A method for capitalizing upon and synthesizing analyses of human interactions

Annie Corbel[1], Jean-Jacques Girardot[1], Kristine Lund[2]
[1]G2I / RIM, École Nationale Supérieure des Mines de St. Etienne, 42023 St. Etienne cedex 2
[2]I.C.A.R. (Interactions, Corpus, Apprentissages, Représentations), UMR 5191, CNRS, Université Lyon, École Normale Supérieure Lettres & Sciences Humaines, 15, parvis René Descartes, BP 7000 69342 Lyon Cedex, France
{Annie.Corbel, Jean-Jacques.Girardot}@emse.fr Kristine.Lund@univ-lyon2.fr

Abstract. It is often the case that analyses of human interactive activity are lost once an article is written about the results obtained. Although it is clear that corpora are gathered in order to answer particular research questions and that already collected corpora are often not adapted for answering new research questions, it is still interesting to reflect upon the capitalization and exploitation of analyses that have been carried out. For example, comparison of analyses, validation of analyses or alternatives modes of visualization could be possible. This article proposes a model of designation and extraction of parts of human interaction corpora using the *anchor* and *link* concepts that allow for experimenting on the reuse of analyses of human interactions.

Keywords: human interaction analysis, anchors, inter-coder reliability

1 Introduction

Many researchers are interested in the diverse forms of cognitive and social activities that take place when people interact together, for example, in teaching-learning situations or during cooperative problem-solving in the workplace. Computer Supported Collaborative Learning (CSCL) platforms, such as DREW[1] [1], [2] allow the researcher to collect and conserve computer-mediated interaction traces in the form of computer files. Researchers in the human sciences create other computer files when they transcribe (most often manually), the recordings of audio and video interactions. These two types of traces of human activity — issued from different sources — are the focus of analysis by researchers with particular objectives. Indeed a researcher will collect his or her data and thus define the type of trace, according to his or her research questions. As it stands today, these analyses, from which Ph.D. theses or articles are written, are not easily reusable and thus do not permit capitalizing upon analyses carried out for a given experiment or observed situation, or between different experiments or situations.

[1] Dialogical Reasoning Educational Website; see http://scale.emse.fr/

E. Tomadaki and P. Scott (Eds.): Innovative Approaches for Learning and Knowledge Sharing, EC-TEL 2006 Workshops Proceedings, ISSN 1613-0073, p. 67-76, 2006.

In this article, we address the possibilities of exploiting the analyses of traces of human interaction, for a single situation or across situations. The hypotheses we make and constraints that we recognize are the following:

1. The traces are available in the XML format, the semantics of which is known, at least informally. This is not a strong hypothesis: many CSCL tools directly produce such formats. In other cases, if the representation and the semantics of the traces are known, it is possible to convert them to XML format without loss of information.
2. The proposed approach does not prejudge the use of a specific tool or a prescribed format; it applies to the conjoint usage of different tools and methods of gathering traces, for example through one or more CSCL tools on the one hand, and by the manual transcription of audio or video, possibly with the help of an appropriate tool like [3] or [4], on the other.

One of the needs of the researcher in human and social sciences is to explore collected interaction traces in a pertinent and efficient manner, to annotate interesting phenomena and to obtain new documents that reflect the result of his or her activity. These new documents, represented in XML, will allow the comparison of these results within and across situations. Conversion into formats more appropriate for visualizing and disseminating results should also be made possible.

2 The form of human activity traces

In the context of previous projects (CESIFS[2], SCALE[3], COSMOCE[4]), the authors carried out different experiments using the DREW platform [2]. DREW proposes different types of interaction (chat, whiteboard, argumentation grapher, text editor) and manages the creation of a trace (in XML) of the computer-mediated human activity that DREW makes possible. This trace is collected in the form of a sequence of events, each event corresponding to a single participant's intervention: a message sent in the chat, an element created in the whiteboard, an argument for or against a thesis put into the argumentation grapher, etc. In the document generated, these events are conserved in the order of their appearance, the DREW server arbitrating between events that are quasi-simultaneous.

In the context of the European project SCALE, a larger platform was developed called the Pedagogical Web Site (PWS [5] [6]). The PWS can replay in real-time a DREW session, carried out, for example, by two learners in a cooperative problem-

[2]The 'CESIFS' project (Conception et Etude de Sites Internet pour la Formation Scientifique) or Conception and Study of Internet Sites for Scientific Training), was supported by the French region Rhône-Alpes 1997-2000.

