1. Acronym of the project: CAP

2. Title of the project suggested: CAP: A Hierarchical Lexical Function Related to Proper Nouns

3 Coordinator: legal organization, organization type, legal address, department or unit, contact person, telephone and email.

- **Prof. Thierry GRASS** ; Université de Strasbourg ; UFR LSHA, LILPA ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Phone : +33 388 417468 ; email : thierry.grass@yahoo.fr
- **Dr. Amalia TODIRASCU** ; Université de Strasbourg ; UFR LSHA, LILPA ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Phone : +33 388 417429 ; email : todiras@umb.u-strasbg.fr

4. List of partners:

- **Equipe LiLPa** (Linguistique, Langues et Parole) ; Université de Strasbourg ; 14 rue René Descartes ; F-67084 Strasbourg Cedex ; Contact Person : Prof. Catherine SCHNEDECKER; Phone : +33 388 417885 ; email : Catherine.Schnedecker@umb.u-strasbg.fr
- **Institut de Traducteurs, d'Interprètes et de Relations Internationales** ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Contact Person : Eckhart HÖTZEL ; Phone : +33 388 417480 ; email : hoetzel@umb.u-strasbg.fr
- **Laboratoire d'Informatique de Tours**, Université François Rabelais ; 64, Avenue Jean Portalis ; F-37200 Tours ; Contact Person : Prof. Denis MAUREL ; Phone : +33 247 361435; email : denis.maurel@univ-tours.fr

5. Financing plan (budget and funding sources)

Our work is fund by Université de Strasbourg, as part of the institutional four-year contract (2008-2011) and by Université François Rabelais de Tours, as part of their institutional four-year contract (2008-2011). The fund cover salaries of permanent staff involved in this project (7 researchers, fig.1), as well as corpora, lexicon licences and tools (Fig.2). One Ph.D. student, who is fund by a private software company, and several MSc students work on the linguistic analysis of data.

We ask for CLARIN assistance to harvest existing resources and tools, by creating adequate WSDL services [WSDL], to develop an unified interface for querying simultaneously available corpora and tools and to integrate our own corpora into an existing resource federation.

Resources	**Languages**	**Category**	**Licence**
EuroWordNet	French, English, German	Lexicons	985 €
JOC	French, English, German	Corpus (tagged and lemmatized)	100 €
Le Monde (2004-2006)	French	Corpus (raw text)	626 €
British National Corpus	English	Corpus (tagged and lemmatized)	537,39 €
Le Monde diplomatique (2003-2004)	French, English	Corpus (raw text)	276 €
WordSmith	French, English, German	Monolingual concordancer	65 €
ParaConc	French-English German-English	multilingual concordancer	74.28 €
CasSys	French	Named Entity recognizer	Free s
ProlexBase	French	Multilingual Ontology	Free

Fig. 1. Available data and tools from LILPA (Université de Strasbourg) and LIF (Université François Rabelais de Tours)

Positions	Nb	Months	Salary	Fund source
Assistant professor	4	18 (25% of full research time)	30000 €	Université de Strasbourg
Assistant professor	1	18 (25% of full research time)	7500 €	Université François Rabelais
Professor	1	18 (25% of full research time)	15000 €	Université de Strasbourg
Professor	1	18 (25% of full research time)	15000 €	Université François Rabelais
Ph. D. Student	1	18 months (30%)	18000€	private enterprise Ready Business System, ltd., France

Fig.2. Human resources from Université de Strasbourg and LIF (Université François Rabelais de Tours)

6. List of languages covered: French, German, English

7. Description of the project:

The first aim of this project is to define a lexical relation called Cap through an exploration of corpora in three languages (French, German and English), from a lexico-semantic point of view. The second aim is to gather linguistic data in the three languages to illustrate the real use of the Cap relation. The third aim is to propose a complete description of the various proper noun categories related by this function, in order to extend the multilingual ontology Prolexbase (Tran et Maurel 2006).

In the field of lexical semantics, several relations, such as hyponymy, meronymy or synonymy, have been studied and described. The results of these studies were used to design electronic dictionaries (i.e. WordNet (Fellbaum 1998, Miller 1995)). However, these relations are not sufficient to organize 'world knowledge' or 'text knowledge'. **We will here focus on a particular relation called "Cap", which means "Head of".** This lexical relation is very briefly described in the *Explanatory and Combinative Dictionary of contemporary French* published by the team of Igor Mel'čuk (1984, 1988, 1992, 1999) and which was originally designed for French. As far as we know, this relation is completely original but it has been poorly described in French and other languages. This relation refers to nouns which express a hierarchy, that means a verticality in space. These nouns can be constituted through affixation (général, généralissime in French), through compositions (général, général en chef), through modification i.e. with an adjective (directeur général), through lexicalization i.e. in army ranks (général vs. colonel), or through acronyms (PDG = président directeur général).

