nothing has really been done on visualization and reuse of the exchanged messages. Actually, messages are tree-displayed and indexed according to standard attributes (date, author, title); it is as though a forum has been linked to a document. In fact, many works outline yet that online discussions are often disrupted and confused because of the numerous and frequent development of discussion threads and parallel talks. We can for example quote [Marcoccia, 2004] who stresses the phenomenon of progressive themes' digression in newsgroups, when each message in a thread introduces a theme development. The result could be a real "topic decay" [Herring, 1999].
In this paper, we first present methodological principles to design a groupware supporting activecollective interpretation of documents (AnT&CoW). Then, we focus on existing works in modeling writing activities. In section 4, we present a model of discourse production stemming from rhetoric, which is adapted in section 5 to a mediatized activity. This model is the basis of AnT&CoW, which features are described in section 7, after a review of existing tools and standards for annotation in section 6. We finally present the tool architecture combining Natural Language Processing (NLP) techniques for text material processing

2 Methodological principles to design a groupware supporting collective interpretation of documents

The context of our research leads us to define new practices to support collective interpretation of digital document. Then, a classical software design process, deducing design principles from a needs analysis or an existing activity analysis, is not suitable.

The design process which we are presenting here draws its inspiration from the methodological positioning in the field of design in Educational Research by [Baker, 2000], carried on, in France, by Tchounikine [Tchounikine, 2002]. These authors distinguish models as scientific tools from models to design systems. The firsts propose a theory to understand or predict a situation or an activity; the seconds translate the firsts in models allowing design and implementation of systems supporting the situation or the activity.
However, theories from humanities usually mobilized to design groupware (activity theory, learning theory, communicative action theory…) are very difficult to use as they are. In fact, it's difficult to deduce principles of design or to adapt the definitions of these theories in a computer-mediatized framework.
Designing consists then in defining new models, with new concepts, in keeping with the theory, in order to describe an artefact supporting and marking interactions. The theory will then help us to analyze these recorded interactions.
We thus propose the following process, illustrated in Fig. 1: From a social science theory fitting to phenomena which one wishes to support/observe, use or define a *descriptive model* of these phenomena which makes the theory operational. This descriptive model allows reasoning about situations in which these phenomena would be mediatized using an information processing system. This reasoning leads to the creation of a *mediatized activity model*. This step involves researchers in humanities and social sciences responsible of the link with the description model, and computer science researchers (designers), understanding and controlling software properties. This mediatized activity model is then materialized in a *design model* describing requirements for a groupware enabling to assist interactions and also to mark them. This groupware will thus be a mean to collect corpus. This corpus, analyzed using the mobilized theory, will allow us to make evolve our comprehension of the phenomena being studied.

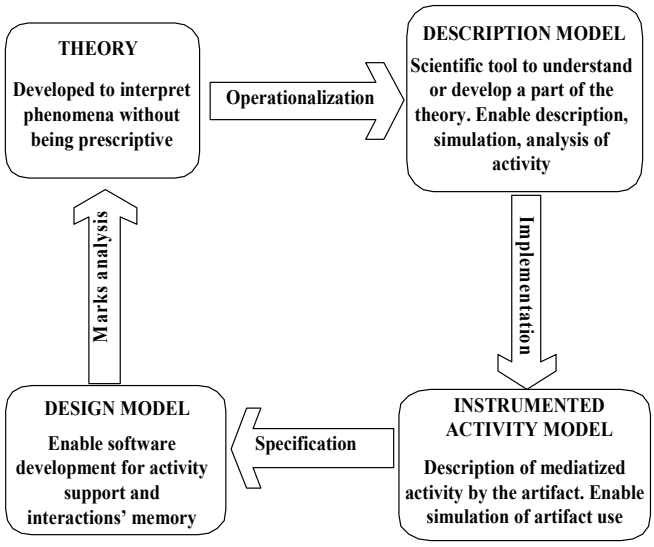

Fig. 1 – Groupware design based on a theoretical activity analysis

It seems to us that although the step of designing mediatized activity is always present while designing software, the activities of this phase are not usually explicit. It occurs as if it was possible to define design principles of an artifact supporting an activity, directly from the descriptive model of the face-to face-activity. However no one would deny that this mediatization has an impact on the activity. During this step of designing the mediatized activity the exchanges between researchers in humanities and social sciences and researchers in information and communication technologies will take place. They will then be able to build a common model reflecting the guidelines of the activity and the ways to assist this activity at the same time. This step allows the next step of design to take place. A design model will then be defined, describing the functions of the tool.
In the following section, in order to define a description model which fits our problematics of collective interpretation of documents, we present existing work on analysis of documents centered activities

3 Which theory to analyze discourse production activity?

In the field of cognitive psychology, many researchers have studied mental activities involved in writing, distinguishing text *comprehension* and text *production*.
With regard to *comprehension* models, the researches focus on memorization of text fragments, necessarily summarized. One of the most quoted model in this field is the Kintsch's constraint-satisfaction process [Kintsch, 1988]; The comprehension of text is described there as a cycle of phases of construction of a coherent mental representation of a text in the course of reading, and of phases of selection (or not) of text fragments for memorization (integration). Researches were undertaken to use this descriptive theory at constructive ends, for example for defining design principles for hypermedia documents to be easily integrated by the reader [Garlatti and Iksal, 2000]. These authors propose a guide for "good practices" in designing documents, particularly to ensure text coherence. These documents are then presented so that the reader receives help in constructing his mental model. The aim is to minimize the cognitive cost while reading the document.
Concerning *production* models, the stress is laid on editorial processes of planning, formatting and reviewing, and the control model which allows to apply these processes. The authors frequently quoted in this field are [Hayes and Flower, 1980] who proposed models of editorial strategies. There again, this descriptive theory was used in works which gave rise to computer-supported editorial processes. In [Piolat *et al.*, 1989] a combination of three pieces of software (scripsis, scripap, scriprev) is used. Each one focuses on a process (planning, formatting, reviewing). However this work doesn't aim at proposing tools for text production within an organization, but at providing a framework for experimental study of text production.
As we presented in section 2, our approach consists in designing a groupware on the basis of the theoretical analysis of the collective activity this groupware intends to support. The descriptive models of comprehension or production offered by cognitive psychology, which we quoted above, do not appear suitable according to us for the design of a tool supporting collective interpretation of documents. In fact, they separate the memorization phase from the text formatting phase. Indeed, collective interpretation of documents mixes written activities during reading - annotations – and reading activity to produce meaning, sense. The reading/memorization phases and writing/integration are thus associated. In researches related to written didactics, reading and writing are also seen as stages of a generic activity related to the written support [Barré de Miniac, 2000]. We thus propose to use a discourse production model stemming from ancient and medieval rhetoric didactics, representing in a whole cycle both memorization and discursive production.