[3]The "SCALE" project (Internet-based intelligent tool to Support Collaborative Argumentation-based LEarning in secondary schools) was financed by the European Union "Information Societies' Technologies (IST) programme (IST-1999-10664) of the 5th framework between 2001 and 2004; http://www.euroscale.net, http://drew.emse.fr.

[4]The 'COSMOCE' project (Conception, Outils, Supports, Médias, Organisation pour la Collaboration des Entreprises) or Conception, Tools, Support, Media, Organization for the Collaboration of Companies, was supported by the French region Rhône-Alpes 2003-2006.

solving situation. It is possible to visualize the trace of their activity in html format and to perform analyses on the nature of their activity (cf. for example the Rainbow framework: [7]). Some of these experiments have also been the object of audio and video recordings, these recordings having been manually transcribed by researchers, in order to obtain documents that can be manipulated on a computer.

The traces that were gathered in the context of these projects were for the most part in XML format. However, if one takes into account the wide variety of CSCL tools and transcription conventions followed by researchers, it seems illusionary to attempt to propose a common transcription/trace format or even hope to define a kind of "pivot format" that can represent human activity, whether it is through an exceedingly complex format that expresses all the nuances and variations possible or whether it is through a simplified format that expresses a lowest common denominator. It is simpler and more reasonable to imagine that the XML trace documents are conserved, unchanged, in their original form, as the researcher chose to record them. Consequently, it becomes necessary to furnish the researcher with a tool that permits him or her to explore the collected corpus through a friendly interface. The minimal functionalities that should be supplied are:

- The visualization of corpus extracts;
- The possibility to annotate elements of the corpus;
- A search mechanism for the corpus.

Some of these functions can be provided with simple programming. Others necessitate the definition of a model of designation and extraction of parts of interaction corpora. It is this last point that we address in the method described in the following sections.

2.1 Analysis of computer-mediated human activity traces

Many researchers are interested in the *processes* that make up social and cognitive human activity in teaching-learning situations or during cooperative problem-solving in the workplace as opposed to being interested solely in a final common product that may be the goal of such situations. Thanks to the automatic chronological recording of human activity mediated by computer, researchers have the technological means since the 1990s [8] to respond to a variety of questions centered on *process*. For example:

- How do learners use the tools put at their disposal in relation to the activities they carry out? [9];
- What is the role of argumentation in the co-construction of knowledge? [10];
- How does structuring computer-mediated communication interfaces change the nature of interaction? [11];
- How do the internal factors of interaction (e.g. social talk) correlate with cooperative profiles (e.g. symmetry of roles) [12].

It is clear that each research question requires obtaining carefully chosen data that through specific analyses allow a response to be formulated. It follows that certain collected traces will not be adapted to addressing research questions for which the traces were not designed. For example, if a researcher is interested in how social talk relates to role symmetry, he or she would need to observe a task where roles can be

either symmetrical or asymmetrical. On the other hand, the gathering of this same data would not help him or her in answering a question pertaining to structuring communication, if in fact learners were given the same communicative interface or indeed if they were speaking unhindered, face to face. However, if the task generated argumentation and involved complex concepts, perhaps the trace would be interesting for studying the co-construction of knowledge, even though it was not originally designed for that purpose.

Despite the constraint of research questions guiding data gathering, and that as a consequence, already gathered data is not systematically adapted to new research questions, it is nevertheless interesting to stock analyses of corpora in a database in order to further exploit and capitalize upon them.

So, what then do we mean by exploiting and capitalizing upon analyses of interaction? Firstly, researchers from different disciplines or researchers using different methodologies have been known to work on the same corpus, see for example [13]. It is interesting to reflect on how one could facilitate the comparison of these different analyses, thus confirming comparable results obtained from different methodologies [14] or generating new research questions. Secondly, when the same analysis method is performed on many interactions by different coders, inter-coder reliability should be performed [15] in order to ensure that the coders agree on how to apply the coding scheme in question and thus guarantee the results and ultimately the coding scheme's replicability. Thirdly, it should be possible to automatically generate visualizations of specific analysis results by translating the corresponding XML documents into formats readable by other software applications.

In order to understand how such issues may be treated by the method proposed in this article, we illustrate an example analysis below, beginning with the Rainbow framework, used for analyzing computer-mediated pedagogical debates [7].