We consider this hierarchy-founding relation as a fundamental relation in the structure of the mental lexicon because hierarchical relations exist in any kind of social organization. From a syntactic perspective, this relation links common nouns to proper nouns (*The new **president** [common noun] of the **United States** [proper noun] is **Barak Obama** [proper noun]*). From a semantic point of view, it may relate human to human (***John** is the **boss** of **Mary***), human to organisations (***Rick Wagoner** is the **CEO** of **General Motors***), but it can also be used with toponyms (in French, the expression "***chef-lieu***" means the main town of a region) or with the names of companies (parent company and subsidiary company).

This relation may help to develop industrial applications as text mining, information extraction and retrieval or machine translation (automatic or computer assisted translation). Concerning computer assisted translation, the Cap relation obviously contributes to the understanding of a text. As mentioned above, it may associate common nouns, which are more or less synonyms of "***chief***"[1], to proper nouns, which have as their properties to

[1] Note that French "chef" is only partially equivalent of English "chief".

represent a specific noun category that refers to a specific entity, not to a class of entities, and to play a significant role in discourse (Grevisse et Goose 1986).

For proper nouns are used to identify a specific entity, the pieces of information that serve as their definition are often an encyclopaedic description, together with morpho-syntactic properties (e.g. the use or the absence of determiner). From a lexicographic point of view, they do not have the same kind of definition as common nouns, they are quite often defined by their relations to other proper nouns.

Several projects study proper nouns to find criteria for an automatic extraction of Named Entities (Friburger *et al.* 2004, Poibeau 2006, McDonald 1996), other projects focus on their role in communication and in improving discourse coherence (Schnedecker 1997), other focus on identifying categories (Magnini et *al*, 2002), defining relations and specific properties (Grass *et al.* 2004, Sekine 2004). Our project aims to complete existing classifications and descriptions by proposing a linguistic description that accounts for the ways of expressing the Cap relation in a multilingual context (what common nouns and/or verbs are used in different languages) and that accounts for the categories of proper nouns that are related. We intend to analyse various linguistic expressions of this function at two levels. At sentence level, we will account for lexico-syntactic combinations; at text and discourse level, we will focus on reference chains and on the role the Cap relation plays in the text.

We plan to obtain a complete description of Cap relation and of its specific properties in the mentioned languages. The data collected from the corpora will be used to refine the set of relations and to extend the multilingual ontology Prolexbase (http://tln.li.univ-tours.fr/tln_prolex/prolex.php). Third, we will propose lexico-syntactic patterns to automatically identify the Cap relation from corpora, for German, French and English.

We intend to make available and to disseminate the results via the CLARIN infrastructure.

The material we plan to obtain may be of great interest for scholars in several fields of humanities, for instance scholars who carry out research about hierarchical structures and about the expression of hierarchical structures, namely linguists, sociologists, historians, but also economists and other scholars in the field of Business Intelligence.

Methods

We proceed to a systematic study of the linguistic expression of this relation, using various monolingual corpora (newspapers articles from the economy or politics domains, scientific-technical divulgation texts, law texts). For our analysis of “Cap” relations and the involved proper nouns categories, we use also multilingual, parallel corpora and some pre-existing lexical bases. We follow the main steps:

A. we prepare and select our corpora :

a) we identify a list of already available annotated corpora (tagged, lemmatized, corpora) from CLARIN project, mainly containing organization and person names;

b) we create a multilingual, parallel corpora, composed mainly of newsletters distributed by multinational enterprises (self made Swiss corpus);

c) we create monolingual specific subcorpora (from economics or politics domains) from online newspapers;

d) we tag, lemmatize, align and normalize (using TEI or XCES standards) our corpora;