4 Discourse production model

Writing is the place of complex and evolutionary interactions between emotional, cognitive and linguistic factors [Barré de Miniac, 2000]. We will be interested more particularly in the cognitive factors as organizing factors of the concepts in memory and text, and in the linguistic factors as marks at the same time of a specific type of discourse and of the semantics of the document in "co-text". As an author's discourse is surrounded by social life and events, a text is surrounded by textual context making its sense.
We find these two types of factors in the rhetoric didactics. From Aristotle rhetorical theories to Hugues de St Victor's ones through Cicero or Quintilianus, discourse production is taught according to a process. Aristotelian rhetoric is focused on a final production of oral discourse (speech) without denying a memorizing phase required for any production. This phase of memorizing is better represented by rhetoric, that we will call memorial, held by thinkers quoted by [Carruthers, 1990], such as Quintilianus (the institution oratory), Cicero (De oratore, De inventione) or Tullius (Ad Herennium), and then Hugues de St Victor (Didascalicon), Fortunatianus (Artis rhetoricae libri tres) or Julius Victor (Ars rhetorica) from Middle Ages. In this approach of rhetoric, we can observe a continuum between the memorial part more "logical" or "dialectical" and the stylistic, editorial part. Rhetoric is regarded as an alliance between structuring and eloquence.
The discourse production process as recommended in this didactical context is made up of two phases: "Divisio" and "Compositio". Divisio is done while reading and consist in dividing a text in understandable units, in memorizable short segments. Compositio is the ordered combination, the suitable arrangement of "res" (conceptual or material objects) contained in the memorized segments (Fig. 2). These memorizing, Divisio, and creation phases, Compositio, are themselves divided into stages supported by the use of annotations.
The first stage of Divisio is Cogitatio. It is an individual memorial stage which consists in associating, by a conscious choice and recall, images and sections of a chrono-

logically divided content of a document in various memorial places. Textual fragments that form the text are then structured and become easily memorizable.
Collatio is the phase where textual fragments stored in several distinct places in memory are combined in a structure. In this phase connections between the various places of contents are created. A co-text is then formed by semantically binding new memorized fragments and fragments previously memorized. This phase is not specifically individual even if it structures an individual memory, insofar as this stage can be related to discursive exchanges, interactions with others.
Compositio is divided into four stages of activity evoking stages of document creation. The stage of Inventio is close to that of Collatio insofar as it is question of creating semantic links between various memorized elements, on the "res" (conceptual objects, idea) level not on the word level. An outline is formed, i.e. a set of ideas hierarchically arranged, an argumentative structure for example.

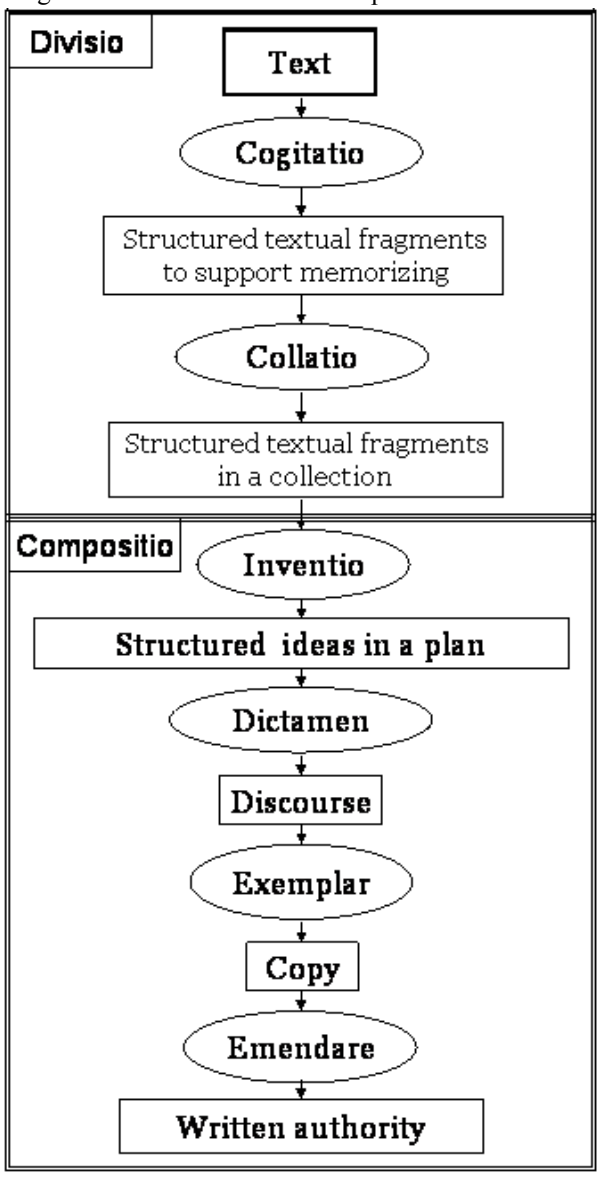

Fig. 2 – Discourse production model

The following phase will be the formatting in word of this conceptual outline. It is a traditional phase of drafting, called "Dictamen". We see with this stage the physical discourse creation, classically done on an adjustable support (a draft), where the style, the choice of the terms, therefore the textual shape of the discourse only can be modified.
The Exemplar phase consists in transforming the draft support of the discourse in a perennial support. The discourse remains strictly identical to the one found in output of the process of Dictamen.
The last phase but not the least in this succession of process is the Emendare where the final copy of the discourse is diffused and then openly commented by the addition of public comments, "notae" or arguments of an author to the original text. This phase thus makes the text become a reference text, a written document being an authority on the field.
This model represents a method of discourse production strongly supported by memory. In a computer supported collaborative work (CSCW) context, discursive creation must be supported by an adequate tool enabling storing, creating and sharing information. In order to design this tool, we first wish to model this mediatized activity of discourse production to represent the functions required for implementing in a tool.

