

A Computer Science Electronic Dictionary for NOOJ

Farida Aoughlis

Faculté de Génie électrique et informatique
Université Mouloud Mammeri,
Tizi Ouzou, Algérie
fariyamo@yahoo.fr

Abstract. An automatic text analysis system cannot lexically recognize a word unless it already exists in the electronic dictionary. Our works applies to the NOOJ system. Work remains to be made to build terminological dictionaries. To build the terms dictionary, it is necessary to identify the terms in the texts, to count and to acquire them, to locate the terms by deciding if a word group constitutes or not an entry in the dictionary, and for all that the problem is to find corpora.

Keywords: Terminology, terminology extraction, terminology acquisition, electronic dictionary, compound noun, local grammar, NOOJ.

1 Introduction

Automatic translating software needs increasingly significant and varied terminological resources. They can be simple lists of terms that are more or less structured (structured indices, thesaurus, lexical networks) used by automatic indexing systems or for information retrieval or more documented terminological reference frames. [54]

The computer science compound words dictionary INFO_COMP that we build will allow the analysis and the automatic computer science texts indexing. The context of works is NOOJ was developed by Silberztein M. NOOJ is a new linguistic development environment, issued of author 12 years experience, as the INTEX user and designer at the LADL [48].

For the construction of our dictionary, we extract manually terminology from texts. We also extract terms automatically with NOOJ but here the terms extracted are examined if they are possible entries in the dictionary.

2 Terminology, Term and Specialized Language

The computer science texts are technical texts and the language is specialized. What do we mean by terminology? ISO, 1990 defines the terminology as the scientific study of the notions and the terms used in the specialist's languages. A specialized language [24], [25] is the use of a language that makes it possible to give an account

technically of specialized knowledge. This specialized knowledge is linguistically called by terms that are words, word groups. In a general language, the lexemes constituting the inputs of the dictionary are words; in the case of a specialized language, the lexemes are terms. In the linguistics dictionary [22], any discipline, and with stronger reason any science needs a set of terms, rigorously defined, by which it indicates the concepts which are useful for him: this set of terms constitutes its terminology. In what follows a term relates to a compound noun. One will find in [36] work an examination of the techniques of extraction of the terminological data and their impact on the terminologist work.

3 Tools for Automatic Extraction of Terminology

A census of the various existing systems is made in [16]. They make it possible to extract new terms starting from texts or corpus. Here also the various linguistic, statistical or mixed methods exist and are used to develop the tools for automatic extraction. Some extraction software evaluation criteria are examined in [35]. In [19], standard techniques for terms extraction are examined. In the thesis of [18] and in [21], specialized lexical pivots are used for the automatic acquisition of terms. In [20] a non-technical corpus is used for the extraction of terminology. An evaluation of acquisition tools for information extraction is made in [43]. In [31] we find studies about terminological variation. The majority of the authors [17] consider it essential to maintain a human activity in the acquisition systems, with an acceptance or a refusal of the results but also as far as being the acquisition centre, the computer processing will thus be reduced to a presentation and data-recording tool. Studies based on a semantic approach and terminology acquisition can be found in [40] and [16]. In the framework of the corpora specialized in terminology and the automatic acquisition of collocations, various works in progress [33], [41] and [45] are presented. A study about collocation can be found in [44]. In [38], Meilland J.C presents an automatic extraction terminology from short textual words. Smadja et al., [53] used statistical methods for translating collocations for bilingual lexicons.

3.1 Tools Using the Linguistic Methods

They are based on a syntactic analysis of the texts. A morpho-syntactic tagging locates all the noun syntagms (candidate terms). The expert retains the terms. Four systems are listed in [54]; we have TERMINO [15], LEXTER [8], [9], XTERM developed in 1999 by Cerbah, F. and FASTER [31]. Another system SINTESI (Systema INtegrato per TESTi in Italiano), is described in [24], the extraction of terminology is used to generate search keys for a database. To acquire English computer science terms the system LEXPRO [46] uses the system INTEX [48]. NOOJ [50] allows terms extraction.

3.2 Statistical Methods

System ANA [23] automatically extracts concepts from texts to produce a semantic network. MANTEX developed by [42] is founded on the repeated segments principle. In his doctoral thesis, [4] evaluates the statistical approaches capital in locating complex lexical units. MANTEX [45] extracts collocations.

In these methods, pertinent words having only one occurrence are not acquired (silence).

3.3 Mixed Methods

System like ACABIT [14] allows curing the problems of noise which arises in the linguistic methods. XTRACT [52] is a generic tool for location of collocations and not only of the terms.