2.2. Taking the Rainbow framework further

The Rainbow framework was developed as part of the European SCALE project (see above) in order to analyze the restructuring of argumentative knowledge during computer-mediated debate [7]. In the context of the method proposed for this workshop, we illustrate how analysis of interaction corpora using Rainbow can be supported and how the analysis of argumentative interactions can be taken further.

There are seven categories within the Rainbow framework (hence the name): 1) outside activity not having to do with the task at hand, 2) social relation, 3) interaction management, 4) task management, 5) opinions, 6) argumentation and 7) explore and deepen arguments. We do not have the space here to further define these categories (but see [7] for a full description); rather we use Rainbow as an example of a coding scheme that can be applied to traces of computer-mediated human activity (cf. Fig. 1) and on which our proposed method of exploitation and capitalization can be applied.

46	10:26:47	Mark	ok let's argue	4. Task management
47	10:26:48	Mike	go ahead	4. Task management
48	10:26:49	Mark	ok	3. Interaction management
49	10:26:53	Nigel	i don't like solution C	5. Attitudes, opinions, agreement
50	10:27:16	Nigel	because we won't have a good driving force	6. Provide (counter-) arguments
51	10:27:22	Mark	ah really I like it	5. Attitudes, opinions, agreement

Fig. 1 An example of the Rainbow framework applied to an extract of computer-mediated human activity translated from data from the COSMOCE project.

Fig. 1 shows how each chat intervention may be categorized according to the Rainbow framework. It becomes clear that once different researchers have coded a number of different interactions making up a single corpus, it would be interesting to automate comparison of analyses in order to perform inter-coder reliability and obtain percentage of agreement on the whole corpus. In addition, other analysis methods can be applied to the same corpus. For example, in the COSMOCE project [16], after performing analysis with Rainbow, we further analyzed Rainbow categories 6 and 7 in order to ascertain the finer relations between arguing and how arguments are discussed within collaborative conception, precisely because Rainbow was not elaborated to analyze situations where design is the task (Fig. 2 illustrates the concept with a short extract).

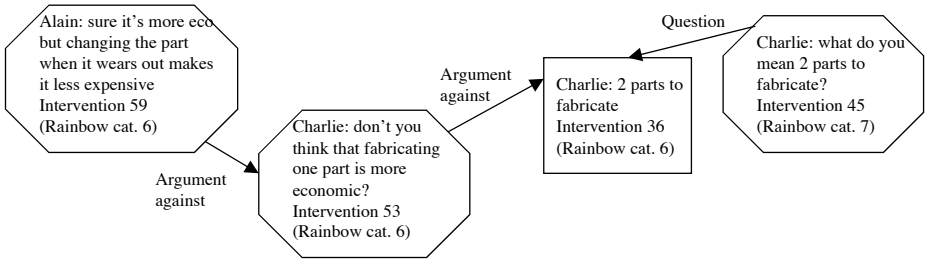

Fig. 2 illustrates a relational graph we produced that shows an example of the proposed relations between Rainbow categories 6 and 7.

In order to carry out this work, we needed to locate the chat interventions analyzed as Rainbow categories 6 and 7 in the original interaction and then propose semantic relations that existed between these interventions as a function of how we understood the designers to interpret their own discussion. We are currently developing a model of reasoning that describes argumentative activities of collaborating designers (cf. for example, [17] for a model of this type) for the situation we studied. We would like to perform these analyses on other interactions that have been analyzed by Rainbow in order to validate our model of reasoning.

The method we propose here (see the section below) is designed to support researchers in these kinds of undertakings: analysis according to a given coding schema, selection of analyses already done in order to perform further analyses, and finally comparison of analyses done by different coders or with different methodologies.

2.3 The proposed method

We begin by defining the term "primary corpus" (cf. [18] for an alternative definition) as the collection of all the documents gathered during the course of an experiment or observation. These typically consist of:

- Auditory or video documents that have been recorded during the experiment or during observation of the situation;
- Transcriptions of these recordings carried out by the researcher;
- Traces of computer-mediated interactions;
- Documents distributed to participants in the experiment/situation;
- Notes taken before, during and after the experiment/situation;
- All other documents judged to be pertinent by the researcher.

These documents are finite in number and will not evolve a posteriori, as they represent all the data gathered during and on the experiment/situation. In practice, we are interested in the documents that exist in computer format (having been originally generated in or translated into XML) for which an informal semantics can be defined.