5 Model of mediatized discourse production

We are interested here in collaborative interpreting numerical document, sense making by several participants. We will not take into account non-textual numerical documents.
The transformation of the discourse production model within a mediatized framework, enables us to define the following stages to recommend (Fig. 3). First, the text of the document is segmented to be stored in a memory as memorizable fragments.
These segments are then indexed to avoid the loss of the document structure as consistent unity. It is important to chronologically index the segments to mark the hierarchy of the various paragraphs in a text document, various words in a paragraph... This type of indexing concerns all metadata which might be automatically associated with element stored (localization, author, date...). Indexation must also be used to bind new fragments laid into the system to the conceptual set already present in the tool. We will then obtain a set of textual segments semantically bound to other textual segments. It is a process of co-textual structure creation organized by socio-cognitive as well as semantic links.
The structuring phase represents a hierarchizing process, organizing ideas according to a chronological outline. A detailed outline is defined, containing all ideas necessary to the formatting phase, the change of concepts, to words, to discourse. It is the phase where the "res" (concepts) contained in indexed textual fragments are re-used and reorganized in a new document.
The writing phase is the one where the outline is formatted in text giving a discourse as a result. This discourse is not the final objective of this activity in this vision of rhetoric, since it is then published to become an amendable object, a

writing improved by reader's feedback, themselves becoming authors in the community.
This phase when the published discourse is assessed by other members of the community is extremely important as it is allowing the validation of the Exemplar, its improvement even, and constituting a written authority, a reference discourse in the community.
Within a collective interpretation purpose, annotating a document thus consists, according to us, in following a process of formatting organized ideas in a discourse. Indeed, following the reading of a document, it comes to engage a process which enables to add an idea or an opinion structured in textual form.

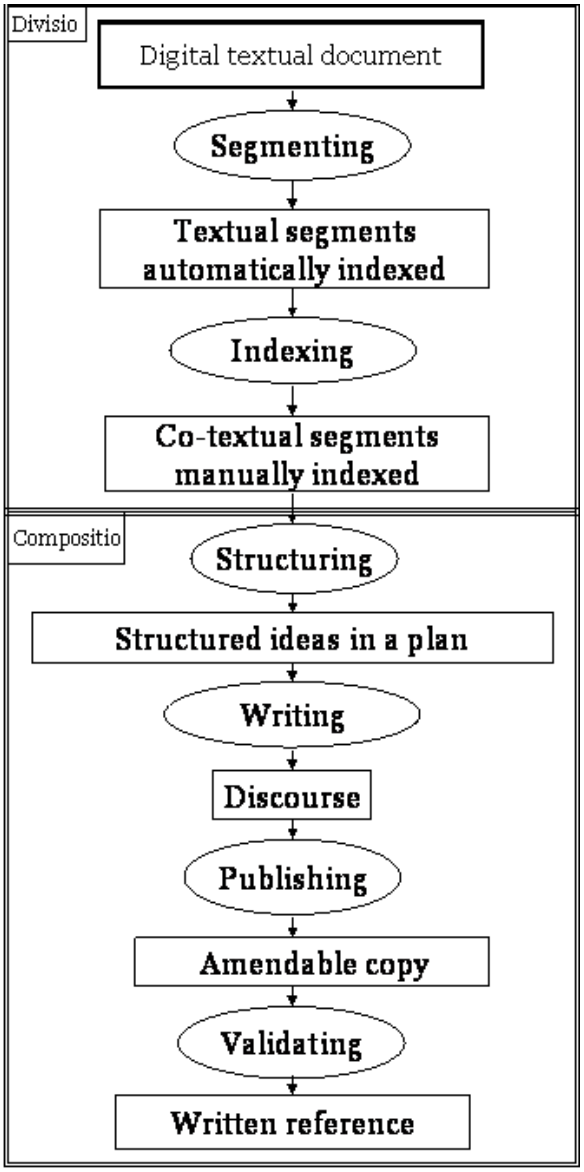

Fig. 3 – Model of mediatized discourse production

For example, in a collaborative work context, one can consider the sharing of a document in order to be commented on. After a visualization phase of the text, a reading, the text read will be segmented to allow the addition of a structured comment, of a discursive annotation. A segment will be emphasized in order to indicate the anchoring of a discursive element linked to this segment. This highlighting could be done by traditional techniques of underlining, circling, colouring segments of unsettled sizes (from a word, or a part of a word, to the paragraph, or set of separated elements). Following the segmentation and the choice of element to be annotated, an indexing phase is required, consisting in connecting segments. The tool should help the user to find semantic links between elements to structure them together and to form an organized set of textual segments according to their meaning. This meaning depends on the user's understanding. Indeed, the annotation consists in an anchor, a geographical relation, in a body, a discourse which creates its meaning amid a "co-text", but also in the whole set of textual segments stored in memory and linked to it, indexed to it by comprehensible key words, structured by and for human user. While writing this annotation, the author should organize his/her discourse to be written. This necessary step is the structuring of "rei", of concepts stored in memory, which will give rise to an outline made up of hierarchically structured arguments. The writing phase will allow constituting the body of the annotation which will be readable by a member of the discussion after publishing and thus spreading this annotation.
Just as a reference text, the annotation can be endorsed thanks to a new link brought to the latter. A reply to a comment allows taking part in the thread of discussion initiated by the first annotation.
This model of mediatized discourse production, resulting from a model of discourse production activity stemming from rhetoric, enables us to describe the requirements of a groupware assisting this type of discursive production by means of annotation.

6 Designing AnT&CoW

6.1 Existing annotations standards

As recommended through the model presented in section 5, the groupware must let users visualize a document, segment it, create various types of associations (indexing, gathering) with the various fragments, write the discourse constituting the annotation body, or publish it. The validation phase (cf. Fig. 3), optimizing collaboration through answers during the discourse, requires a specific association function as a "reply to" function to an annotation. The discursive model allows a continuous look back on the document when reading and writing, so the visualization function is predominant. The visualization is supported by the use of a plug-in into a navigator. Indeed, being an extension of a naturally used navigator and giving access to a lot of Web documents to be read, this plug-in enables visualizing simultaneously the document and the body of the annotation while writing or indexing the annotation. This annotation is captured by a "pop-up" window, then indexed to entitle its recovery after publication and creation of a set of structured documents.

In an annotation activity, several problems arise: the question of anchoring the annotation and the forms of its meta-information in the original document. These problems are tackled in the field of Semantic Web (SW), which goal is to enrich Web resources with structured descriptive information to improve their accessibility, their retrieval and the use of information. We now will describe some existing tools from this field which we can re-use and enrich in our project.

The SW identifies three types of annotations: simple metadata (modification date, author, etc.) ; annotations which we would describe as "computational" insofar as they are addressed to programs enabling them to take a profit from annotated resources [Bremer and Gertz, 2001], [Volz *et al.*, 2003], [Roussey *et al.*, 2001]; and annotations which we would describe as "social" since they are addressed to the reader, to an human user, enabling her/him to be an active Web participant.

Tools developed since the beginning of the 90's allow reviewing texts using comments or explanations, to justify decisions... In general, they consist of various elements permitting to visualize, to create, to store and to search the annotations. Annotations are defined by an anchor, some attributes and a body. They are stored on a dedicated server (annotations server), and can be classified according to their attributes, their public/private/group shared status. The annotations server contains information about the annotation localization (the document on which the annotation was created or its place in the document), its style (font, color...), its contents (text and attributes), and its function (if it is an explanation or a proposition for example). The annotations are generally tree organized. This configuration facilitates navigation in the set of annotations and their management.