B. we search the corpora to identify linguistic expressions of the function "Cap":

e) we establish manually a multilingual list of lexical units which are synonyms to "chief/chef" and to other nouns denoting a hierarchy-based function using lexical bases (WordNet and EuroWordNet) or multilingual corpus-based dictionary (Wortschatz http://corpora.informatik.uni-leipzig.de/) to identify synonyms from synsets. We use these synonyms to define lexico-syntactic patterns and we use these patterns to extract Named Entities related by these synonyms from the tagged corpora. We check manually the extracted data to identify valid linguistic expressions of the Cap relation;

f) we use a parallel concordancer on multilingual corpora (CLUVI corpus, Oslo corpus, Swiss corpus) to find equivalent expressions of the relation in the three languages;

g) we define several couples of proper nouns known to be related by the Cap relation (as *Martine Aubry* and the *Socialist Party* in France) and locate their contexts by using a stand-alone (WordSmith) or an integrated concordancer. We then analyse the extracted contexts to identify various linguistic expressions of the same relation.

h) we automatically extract person and organisation names using CasSys (Friburger et Maurel 2004) from the French subcorpus of our multilingual self-made corpus. Then, we exploit sentence alignment and person names to extract the contexts between the person and organisation names for the two other languages, in order to identify more linguistic expressions of the Cap relation.

C. we proceed to data extraction and analysis

i) we use the list of lexical units established in e) to search monolingual corpora and to extract contexts. We then manually analyze the extracted data in order to describe the relation and to propose examples to illustrate this relation;

j) we identify some subcategories of proper nouns involved by these relations using the Prolexbase (Tran et Maurel 2006, Grass *et al* 2004, http://tln.li.univ-tours.fr/tln_prolex/prolex.php) classification. On the basis of the extracted and analysed data, we propose a hierarchy of relations and we refine existing Proper Nouns categories. In addition, we define lexical patterns to automatically identify the Cap relation.

Data and tools

One of our aims being to study the lexical relation Cap, from a multilingual perspective, we focus on data identification and preprocessing. This work requires access to various monolingual tagged and lemmatized corpora, available in three working languages: French, German and English, as well as multilingual, aligned corpora. For our project, we need some tools to tag, to lemmatize, to align our corpora and to explore these corpora, so we chose a set of open-source tools or some tools already available in our laboratory (WordSmith, ParaConc).

A. Corpus preprocessing and selection

We create our own corpora and we use open-source tools for tagging and lemmatization. We create monolingual corpora (in XML format), by a manual selection of articles from economics or political fields extracted from on-line newspapers :

- The Guardian,
- The Observer
- The Independent
- Marianne
- La libre Belgique
- Journal interparlementaire franco-allemand
- Le Figaro
- Mouvements
- 24 Heures
- Der Spiegel
- Focus
- Vorwärts

We create a multilingual corpus from a set of newsletters distributed by the airline company Swiss, available in English, German and French (source language : German). We tag and lemmatise this parallel corpus (50 000 word/language). For this purpose, we used an open-source tagger available for the three languages French, English, German: TreeTagger (Schmid 1994). The input of this tagger is a raw text file, encoding in Latin-1 (English, German, French) or in UTF-8 format (for French only). The output of the tagger is a three-column text file, representing information about the token, the POS tag and the lemma, respectively. The POS tagset for English is the same tagset adopted by PennTreebank project (Mitchell et al, 1993). For German, we use the STTS tagset (Schmid, 1994) and for French we use a simplified tagset, containing mainly the lexical category and some morpho-syntactic properties.

In order to align our multilingual corpus, we use as well an open source linguistic platform, Unitex (Paumier, 2000). Unitex is a complex system containing a set of language tools and resources: dictionnaries, taggers, sentence alignment tools, available for several languages (French, German, English are supported). Unitex applies the tools on raw text in Unicode (Little Endian) format or on tagged corpus (the output of TreeTagger might be processed as well by this system). Moreover, Unitex contains a sentence aligner available for all language pairs and we apply it to align the Swiss corpus.

Validating the tagged and lemmatized corpora is a time-consuming task, so we would like to work with standardized tools and corpora in order to focus on the analysis of linguistic data, rather than creating new corpora. Thus, we selected the following list of corpora, freely available or provided by CLARIN project, relevant for the fields of economics and politics:

- Monolingual corpora:
    - L'Est républicain (CNRTL) (a local newspaper of the East of France) (French), raw text using TEI standard ;
    - Corpus Tècnic de l'IULA (English) http://bwananet.iula.upf.edu/indexen.htm, economics and law subcorpora, tagged and lemmatized corpora, Corpus Query Processing (Christ 1994) (CQP)-based interface ;
    - ZEIT-Corpus (Corpus of the weekly Die Zeit from 1946 - present day (complete runs from 1996)), tagged and lemmatized, using TEI standard (TEI).
- Multilingual corpora:
    - CLUVI: http://sli.uvigo.es/CLUVI/index_en.html (raw text) ;

- Oslo: http://www.hf.uio.no/forskningsprosjekter/sprik/english/corpus/index.html (SGML encoded, sentence aligned, tagged) ;
- JRC-Acquis corpus (French, German, English), already tagged, lemmatized and aligned for another project (Todirascu et *al*. 2008) (http://langtech.jrc.it/JRC-Acquis.html), in XCES format (Ide et Dorman, 2000).