We make the hypothesis that this primary corpus will be considered as fixed and unchangeable. All other documents created at a later date from this primary corpus will be an extract, a comment or an interpretation of the primary. Any annotations to the documents in this base will be expressed through an intermediary document (the "anchors document") that will create references to the primary corpus. It could be the case that a study is performed on different primary corpora, these will be globally called an "observation base".

The methodology described above allows us to constitute a corpus that contains all of the available data, without any information loss as no data is translated from one format to another. As mentioned previously, this corpus should be visualized and explored by the researcher. He or she should also be able to designate particular elements, annotate them and extract these elements or parts of them.

However, we cannot expect the human and social sciences researcher to master the different representations linked to specific software, even through the most friendly of XML editors available. We must therefore provide him or her with a tool that allows a visualization of the corpus he or she wishes to analyze.

Following an initial analysis of research practices, needs and existing tools, we propose the following tentative solution:

- The development of a generic browser, allowing for the visualization and the mark-up of the different documents that are part of the primary corpus.
- The development of an annotation tool, allowing for the linking of annotations to elements of the primary corpus.
- The development of an analysis tool allowing for the creation of links between elements of the corpus (a given chat intervention for example) and elements of the analysis method (for example, the task management category in the Rainbow method).

Documents pertinent to the analysis method (such as the enumeration of categories in Rainbow) constitute the *Analysis Base*.

Technological aspects

The use of XML [19] and the existence of related technologies allow us to list the specifications of these different tools.

Generic Browser

The use of formatting procedures for representing data contained in XML documents forms the basis of the Generic Browser. In our prototype, these procedures are written in Xquery [20], a language of interrogation and conversion, adapted to XML documents. Each particular type of XML document (for example a DREW activity trace) has an external file associated with it that describes which kinds of elements (in an XML sense) are considered as interesting by the researcher. A procedure for showing information (as defined by the researcher) is associated with such elements. It is the result of this procedure that is shown in the Generic Browser.

Mark-up Tool

The researcher in human and social sciences may at any time decide to mark up a specific element of the corpus. This mark-up process results in the creation of an *anchor*: a spatio-temporal designation of a corpus element. The anchor is an XML element that gathers diverse resources such as its type, a reference to a specific document in a primary corpus in the observation base, a geographical and/or temporal point in that document and complementary information (hour, date, author of anchor).

Each anchor is of a specific type, which describes how to interact with this anchor; this behavior is defined in in `anchor-type` XML elements, where, for example, an XQuery expression describes how to display the anchor in the Generic Browser.

The collection of anchors is conserved in an independent document. This document can also be explored with the Generic Brower, thus allowing the researcher to immediately bring up the *anchored* elements.

Link Creation

A link is a simple XML structure, made up of a group of labeled anchors. Each anchor designates an element of the observation base or an element of a primary document. The label of an anchor is an identifier that indicates the role of the anchor within the link. Each type of link is described by a `link-type` XML element that indicates the set of anchors that can be put in the link and how these anchors can be validated, and describes how the link should be displayed in the Generic Browser. Here again, XQuery is used for validating and displaying information.

Annotation Tool

The annotation tool is a simple structured text editor that allows the researcher to create an annotation document in XML. Each annotation is represented by an XML element and is designated by an anchor. Annotating a corpus consists in to creating the desired textual information and building a link between this information and the part of the corpus that is annotated

In this way, an annotation can be represented by a link that contains:

- An anchor on the comment created by the researcher
- An anchor (or more) on elements of the corpus
- An anchor on the document describing the researcher him or herself and the general objective of his or her work

(cf. Fig. 3 for an illustration of the relations between all the technological aspects described in this section).