These researches lead to the definition of the W3C's Annotea standard [Annotea, 2003] [Kahan *et al.*, 2001], based on a RDF annotation description [Brickley and Guha, 2004]. This standard improves collaboration through shared metadata based on Web annotations, bookmarks, and their combinations. Several annotations servers (ZAnnot, Annotea...) and annotations clients (Annozilla, Amaya...) implement the Annotea standard. The annotations server ZAnnot [Zannot, 2003] stores annotations in a RDF database. Users can interact with Zannot server by Annozilla client [Annozilla, 2004], the Mozilla navigator's plug-in, in order to search for an annotation, to create a new one or to remove another. An annotation is described by a set of metadata (its attributes defined by a RDF diagram) and a body. The RDF notation's advantage is that it is possible to personalize it, for example by adding to the annotation diagram, attributes or a set of values of attributes. This technical solution is thus interesting since it is possible to adapt the model to a need of multidimensional indexing. These dimensions supplement Annotea already existing attributes and are related to a "socio-semantic" use of the annotations in our project.

We are now going to describe and classify these existing annotations tools, and we will clarify our positioning.

6.2 Existing annotations tools

At present, several annotations clients are available, stemming from SW initiatives. Most of them adopt what we would call a "computationally-semantic" approach. This approach has, as main objective, to index Web pages more or less automatically. These tools are used for metadata creation and some are based on ontologies to support the computational annotation: OntoMat-annotizer [Handschuh *et al.*, 2002]; Melita [Dingli, 2003]; MnM [Domingue *et al.*, 2002]. Computational annotations are geographically dependent on a part of a Web page, but they only enrich the page with concepts for automatic indexing and do not either contribute towards to co-operate or interact between readers of a same page. In fact metadata index a page, and allows the search engines a better information or pages recall.

Other annotations clients adopt a more social approach, aiming at facilitating human communication, without considering indexing features or annotation recall. In this software, these annotations can only be sorted on rudimentary metadata such as the creation date or the author: Yawas [Denoue, 2000]; CritLink [Ka-Ping, 1998]; XLibris [Price *et al.*, 1998]; etc. These annotations tools regard the annotation as a comment, a way of looking at annotation shared by some proprietary software or some plug-in application software, where the comments are neither indexed nor differentiated from the document [Windows Word comments, 2003]. The annotations are sometimes stored apart on annotations servers [Acrobat pdf, 2004] and organized in a minimalist way. However, these annotations tools do not allow connecting annotations. These tools cannot then represent a structured set of exchanges between users related to a document.

We are considering documents as mediators of discourse as KMI's D3E [Sumner *et al.*, 2000] considers. However, this tool does not allow a rich indexing of annotations, and then it will be difficult to understand the design rationale of the discussion, of a new document or even of a new concept.

Thus, even if these annotations tools support the interaction more easily than the computational annotations tools, they are not sufficient to implement our model.

We finally can classify annotations tools in two families; one concentrates on the Web pages indexing, supporting their recall, while the other concentrates on the human communication through comments. In a collaborative environment design aim, we can deplore the lack of annotations management or co-operative work possibilities in these two tools families. We thus propose to enrich them thanks to the SW indexing techniques and to support user in her/his activity of documentary annotation. Supporting this documentary activity will help her/him working in a collaborative way. Moreover, we propose other annotations functions such as multi-anchoring (allowing connecting several fragments of documents) or the answering possibility to an annotation.

In the following part, we then expose the features of an application supporting cooperation around a document, in a "Socio-Semantic Web" approach.

7 AnT&CoW requirements

Following [Zacklad et al, 2003], we define annotation as a type of located metadata, connected to another document. This unit is connected to various parameters such as time, place, participants, its public or private status, its meaning... which means that annotation is an entity made up of several parts such as its anchor (or its anchors) in a document, its attributes, and a body (the text of the annotation). We also consider that the annotation is a mark of the collaboration process which has two principal functions: planning (project management, micro-organization) and the reviewing (argumentation, annotation constituting a document body...)
Metadata suggested by Web standards (for example Annotea described above) to index annotations (name of the author, date, topic, type of annotation, etc.) are thus not sufficient for our project. In fact, with this type of index, we cannot store the organizational context (roles, profile of the participants, etc), the contextual field (specific lexicon, keywords of the field, concepts, etc), nor the type of argumentation (suggestion, opposition).
In order to allow a more subtle classification of these annotations, we thus propose to extend the collaborative annotation indexing not only by domain specific dimensions (topics), but also by a cognitive dimension thanks to an argumentative dimension (preserving the rationale of the decisions and negotiations between human participants) and an organizational dimension, using the participant's role to stress the importance of a decision.

7.1 Semiotic ontologies for multi-dimensional indexing

The three dimensions defined above are described by an ontology. From a SW point of view, ontologies are supposed to represent exhaustively the knowledge of a specific field, structuring concepts in a hierarchy by relations between them. Each concept is well defined by all its properties and the expert must thus entirely specify the relations between the concepts. However, human experts often have conflicting definitions of some concepts for which several definitions are in competition. Concurrently, specific inference mechanisms calculate the coherence and the consistency of these ontologies. Building such ontologies is a time-consuming and expensive task. Plus, on one hand, generic ontologies (EuroWordNet, DOLCE [Gangemi *et al.*, 2003]) are not adapted to domain-specific applications; they do not contain domain-specific concept definitions. On the other hand, domain-specific ontologies are not available or they are very expensive, even if their portability is increased by the use of W3C standards (OWL, RDF). Thus, it is difficult to work out a representation of the semantic contents of Web pages, even using ontologies.
To avoid this drawback, a more socio-semantics approach of the Web proposes the use of less formal ontologies, which main purpose is to help user navigating through Web pages and not to compute automatically the semantic representation of the document content. From this perspective, the concepts should be less-specified; there is no need to identify all the concepts' properties. Standards as Topic Maps (TM) (Standard ISO, [Biezunski *et al.,* 1999]) are defined for these semi-formal ontologies. TM formalism defines a network of topics covering domain-specific knowledge. Topics are defined via simple URL, so all the users share the same definition. The topics are hierarchically organised (related by "isa" relations) and associated by horizontal relations ("partOf", "used") (Fig. 4). No coherence checking mechanism is done.