4 Compound Nouns

In order to constitute our dictionary, a manual collection of the terms is carried out. It is significant to know if a compound noun can constitute an input in the dictionary. We are interested in the linguistic aspects of the terminology and particularly in the compound nouns composition.

4.1 Concept of Composition and Compound Nouns

All the completed research tasks try to define the concept of a compound noun composition but do not provide a magic recipe for their recognition. There is not a single definition of a compound noun but a certain number of common properties. The concept of nominal composition, very much discussed was approached by [10], [11], [29], [5], [6], [7] and [34] who admit that it is not possible to distinguish the compound nouns from free sequences. We will quote other work [37] and [39] for the syntax of the compound nouns, [32] on the lexicon of the compound nouns and [30] for syntactic and semantic automatic processing. Work on compounds N “of” N was studied without giving operational definition of “freezing” [1]. French compound nouns “freezing degree” notions for NN (Noun Noun) and NDN (Noun “of” Noun) categories are defined in [27].

4.2 How to Decide if a Compound Noun Is an Entry of the Dictionary?

A term can be simple if it contains one word or compound if it contains more than one. A compound word is built starting from simple words. Silberztein M. defines a compound noun as a consecutive sequence of at least two simple forms and blocks of separators. A simple form is a nonempty consecutive sequence of characters of the alphabet appearing between two separators. A simple word is a simple

form that constitutes an input of dictionary. We will use indifferently term or compound noun to indicate the same concept within the selected technical language (computer science).

A terminological bank is essentially composed of compound nouns. Each linguist uses his own terminology to define the compound words and proposes his own criteria. In our case the simple words are partly listed in the electronic simple words dictionary, the DELAS. The compound words are listed in the DELAC, but the terms concerning the specialized vocabularies are there in a very small number. Our task is to set up the computer science French terms dictionary, this dictionary will be thus a specialized dictionary. Will all the compound nouns found in the texts constitute an input in the dictionary? In order to decide if a word is or not an input of the dictionary we have to know what is a possible lexical entry? In [27] work morpho-syntactic, or semantic criteria are defined to make the distinction between free nominal group and compound noun. In [48] criteria defined a compound word and in [49] work, semantic is used and productive nominal groups and the lexicalised compound nouns are presented. We can find in [26] a study about “frozen sequences” and the semantic factors.

5 The Electronic Dictionary of Computer Science: INFO_COMP

NOOJ [51] is a development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to large corpora, in real time. The description of natural languages are formalized as electronic dictionaries, and grammars represented by organized sets of graphs.

For our specialized dictionary, at first we collected terms from texts, we studied the shape of an input, the compound nouns gender and number, the possible determinants, the particular inflection of the compound nouns and their local grammars. In [47], we can find the plural of some compound nouns. We defined the compound words classes for computer science terminology in [3], [2].

Our dictionary contains about 10 000 compound words. 30 000 compounds are extracted manually and will be added to the dictionary. With NOOJ, we also extract automatically candidate compounds, from texts and corpora. We add manually the terms retained in the dictionary with codifying the entries, it is the terminology acquisition.

5.1 NOOJ Grammars

A NOOJ grammar makes it possible to gather the terms by family; here, a family is the principal word (noun) in the compound, called the “head”. A grammar can contain several graphs and can be used for removing ambiguous forms. For the term card we have: card to band, card to card, card to disk...We give the local grammar “carte” (card) in the Fig. 1:

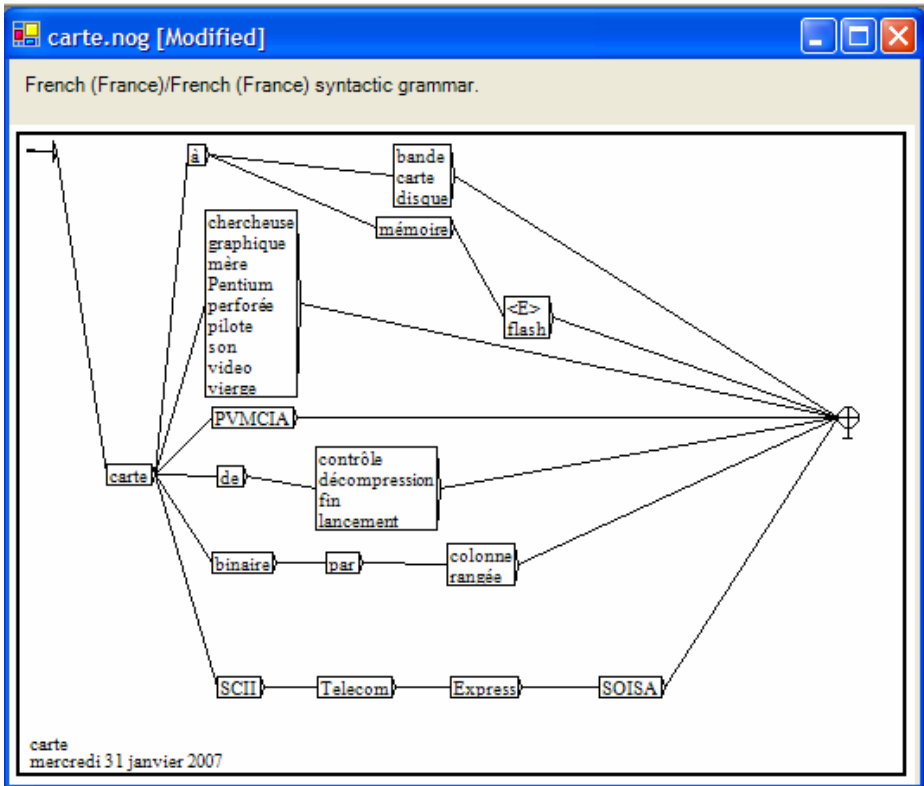


Fig. 1. A local grammar for the term “carte” (card)

5.2 NOOJ Dictionaries

NOOJ dictionaries [51] are a great enhancement over INTEX DELA-type dictionaries [12] as well as lexicon grammars. NOOJ dictionaries can represent spelling and terminological variants. DELAC is the dictionary of the compound nouns. Courtois B. studied binary compounds [13], Gross M. ternary compound nouns [28] and longest (4, 5 and 6 full words), only for French natural language.

In NOOJ, the INTEX dictionaries are represented in one unique format [50] the full description of the inflexion and derivation is encoded inside NOOJ dictionaries for the entries.

5.3 INFO_COMP Dictionary and Flexional Models .FLX

In the Fig. 2, For the term “accès aléatoire” (random access), we have the entry :

accès aléatoire, N+NA+info+FLX=AccesAccordé

accès aléatoire is a term, for this dictionary entry, we have the category N or NA:

N+NA; info is a semantic information: computer science term;

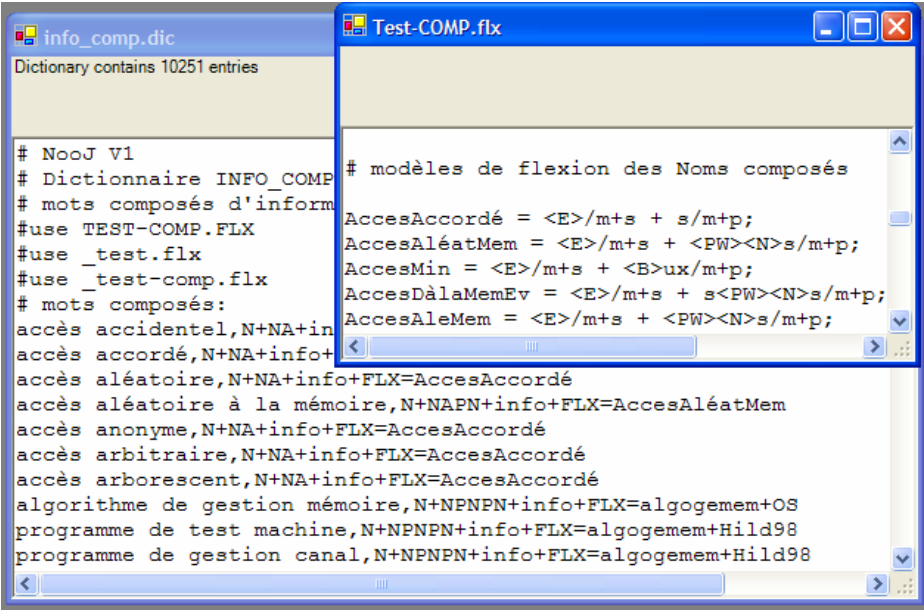


Fig. 2. Extracts from the dictionary INFO_COMP; the inflectional description Test-COMP

FLX gives the name of the flexional model here *AccesAccordé*, defined in the inflexional description file *Test-COMP.flx*:

$$AccesAccordé = \langle E \rangle / m + s + s / m + p;$$

The lemma “accès aléatoire” has two forms:

m+s => accès aléatoire	masculine (m)	singular (s)
m+p => accès aléatoires	masculine (m)	plural (p)