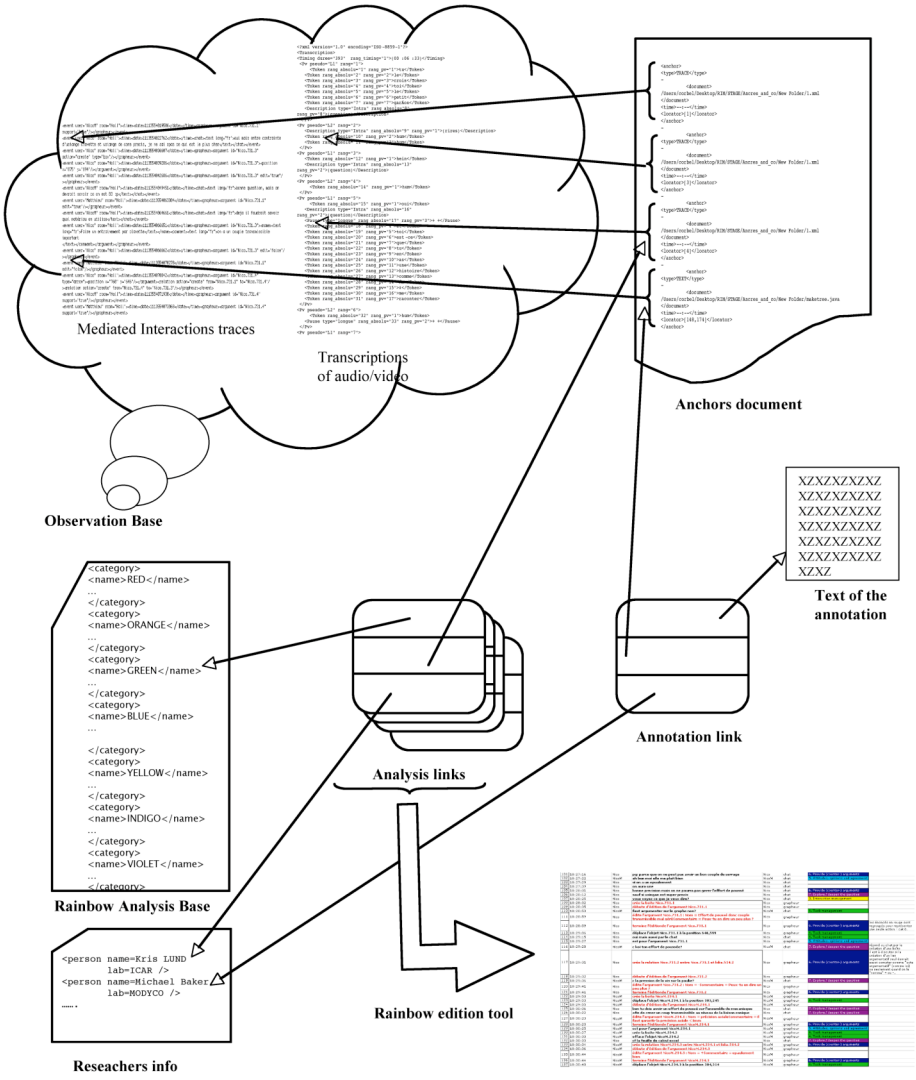

Fig. 3 illustrates the different components of the proposed method.

Analysis Tool

An analysis such as Rainbow (see the section *Analysis of computer-mediated human activity traces*) can thus be carried out with the help of the tools described above:

- The researcher can analyze the primary corpus by using the Generic Brower; he or she can create anchors on the elements deemed interesting;
- The researcher can also access an analysis base, a document in which the seven Rainbow categories are represented by anchors;
- It therefore becomes possible to place links between corpus elements and analysis categories.

The group of links thus created is in fact the analysis carried out by the researcher on the corpus. Once the analysis has been done, performing inter-coder reliability becomes straightforward. Analyses by different researchers on the same corpus can be compared and percentages of agreement calculated.

Since numerous kinds of computations and transformations can be performed on XML documents, the links resulting from an analysis can be used to provide usefull representations of this analysis; XQuery procedures can be designed to generate a representation of the result of an analysis in Word or Excel format, or create inputs for a graph drawing software such as Graphviz [21] (used in fig 2).

Computations can also be performed to provide global perspectives, like the summary of activities of individual participants, time spent in specific tasks, etc.

3. Conclusions and perspectives

A model of designation and extraction of parts of human interaction corpora was proposed. An initial prototype has been built according to the proposed model and will firstly be tested on a selection of computer-mediated human interaction traces by researchers using the Rainbow framework. Next, we will develop a second analysis base, based on a different analysis method and test its use by researchers. Our ultimate goal is to provide an observation base of primary corpora that through the definition of anchors, allows researchers to annotate, analyze, validate analyses and visualize data using a single adaptive tool with provision for future reuse of the work done.

Acknowledgments. The authors would like to thank their colleagues from the COSMOCE project, the EAIH project and the European project LEAD for the inspiration for this work.