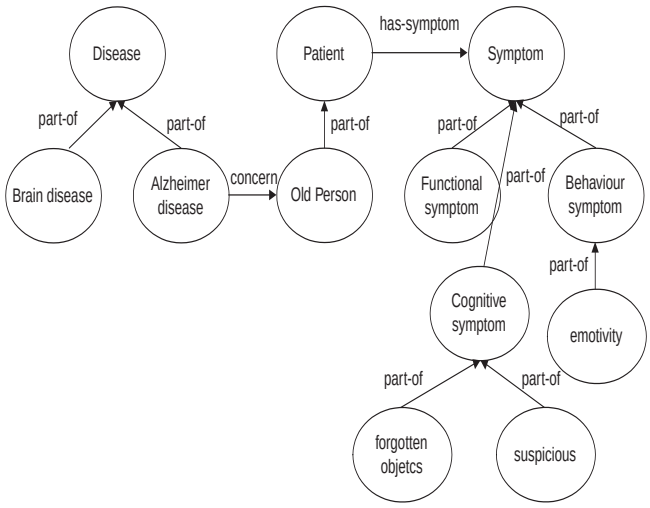

Fig. 4 – Medical domain ontology fragment in Topic Maps

While TM do not require a precise definition of concepts, and are designed to support user browsing Web pages; we adopted this formalism for representing the various dimensions of our ontology.
In our system, the organizational and argumentative dimensions are built manually. The first one is based on a social analysis of the network, and the second one is based both on a cognitive and a pragmatic analysis of interactions in the network. The domain-specific dimension requires a combination of Natural Language Processing (NLP) techniques and manual choice of terms and concepts. This ontology is stored on an ontology server which allows an easy recall of the concepts. We focus now on the NLP techniques.

7.2 NLP tools and methods for domain contextual ontology building

Due to the low availability of domain-specific ontologies and to the fact that generic ontologies are of little use for domain specific applications, many projects aimed to use NLP techniques to extract semi-automatically terms (concept instances) [Jacquemin and Bourigault, 2003] to create term clusters (concepts) [Cimiano and al, 2004] as well as to extract relations between terms [Buitelaar and al, 2004]. The expert should name the clusters as concepts and eventually should define relations between concepts.
In our system, NLP techniques are used for two main purposes: building and maintaining the domain-specific ontol-

ogy from corpora, but also for browsing and indexing annotations.
The annotation indexing can be done automatically by the tool (date, author, answered annotations codification, automatic chronological thread of discussion) or manually by the user. The annotation manual indexing phase by the user regarding to three dimensions (choice of a value representing the annotation content according to each dimension) can be tedious and we thus wish to support it thanks to NLP tools.
The *first* task, concerning ontology building is done off-line, by extracting terms from a selected corpus and by proposing a simple topic hierarchy (a term is equivalent to a topic).
Tests were carried out in the medical field (Alzeihmer's disease and memory troubles), for an Electronic Patient File (EPF) project. An EPF is a patient file created and maintained by a medical group to follow a patient and improve its cares. To be easily followed by distant members, this file is shared by means of Web interface.
It was not possible to use medical ontologies [MeSH, 2004], [UMLS, 2004] insofar as they are too generic or cover a swarms of domains (MENELAS, [Zweigenbaum and al, 1994]) far away from the application's use in the project.
For building a semi-formal ontology (structured in topics) from corpora, we identify candidate terms by using a term extractor. Among the term extractor available, we tested LIKES [Rousselot and al, 1996] which is a simple repeated segment extractor identifying sequences of words (collocation, repeated segments) occurring in the corpus. The repeated segments are potential candidate-terms, and they are organized in a tree, gathered according to their head and displayed according to their frequency of their occurrences. The candidate-terms are used to select the topics of our ontology. The outputs are filtered in order to eliminate the incorrect candidate-terms (terms finishing by a preposition, a conjunction). The majority of the candidate-terms correspond to a Head + Modifier pattern.
We carried out tests on a small medical corpus (14000 words) and obtained an approximately 100 topics ontology. The sizeable drawback of this tool remains the significant number of candidate-terms, which requires a stage of manual cleaning of the resulting hierarchy.
We developed a tool (GenTMInd), identifying hierarchical relations between terms via heuristic rules and structuring them in Topic Maps format. For example, a term matching a pattern Head + Modifier is a subconcept of the Head concept. For the moment, candidate topics should be identified among simple noun phrases (a noun phrase followed by only one prepositional phrase).
These assumptions and heuristic rules are not sufficient to identify all the hierarchical relations or all the relevant candidate-topics. User thus can manually update the ontology by adding relevant topic-keys indexing her/his annotation and by organizing them in the existing TM.
However, after a relevant corpus is gathered, we will extend the search for candidates to a set of domain-specific verbs. We will explore the context of each topic-candidate in order to identify more relations between the topics. If it is possible to find out candidate-topics frequently co-occurring (related by a syntactic relation as predicate-argument or head-modifier) in the text, it would mean that horizontal relations must be added between two candidate-topics. For example, in the context of the disease of Alzheimer, the corpus of test contains "old person", which means that relation "concern" between two topics could be added (Fig.4).
The *second* task is to help the user indexing his annotation regarding to three dimensions (other indexes like author name, date, title, are automatic), by proposing him/her a semi-automatic indexation of his/her annotation (indexes as name of the author, date or title are automatic). NLP tools scan the annotation submitted by the user, identify some relevant terms candidates and match these terms to the concepts of the ontology for each dimension. The matching process uses three resources: the indexation context, the annotation co-text and the ontology. Ontology is a vertical representation of the concepts, i.e. with paradigmatic links, while the indexation context and the annotation co-text are syntagmatic links database. The indexation context is a database storing textual contexts frequently co-occurring with the ontology topics. The annotation co-text is a database storing textual bodies of annotations and textual fragments where these are anchored (fragment of documents). Indeed, to process this mapping, we have several relations databases allowing combining paradigmatic and syntagmatic relations to improve lexical access, data recall. The mapping algorithm checks the contexts of the ontology topics and the contexts of term candidates. If similar context are found [Harris, 1988], the topic is proposed to index the candidates.
The annotation tool will then propose domain specific keywords or "keysyntagms" as well as argumentative types to the user. The user will then decide if the index suggested is relevant and if s/he wishes to preserve it as metadata of her/his annotation.
By creating his/her annotation, the user decides if the annotation is anchored to one or more parts of the document or of several documents. Thus, we consider a complex annotation indexing and multi-anchoring, defining more precisely the co-text of the annotation. Once the validation is done by the user, the annotation is stored with its metadata on the annotations server.
The next step in this tool implementation is to adapt a more effective term extractor in our system,, as FASTR is [Jacquemin and Tzoukermann, 1999], in order to identify the candidate-terms in the annotations bodies and to extract a concept hierarchy by the clustering techniques [Cimiano and al, 2004].
We will now present our distributed architecture and some visualization features of our annotations tool, following W3C standards and integrating NLP tools.