In addition, we use existing corpora from our research team or free corpus, available online :

- JOC corpus (French, English, German), tagged and lemmatized following XCES standard ;
- Le Monde (French) (2004, 2006), raw text in XML format ;
- Le Monde diplomatique (French, English) (2003-2004), raw text in XML format ;
- COSMAS (Institut für Deutsche Sprache Mannheim), (German) http://www.ids-mannheim.de/cosmas2/uebersicht.html (SGML encoded, tagged and lemmatized) ;

B. **Detection of linguistic expressions of the Cap relation**

Once the corpora identified and preprocessed, we explore these corpora to search various linguistic expressions of the Cap relation, in the three languages. Before collecting various linguistic expressions from existing corpora, we consult as well lexical databases as thesauri and dictionaries to find synonyms and translations of the word "chef/head of/ Vorsitzender". We thus access to the multilingual dictionary (Wortschatz), available in CLARIN infrastructure. We use as well some existing lexical bases as EuroWordNet (Vossen, 1998) to identify synonyms of the various nouns denoting functions. We combine these synonyms and POS tagging (proper names) to to extract relevant data from the corpora, using existing online or independent concordancers. As well, we use the WordSmith concordancer and ParaConc concordancer to explore our standalone corpora, to extract contexts or to check the defined patterns. We analyse the extracted contexts to see if there are valid linguistic expressions of the Cap relation.

To explore our self-made parallel corpus, we apply ParaConc, a parallel concordancer identifying equivalent expressions in all the languages, using a set of ANSI texts. This tool requires a preprocessing step, preparing the aligned corpus as 1-1 associations.

For the further step, which consists of extracting relevant contexts of the person and organisation names known to related by a Cap relation (*B.Obama*, president of the *United States*), we use a stand-alone concordancer WordSmith to identify relevant contexts of these

Named Entities, from our own monolingual tagged and lemmatized corpora. The concordancer applies as well on raw or tagged texts, in Latin-1 or UTF-8 format. Moreover, we use online concordancers to explore selected online corpora. The output is a list of contexts, which are manually analyzed, to extend our list of linguistic descriptions of the Cap relation.

Beside the existing concordancers, we use as well the open source linguistic platform, Unitex (Paumier, 2000), presented in the previous section, to extract relevant data from our standalone corpus. One of the interesting functions provided by Unitex is the possibility of defining regular expressions and finite-state transducers, combining lexical and word information to search specific patterns in the corpus.

To extend the list of linguistic expressions of the lexical function "Cap" we apply CasSys, based on finite-state automata (Friburger *et al.* 2004) to identify and to categorize Named Entities as persons or organisations from French corpora. CasSys provides a set of transducers in Unitex graphs format and outputs an annotation of the Proper Names in XML format. We then exploit the sentence alignments and the person names identified by CasSys in the French corpora to identify the organisation names and the translated expressions in English and German.

C. Data extraction and analysis

We check the list of linguistic expressions on monolingual corpora, with the help of online or stand-alone concordancers (WordSmith, Unitex).. We manually analyse the extracted contexts of this relation in order to select examples of the relation. We proceed to a manual analysis of the proper names involved by this relation, in order to refine ProlexBase ontology, by adding categories and relations. Finally, we use the list of linguistic expressions identified in the previous step to define lexico-syntactic patterns describing the Cap relation in English, German, French.

Specific needs : *advice, tools, language data, transfer of skills* **from CLARIN**

CLARIN project aims to the development of a resource network (data and tools), useful for NLP systems, as well as for humans. For our project, we identified two categories of CLARIN support: 1) advice and technical support to selecti relevant source data; 2) technical support for querying selected resource.