References

1. Quignard, M. (2000). Modélisation cognitive de l'argumentation dialoguée. Etudes de dialogues d'eleves en resolution de probleme de sciences physiques. Thèse de doctorat de sciences cognitives. Grenoble : Université Joseph Fourier. [Cognitive modelling of argumentation dialogue. Studies of students in physics problem-solving].

2. A.Corbel, JJ.Girardot, P.Jaillon : DREW: "A Dialogical Reasoning Web Tool", ICTE2002, Int. Conf. on ICT's in Education. Badajoz, Espagne, 13-16 Novembre 2002.
3. CLAN: http://childes.psy.cmu.edu/clan/
4. TRANSANA: http://www.transana.org/
5. A. Corbel, P. Jaillon, H. Proton, X. Serpaggi : "A Guide for the PWS User, Scale Project". Research Report G2I-EMSE 2004-400-009, December 2004, Ecole des Mines de Saint-Etienne, 18 pages. http://www.emse.fr/g2i/publications/rapports/RR_2004-400-007.pdf
6. A.Corbel, P. Jaillon, X. Serpaggi : "PWS : An open tool to construct, use and study pedagogical sequences in collaborative argumentation training" CSCL Conference and Workshop, Budapest - Számalk 20-21 February 2004.
7. Baker, M. J., Andriessen, J., Quignard, M., Amelsvoort, M., Lund, K., Salminen, T., Litosseliti, L., Munneke, L. A framework for analyzing pedagogically oriented computer-mediated debates: Rainbow. Research Report IC-3. ICAR Laboratory, Lyon, France.
8. Jordan, B.,& Henderson, A.(1995).Interaction analysis: Foundations and practice . The Journal of the Learning Sciences, 4, 39–103.
9. Bouvery, J. (2006). Activite et Usage Synchrones en Situation de Conception Collaborative à Distance d'un Texte Procedural [Activity and Synchronous Use in the Collaborative Design of a Procedural Text at a Distance]. Master's Thesis in Cogntiive Science. University of Lyon, Lyon, France.
10. De Vries, E., Lund, K. & Baker, M.J. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. The Journal of the Learning Sciences, 11(1), 63—103.
11. Baker, M.J. & Lund, K. (1997). Promoting reflective interactions in a computer-supported collaborative learning environment. Journal of Computer Assisted Learning. 13, 175-193.
12. Lund, K., Rossetti, C., & Metz, Stéphanie (2006). Les facteurs internes à la coopération, influencent-ils l'activité médiatisée à distance? [Do internal factors of cooperation influence computer-mediated activity at a distance?] , M. Sidr, E. Bruillard & G.-L. Baron (Eds.) Actes des Premières Journées Communication et Apprentissage Instrumentés en Réseau JOCAIR '06, 6-7 Juillet, 2006, Université de Picardie Jules Vernes : Amiens, pp. 310-329.
13. Traverso V., Détienne, F. (forthcoming), Méthodologies d'analyse de situations coopératives de conception, Nancy : PUN.
14. Salomon, G. (1991). Transcending the Qualitative-Quantitative Debate: The Analytic and Systemic Approaches to Educational Research. *Educational Researcher*, Vol. 20, No. 6 (Aug.-Sept., 1991), 10-18.
15. B. De Wever, T. Schellens, M. Valcke, H. Van Keer, (2005) Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education* 46(1), 6–28.
16. Cassier, J. L., Lund, K., Prudhomme, G., & Pourroy, F. (2006, 28-30 juin). L'argumentation comme support à l'analyse de la dynamique des connaissances dans une situation de co-conception synchrone à distance [Argumentation as support for the analysis of knowledge dynamics in a synchronous distant design]. Poster presented at the 17e journées francophones d'Ingénierie des connaissances, Nantes, France.
17. Conklin, J., & Begeman, M.L. (1989). Gibis: A tool for all reasons. *Journal of the American Society for Information Science*, 200-213.
18. Balthasar, L. Bert, M. (2005). La Plateforme "Corpus de Languages Parlée en Interaction" CLAPI, historique, état des lieux, perspectives [The Platform "Corpus of Spoken Language in Interaction", history, state of the art and perspectives]. *Lidil : Corpus oraux et Diversité des Approches*, 31, pp. 13-33.
19. Extensible Markup Language (**XML**) : http:// www.w3.org/XML/
20. XQuery 1.0: An XML Query Language: http://www.w3.org/TR/xquery/ and http://www.w3.org/XML/Query/
21. Graphviz - Graph Visualization Software, http://www.graphviz.org/