8 AnT&CoW: Architecture and visualization

Following the Annotea W3C standard, our client/server annotation system implements a distributed architecture (Fig.5):
The client's goal is to annotate documents (for the moment limited to annotate text or HTML pages due to format con-

straints), which are accessible by a Web navigator. Mainly for this reason, we chose Annozilla, a Mozilla navigator plug-in which is an Annotea client following our aim. Using XPointer, DOM standards and many functions of the Mozilla infrastructure (XPConnect, XPCom components), Annozilla offers possibilities of creating, updating and deleting annotations on a document or a part of document and gives possibilities in storing them on a local server (individual use) or a distant one (shared use).

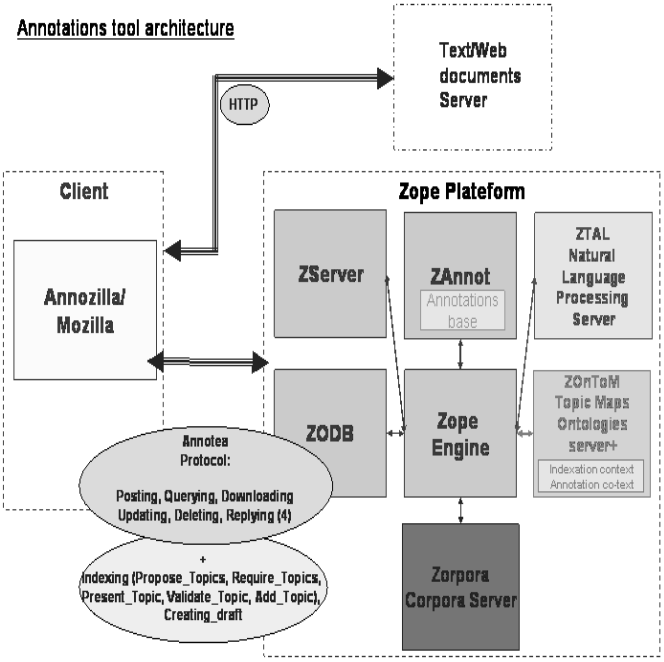

Fig.5 – AnT&CoW annotation tool Architecture

We chose a server respecting the Annotea standard, ZAnnot, developed on the Zope platform [Latteier and al, 2003] which has a Web server and several other components managing contents servers or databases. ZAnnot derives benefits from the Zope platform and manages at the same time queries sent by the Annozilla client and the reply function to an annotation.

On this platform, we encapsulate the ZTAL server for the natural language processing whose functions are defined above, the ZOnToM ontologies server represented out of TM also containing the indexation context and annotation co-text. The Zorpora server is a corpora server which contains not only the basic documents text used to constitute the domain-specific ontology dimension, but also the documents created by the project participants and eventually the authority documents shared in the project.

Since it is necessary to adapt the annotations client Annozilla for our annotation's purpose such as previously defined, we implemented the reply function from annotation to another and the indexing mechanism. To classify annotations, we extended the Annotea annotation diagram by adding metadata corresponding to our three dimensions which will be saved at the RDF format, as the other metadata and annotation bodies. For coherence reasons, our multi-dimensional Topic Map ontology is currently stored in a XTM (XML) format and is not modifiable by the user.

We provide an interface for the user allowing her/him to manage the topics of the different dimensions and to navigate through stored annotations. Navigation consists of a reading of the annotations arranged in one or more visible windows at the same time. Thus the user can, if s/he wished, display in the same document a set of annotation indexed by the same topic(s), annotation textual body and other fragments to which it is connected. (Fig.6) She/He has also possibility of recording elements gathered in only one new working document, a draft or a discussion paper shareable by the project.

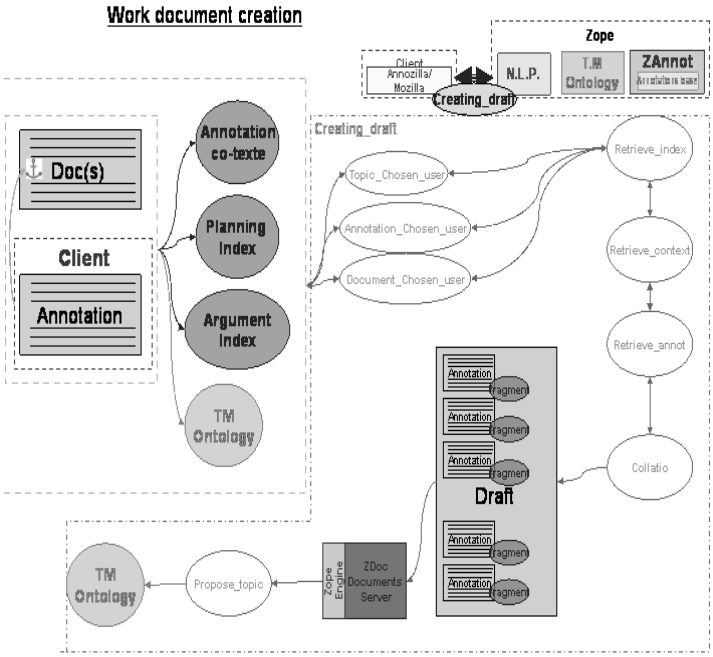

Fig.6: Work Document Creation

When a member of the project group is opening a document, s/he may open in the left side of the Web navigator main window, the Annozilla plug-in, which allows her/him to annotate as well as to retrieve and read organized annotations by means of their attributes defined above. If the author decides to create a new annotation, this annotation appears in a new window containing its body and the indexation fields in a pull-down menu as in this example with an electronic patient file (fig.7).

The next step in the tool development consists in integrating in our architecture the indexation elements, i.e. dimensions of the ontology and NLP tools; ZOnToM must be connected to the annotation server Zannot so that the TM ontology representing dimensions and the contexts/co-texts can be used for a semi-automatic indexing. The ontologies server installation in an on line process will also allow the ontology update, by way of user or of NLP tools.