One of the first requirements of our project is to find large monolingual corpora, provided with tagging and lemmatization information, available for the three languages, from the economics or politics fields. We want to avoid doing ourselves tagging and lemmatization, due to important costs of data validation. While we have already several existing corpus for French and English (tagged and lemmatized), we lack of German data as well as of multilingual corpus. We identify a set of existing online corpora, with an existing query interface, in the LRT inventory proposed by CLARIN. While our project concerns the study of the Cap relation from a multilingual point of view, we also retrieve in the CLARIN repository several multilingual parallel corpora. These online corpora could be explored by means of parallel concondancers, included into the search interface (CLUVI, Oslo).

Even if the LRT inventory is very useful to find relevant and freely available data, by browsing existing resources, the interface of this inventory proposes few search criteria, which are limited to language, resource category (written, spoken, annotated) or country. Due to the poorness of search criteria (we use only language and type criteria), we obtain an important amount of results and we have to read the description of each resource, and we select manually relevant data and tools. Thus, we expect from CLARIN to develop a federation of resources, and to propose tools to harvest existing data. In this context, CLARIN might propose a richer ontology of resources and NLP tools, indexing the existing data and tools. The ontology should be used by existing metadata components (IMDI, Broeder et Wittenburg, 2006) in order to help the user browsing the LRT inventory. We would then ask from CLARIN technical support to develop a search interface allowing us to select the corpora by domain or annotation standards, or to search the tools by providing the required input/output format.

While we use several categories of monolingual and multilingual corpora (online orself-made), we have no unified tool to explore these corpora and to extract relevant data. Online corpora provide project-specific interface to find concordances. These query languages offer the possibility of searching by word, tag or lemma information and the query interface proposes the selection of the interesting subcorpora, by domain, date or genre criteria. Meanwhile, it is difficult for us to collect data from these various corpora, because we have to translate first our patterns into each project-specific query language and then to analyse the extracted data. Thus, we ask from CLARIN to federate existing corpora by means of Web services (described in [WSDL] format) in order to launch simultaneously the same query on the available monolingual or multilingual corpora. These Web services should propose a detailed description of the data and the possibility of translating the query into the project-specific format query. While most of the selected corpora adopt Corpus Query Processing

(CQP) format for searching the data, the Web services should provide an automatic translation of the query into the project-specific searching language. Thus, we need advices and technical support from CLARIN to develop the query module searching for available resources, via Web services and to translate the CQP format into the project-specific language and to collect these results.

As well, we intend to distribute our tagged multilingual corpus through CLARIN infrastructure. We thus require technical support to create our own metadata, our own repository and our Web service in WSDL format.

8. Plan of activities

July 2009 – December 2009:

Pre-processing corpora: tagging, lemmatization, normalization, alignment (steps A: a), b),c),d)) ;

August 2009 – February 2010:

- Development of the federation of Web services to select available corpora and the unified query interface (CLARIN technical support);

October 2009 - March 2010:

- Identification of linguistic expression of the “Cap” relation, through existing corpora (step B, e), f), g), h)) ;
- Evaluation of the list of lexical units through monolingual corpora (step C, i) ;

March 2010 – June 2010:

- development of a web service for our corpus to integrate our resources into CLARIN resource federation (CLARIN technical support);
- Analysis of the extracted data in order to describe the relation and to propose examples to illustrate this relation (step C, i)) ;
- Classification of new Proper noun categories (step C, j));

July 2010 – October 2010:

- Definition of lexical patterns for English, German and French **;**

November 2010 – December 2010:

- Evaluation of lexical patterns and final report.

9. In France, several lexical semantics studies proposed complete linguistic description of generic relations as hyperonymy, hyponymy, meronymy or synonymy to structure dictionary knowledge (Polguère 2003, Mel’čuk 1984, 1988, 1992, 1999). As well, many linguistic studies focus on the definition of linguistic criteria to classify lexical categories (nouns, proper nouns), with respect to their morphosyntactic and semantic properties (Kleiber 1990, Kleiber 1999, Jonasson 1994, Vaxelaire 2005) or with respect to their behaviour in the discourse (Schnedecker 1997). Some of these results have been exploited to create electronic lexical databases as EuroWordNet (Vossen 1998), Synonym dictionaries (Crisco) or multilingual ontologies as Prolexbase (Tran et Maurel 2006, Grass *et al* 2004). Our work contribute to complete existing linguistic descriptions as well as to extend existing electronic dictionaries.

At a European level the use of the patterns which will automatically identify the Cap relation could be extended to operate on the standardized data the CLARIN infrastructure intends to provide. In the framework of this infrastructure the result of our research could become a tool to be re-used by other communities than linguists. These communities (sociologists, historians) might be interested in analyzing hierarchical structures synchronically or diachronically and could get linguistic data (from a diachronic point of view at least those translated from previous stages of the mentioned languages) for their research by using the lexical patterns our project aims to create.