5.4 Automatic Extraction of Compound Terms with NOOJ

A pattern is a NOOJ expression and we can locate any pattern in the text. We use “**locate a pattern**” to extract compound nouns from texts or corpus, for example, in Fig. 3, a linguistic analysis of the text *uc.not* is made with NOOJ, the option *locate a pattern* is used to find the <N><A> (Noun Adjective) terms in the text. The list of concordance (candidate terms) is given, here 472. From the concordance for the text, the linguist selects the terms which are entries for the INFO_COMP dictionary and adds them manually to it (acquisition) with the format given at 5.3. We can locate any pattern we want [51].

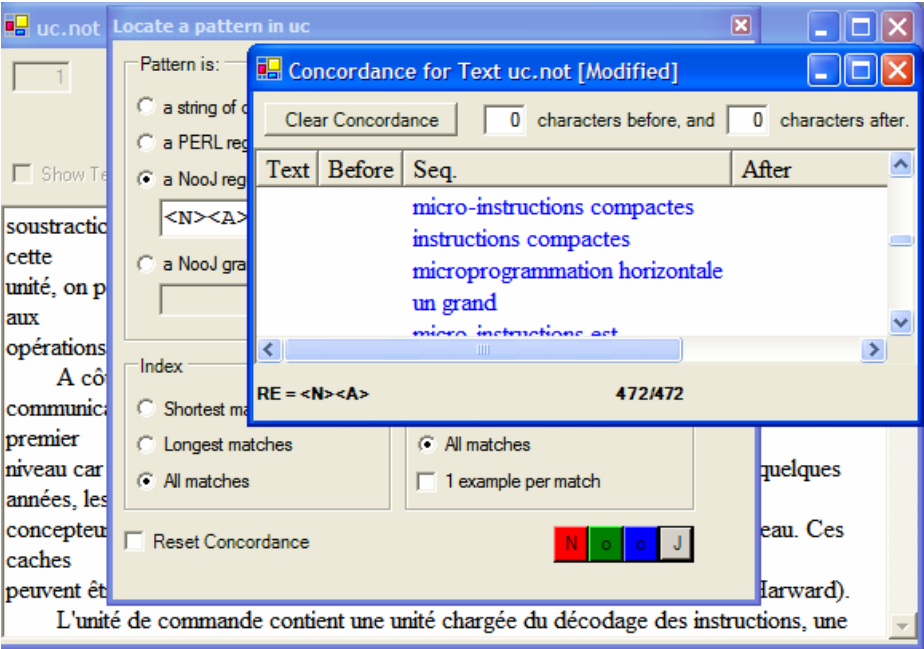


Fig. 3. Locating terms with a pattern here <N><A>

6 Conclusion

We started with newspapers terms were collected. These newspapers made it possible to collect terms but not much (3000), in spite of the size of the texts. Specialized books of computer science were taken as corpus and the number of terms clearly increased, because these books are of the studied speciality. Actually we collect manually terms from Hildebert dictionary and automatically from texts with NOOJ. We are building a big corpus of computer science with different texts of computer science from PDF, word and text files. The corpus contains actually 1071 text files; its size is 90 Mo. It remains to test terminology extraction from this corpus.

More than 10 000 terms were extracted and listed manually and added (terminology acquisition) to the INFO_COMP dictionary. 30 000 terms are collected and will be added to the dictionary. The manual collection of the terms is long and setting up a corpus for French computer science terminology is long and not easy. Manual extraction is long and needs to read big corpus. Semantic is often used to decide if a compound is a term. Others information (conceptual, syntactic, synonymous, links, translation in English...) can be added to the entry of the dictionary according to the use that one will make of it.

The elaborate dictionary INFO_COMP and the local grammars will make it possible to analyse computer science corpus, texts. With NOOJ, one will be able with this

new dictionary of terminology to treat computer science technical texts and to use them in various applicability such are the automatic indexing, the information retrieval, the machine analysis of texts, and machine translation. A translation in English will be added for each term.

It remains to finalize the coding of all the terms of the dictionary, to set up all the other grammars. Tests are then designed to analyse computer science texts, to index them. We expected to compare between NOOJ linguistic method of indexing and extraction with statistics methods like ANA.

Acknowledgements

I would like to thank Max Silberztein, from the LASELDI laboratory at the Franche-Comté University for his help and Elisabeth Metais from the CEDRIC Laboratory at the CNAM of Paris for all her remarks and help.