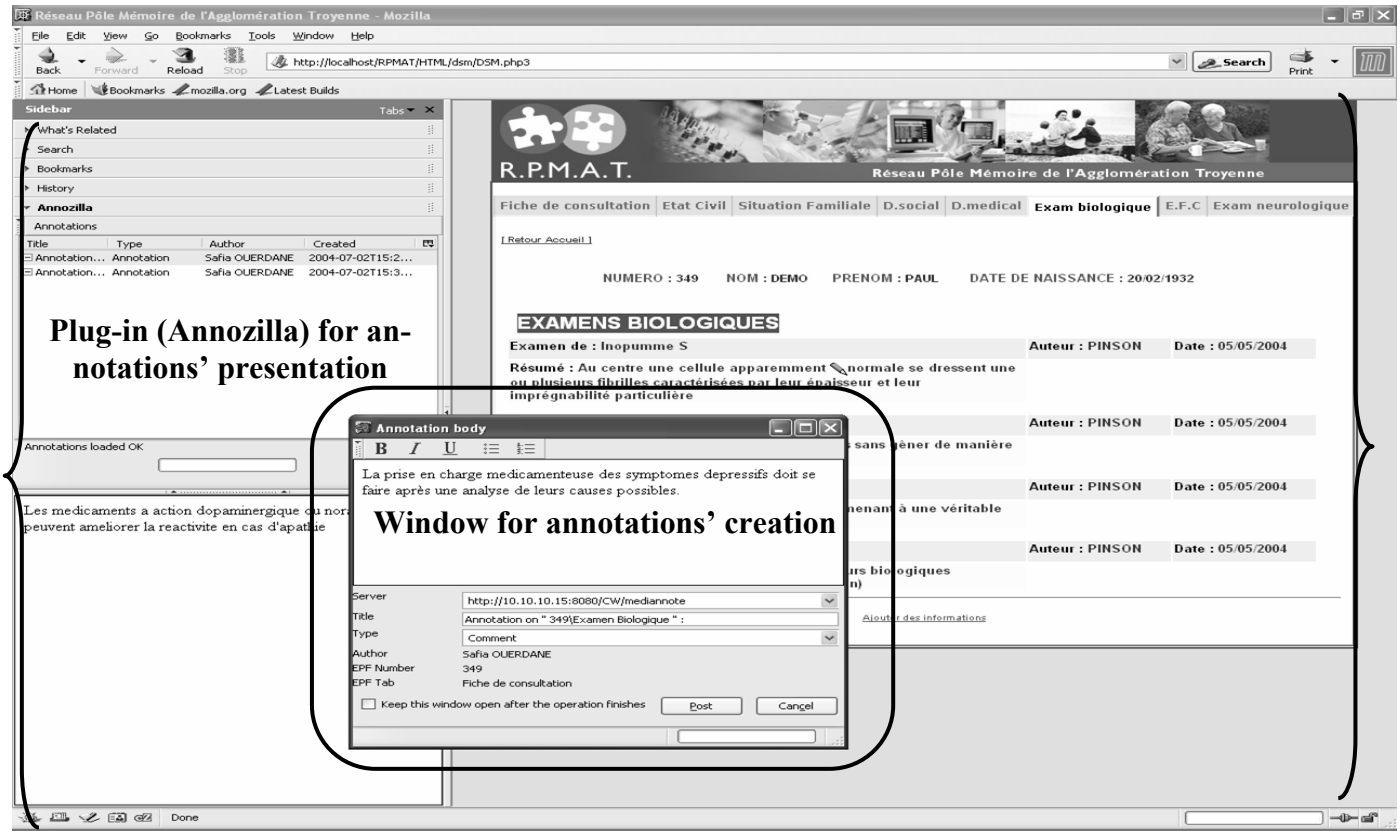

Fig.7 – AnT&CoW Interface for Electronic Patient File

9 Conclusion and prospects

The increasing number of electronic documents forces today's reader to adapt her/his practices. Traditional collective interpretation of texts by use of annotations then becomes an activity to be mediatized. Annotating is an activity mixing writing and reading and allows annotation's author to communicate with members of interest. We propose to define annotation as a kind of discourse, a structured set of memorized concepts which are reorganized as an editable structure aiming at communicating about a document.

To represent this discursive annotation activity and so collective interpretation of documents, we chose a classical rhetorical model of discourse production. Adapting this model to electronic document customs allowed us to design a groupware supporting sharable annotations for document based sense making within a group: AnT&CoW. Deriving from existing annotation's standards and tools, we drew some requirements for AnT&CoW, meeting our theoretical model.

AnT&CoW is a client/server application based on a multidimensional ontology. Our tool's features are supported by Natural Language Processing tools and techniques.

A first version of this tool is in development being in keeping with an iterative design approach. This tool will allow us evaluating our hypothesis not only on discourse production model, but also on annotations status and aims.

Acknoledgments

This research carries a CNRS (National Center for Scientific Research)/STIC (Communication and Information Science and Technology) department funding as part of TCAN (knowledge processing, learning and new information and communication technologies) pluridisciplinary project (Mediannote project).

References

[Acrobat pdf, 2004] *Acrobat PDF*, http://www.adobe.com/support/techdocs/ac76.htm, 2004.

[Annotea, 2003] Annotea http://www.w3.org/2001/Annotea/, 2003.

[Annozilla, 2004] Annozilla, http://annozilla.mozdev.org/, 2004.

[Baker, 2000] Baker M., The roles of models in Artificial Intelligence and Education Research: a prospective view. International Journal of Artificial Intelligence in *Education Research*. Vol 11(2), p. 122-143, 2000.

[Barré de Miniac, 2000] Barré de Miniac C.*, Le rapport à l'écriture : Aspects théoriques et didactiques* coll. Savoirs mieux Ed. Septentrion Presses Universitaires, Ch. Barré de Miniac, 2000.

[Biezunski *et al.,* 1999] Biezunski, M., Bryan, M., et Newcomb, S. R., *« Topic Maps », spécification ISO/IEC 13250*, 3 Décembre 1999.

[Bremer and Gertz, 2001] Bremer J.M., and Gertz M., Web Data Indexing through External Semantic-carrying Annotations. In *11th IEEE Int'l Workshop on Research Issues on Data Engineering: Document management for data intensive business and scientific applications (RIDE-DM'2001), IEEE Computer Society*, pp. 69-76, 2001.

[Brickley and Guha, 2004] Brickley D., and Guha R.V., *Resource Description Language* - http://www.w3.org/TR/rdf-schema/, February 2004.

[Buitelaar *et al.*, 2004] Buitelaar P., Olejnik D., Hutanu M., Schutz A., Declerck T., and Sintek, M., Towards Ontology Engineering Based on Linguistic Analysis, in *Proceedings of LREC'2004*, Lisbon, ISBN 2-9517408-1-6, pp.7-11, may 2004.

[Carruthers, 1990] Carruthers M., *The Book of Memory: A Study of Memory in Medieval Culture*. New York: Cambridge University Press, 1990.

[Cimiano *et al.*, 2004] Cimiano, P. Hotho, A., and Staab S., Clustering Concept Hierarchies from Text, in *Proceedings of LREC'2004*, Lisbon, ISBN 2-9517408-1-6, pp. 1721-1724, may 2004.

[Denoue, 2000] Denoue, L., et Vignollet, L., An annotation tool for Web browsers and its applications to information retrieval, in *proceedings of RIAO 2000*, 2000.