10. References

Broeder, D. Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2), p.119-132.

Christ, O. (1994) *A modular and flexible architecture for an integrated corpus query system.* In COMPLEX'94, Budapest

Fellbaum, C.(1998). *WordNet, An Electronic Lexical Database*, The MIT Press.

Friburger N., Maurel D. (2004). Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, p. 94-104.

Grass T., Maurel D., Tran M. (2004). Prolexbase : Une ontologie pour le traitement multilingue des noms propres, *Linguistica Antverpiensia*, NS3:293-309.

Grass, T., Maurel, D., Piton, O. (2002). Description of a multilingual database of proper names. In Ranchod, E. & Mamede, N. (éds.) Advances *in Natural Language Processing*, *Lecture Notes in Artificial Intelligence* 2389, Berlin, Springer Verlag, p.137-140.

Grevisse M., Goose A. (1986), *Le bon usage*, Bruxelles, Duculot.

Grishman R., Sundheim B. (1996) Message Understanding Conference - 6: A Brief History. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, I, Kopenhagen, p. 466–471.

Ide N., Priest-Dorman, G. (2000). Corpus encoding standard - document CES 1. Technical report, Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, New York, USA and LORIA/CNRS, Vandoeuvre-lés-Nancy, France.

Jonasson, K. (1994). *Le nom propre. Constructions et interprétations*, Paris, Duculot.

Kleiber, Georges (1990). *La Sémantique du prototype , catégories et sens lexical*, PUF

Kleiber Georges (1999) *Problèmes de sémantique, la polysémie en questions*, Presses Universitaires du Septentrion

Magnini, B., Negri, M. Prevete, R. Tanev, H (2002). A Wordnet-based Approach to Named-Entites Recognition. In: *Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks"*, Taipei.

McDonald, J. (1996). Internal and external evidence in the identification and semantic categorisation of proper names, in: Boguraev, B., Pustejavsky J. (Eds.), *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, 1996, p. 32–43.

Mel'čuk, I. et al (1984-I, 1988-II, 1992-III, 1999-IV). Dictionnaire explicatif et combinatoire du français contemporain, Les presses de l'Université de Montréal.

Miller, G. (1995). WordNet: a lexical database for English. In: *Communications of the ACM* 38 (11), p. 39–41.

Mitchell, P. M., Santorini B., Marcinkiewicz, M.A. (1993) Building a Large Annotated Corpus of English: The Penn Treebank, in *Computational Linguistics (Special Issue on Using Large Corpora),* Volume 19, Number 2, p.313-330

Paumier, S.(2000). Nouvelles méthodes pour la recherche d'expressions dans de grands corpus, In Dister, A. (éd.), *Revue Informatique et Statistique dans les Sciences humaines* 36, *Actes des troisièmes journées INTEX, Liège, 2000*, p. 289-295.

Poibeau, T. (2006). Dealing with Metonymic Readings of Named Entities. *The 28th Annual Conference of the Cognitive Science Society* (COGSCI 2006). Vancouver.

Polguère A. (2003), *Lexicologie et sémantique lexicale. Notions fondamentales*, Presses de l'Université de Montréal

Schnedecker, C. (1997). Noms propres et chaînes de référence, Recherches linguistiques, no 21, Université de Metz.

Sekine S. (2004). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. Proceedings of LREC 2004 Conference, Lisbonne.

Schmid H (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

TEI - Text Encoding Initiative http://www.tei-c.org

Todirascu, A., Heid, U., Stefanescu, D., Tufis, D., Gledhill, C., Weller, M., Rousselot, F (2008) Vers un dictionnaire de collocations multilingue. In Blanco X ., L'Homme, M.-C., Campenhoudt, *Lexique, dictionnaire et connaissance dans une société multilingue*, *Cahier de Linguistique*. Revue de sociolinguistique de la langue française, Université de Louvain-La-Neuve.

Tran M., Maurel D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, Traitement automatique des langues, 47(3):115-139

Vossen, Piek (ed.) (1998). EuroWordNet: a multilingual database with lexical semantic networks for European Languages. In: *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*.

Vaxelaire Jean Louis (2005). Les noms propres — Une analyse lexicologique et historique, Paris, Honoré Champion, 952 p.

WSDL – Web Services Description Language http://www.w3.org/TR/wsdl20