[Dingli, 2003] Dingli A., Next Generation Annotation Interfaces for Adaptive Information Extraction. In *6th Annual Computer Linguists UK Colloquium (CLUK 0*3), January, 2003, Edinburgh, UK, 2003.

[Domingue *et al*., 2002] Domingue J.B., Lanzoni M., Motta E., Vargas-Vera M., et Ciravegna F., Mnm: Ontology driven semi-automatic or automatic support for semantic markup. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02),* October 2002.

[Fayol, 1997] Fayol M., *Des idées au texte: psychologie cognitive de la production verbale, orale et écrite*. Paris: PUF, 1997.

[Gangemi *et al*., 2003] Gangemi, A., Guarino, N., Masolo, C., et Oltramari, A. Sweetening WordNet with DOLCE, *AI Magazine* 24(3): Fall 2003, 13-24, 2003.

[Garlatti and Iksal, 2000] Garlatti S., Iksal S., Méthodologie de conception de documents électroniques adaptifs sur le Web. in GAIO, M., TRUPINCIDE, E., *Document Électronique Dynamique, Actes du troisième colloque international sur le document électronique : CIDE'2000*, 2000.

[Handschuh *et al*., 2002] Handschuh, S., Staab S., et Ciravegna, F., S-cream - semi-automatic creation of metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, October 2002.

[Harris, 1988] Harris Z., *Language and Information Columbia University Press*, New York, 1988.

[Hayes and Flower, 1980] Hayes J. R. & Flower, L. S., Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, NJ: Lawrence Erlbaum, 1980.

[Herring, 1999] Herring, S.C., Interactional Coherence in CMC. *Journal of Computer-Mediated Communication* 4(4) : www.ascusc.org/jcmc/vol4/issue4/, 1999.

[Jacquemin and Bourigault, 2003], Jacquemin C. and Bourigault D., Term Extraction and Automatic Indexing, in Mitkov R. (ed), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, pp. 599-615, 2003.

[Jacquemin and Tzoukermann, 1999], Jacquemin, C., and Tzoukermann, E., NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25-74, Kluwer, Boston, MA, 1999.

[Kahan *et al.*, 2001] Kahan J., Koivunen M.-R., Prud'Hommeaux E., and Swick R.R., Annotea : an open RDF Infrastructure for Shared Web Annotations, *Proceedings of WWW10*, Hong-Kong, pp. 623-632, May 1-5 2001.

[Ka-Ping, 1998] Ka-Ping Y., *CritLink : Better hyperlinks for the WWW*. http://crit.org/ping/ht98.html, 1998.

[Kintsch, 1988] Kintsch W., The role of knowledge in discourse comprehension: A Construction-Integration model. *Psychological Review*, 95, 163-182, 1988.

[Latteier *et al.*, 2003] Latteier A., Pelletier M., McDonough C., and Sabaini P., *The Zope Book*, Edition 2.6. http://zope.org/Documentation/Books/ZopeBook/2_6Edition/ZopeBook-2_6.pdf, 2003.

[Marcoccia, 2004] Marcoccia M, On-line polylogues: conversation structure and participation framework in *internet newsgroups, Journals of Pragmatics*, 36 (2004) 115-145, 2004.

[MeSH, 2004] *MeSH, Medical Subject Headings*, http://disc.vjf.inserm.fr:2010/basismesh/meshv04.html, 2004

[Pédauque, 2003] Pédauque, R.T., *Document : forme, signe et médium, les re-formulations du numérique, working paper,* version 3- 8 juillet 2003, http://rtp-doc.enssib.fr, 2003.

[Piolat *et al.*, 1989], Piolat A., Farioli F., and Roussey J.-Y., La production de texte assistée par ordinateur. In G. Monteil, & M. Fayol (Eds.), *La psychologie scientifique et ses applications* (pp. 177-184). Grenoble : Presses Universitaires de Grenoble, 1989.

[Price *et al*., 1998] Price, M., Schilit, B., et Golovchinsky, G., XLibris: The active reading machine. In *proceedings of CHI'98 Human factors in computing systems,* Los

Angeles, California, USA, vol.2 of Demonstrations: Dynamic Documents, pages 22-23, 1998.

[Rousselot *et al.*, 1996] Rousselot, F., Frath, P., and Oueslati, R., Extracting concepts and relations from Corpora. In *Proceedings of the Workshop on Corpus-oriented Semantic Analysis*, European Conference on Artificial Intelligence, ECAI 96, Budapest, 12 August 1996.

[Roussey *et al.*, 2001] Roussey C., Calabretto S., et Pinon J.-M., SyDoM: A Multilingual Information Retrieval System for Digital in *proc. International Conference ICCC/IFIP On Electronic Publishing (ELPUB'2001)*, Canterbury (UK), 5-7 july 2001, p. 150-164, 2001.

[Sumner *et al.*, 2000] Sumner T., Buckingham Shum S., Wright M., Bonnardel N. , Piolat A. & Chevalier A., Redesigning the peer review process : A developmental theory-in-action. In R. Dieng, A. Giboin, G. De Michelis & L. Karsenty (Eds.), *Designing cooperative systems: The use of theories and models* (pp. 19-34). Amsterdam : I.O.S. Press, 2000.

[Tchounikine 2002] Tchounikine P., Pour une ingénierie des Environnements Informatiques pour l'Apprentissage Humain. *Revue I3 Information-Interaction-Intelligence*. Vol. 2, n°1, Cepadues Editions. 2002.

[UMLS, 2004] UMLS, Knowledge Source Documentation, 2004. http://www.nlm.nih.gov/research/umls/umlsdoc.html, 2004.

[Volz *et al.*, 2003] Volz R., Oberle D., Motik B., et Staab S., KAON SERVER - A Semantic Web Management System? In: *Proceedings of the 12th World Wide Web, Alternate Tracks - Practice and Experience,* Hungary, Budapest, 2003.

[Weick, 1979] Weick K.E, *The Social Psychology of organizing*, New York, Random House, 1979.

[Windows Word comments, 2003] *Windows Word*, http://office.microsoft.com/fr-fr/assistance/HA010714941036.aspx, 2003.

[Zacklad *et al.*, 2003] Zacklad M., Lewkowicz M., Boujut J-F., Darses F., and Détienne F., Formes et gestion des annotations numériques collectives en ingénierie collaborative, *actes des journées Ingénierie des Connaissances*, Laval, 2003.

[ZAnnot, 2003], ZAnnot, http://www.zope.org/Members/Crouton/ZAnnot/, 2003.

[Zweigenbaum *et al.*, 1994] Zweigenbaum, P; et Consortium MENELAS, MENELAS : an access system for medical records using natural language. In *Computer methods and programs in Biomedicine*, 45:117-120, 1994.