References

1. Ancombre, J.C-L.: Pourquoi un moulin à vent n'est pas un ventilateur. In: *Langue française* 86, Sur les compléments circonstanciels, Paris (1990)
2. Aoughlis, F.: Building an Electronic Dictionary of Computer Science. In: 8th INTEX NOOJ workshop, Besançon (2005), Presentation at <http://www.nooj4nlp.net>
3. Aoughlis, F., Métais, E.: Computer Science Terminology Extraction and Acquisition. *WSEAS transactions on Computers* 5(10), 2472–2478 (2006)
4. Balvet, A.: Approches catégoriques et non catégoriques en linguistique des corpus spécialisés, application à un système de filtrage d'information, doctorat thesis, université Paris X, Nanterre (2002)
5. Bauer, L.: *English Word Formation*. Cambridge University Press, Cambridge (1983)
6. Bauer, L.: *Introducing linguistic morphology*. Edinburgh university press, Edinburgh (1988)
7. Benveniste, E.: Fondements syntaxiques de la composition nominale » et « formes nouvelles de la composition nominale. In: *Problèmes de la linguistique générale*, Gallimard, Paris, vol. 2, pp. 145–176 (1974)
8. Bourigault, D.: LEXTER, un logiciel d'extraction de terminologie. Application à l'extraction de connaissances à partir de textes, doctorat thesis, Paris, Ecole des hautes Etudes en Sciences Sociales (1994)
9. Bourigault, D., Lepine, P.: Utilisation d'un logiciel d'extraction de terminologie (LEXTER) en acquisition de connaissances. In: *Acquisition et ingénierie des connaissances: tendances actuelles*, Cepaduès (1995)
10. Cadio, P.: À entre deux noms : vers la composition nominale. In: *Lexique 1*, PUL pp. 193–240 (1992)
11. Corbin, D.: Hypothèses sur les frontières de la composition nominale. In: *Cahiers de grammaire*, 17, Novembre 1992, le MIR, université de Toulouse (1992)
12. Courtois, B., Silberztein, M.: Dictionnaires électroniques du français. In: *Langue Française*, vol. 87, Larousse, Paris (1990)

13. Courtois, B., Garrigues, M., Gross, M., Gross, G., Rung, M., Mathieu-Colas, Silberztein, M., Vives, R.: Dictionnaire électronique des noms composés DELAC : les composants NA et NN, version 4.0, technical report LADL N° 55 (1997)
14. Daille, B.: Approche mixte pour l'extraction automatique de terminologie: statistique lexicale et filtres linguistiques, Thèse de doctorat en informatique fondamentale, université de Paris7 (1994)
15. David, S., Plante, P.: De la nécessité d'une approche morpho-syntaxique en analyse de textes. In: OICO, vol. 2(3), Québec, pp. 140–155 (1990)
16. De Chalendar, G.: STEVLAN: un système de structuration du lexique guidé par la détermination automatique du contexte thématique ».Thèse de docteur de l'université de Paris XI, Orsay, (2001)
17. De Chalendar, G.: Les systèmes d'acquisition de connaissances à partir de textes, (2002) <http://www.limsi.fr/individu/gael/manuscritThèse/html/node64.html>
18. Drouin, P.: Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés, Thèse de doctorat présenté à l'Université de Montréal, Montréal (2002)
19. Drouin, P.: Extraction de termes : techniques courantes. In: Ateliers sur les corpus spécialisés en terminologie, (2003) <http://www.ling.umontreal.ca/lhomme/cgi-bin/PD.zip>
20. Drouin, P.: Term extraction using non technical corpora as a point of leverage. Terminology 9(1), 99–117 (2003)
21. Drouin, P.: Acquisition des termes simples fondé sur les pivots lexicaux spécialisés ». In: Actes des cinquièmes rencontres Terminologie et intelligence artificielle (TIA), Strasbourg, pp. 183–186 (2003)
22. Dubois, G., et al.: Dictionnaire de linguistique, édition Larousse (1973)
23. Enguehard, C.: Acquisition naturelle automatique d'un réseau sémantique, thèse de doctorat de l'UTC, Compiègne (1992)
24. Fotopoulou, A.: Analyse automatique des textes: Dictionnaires et thésaurus électroniques des termes des télécommunications, Actes de la 5-ème journée ERLA-GLAT Brest, université de Bretagne occidentale (1994)
25. Fotopoulou, A.: Elaboration d'un thésaurus des termes des télécommunications, rapport, Institut National des télécommunications, Program Human Capital & Mobility (1995)
26. Gonzales Rey, M.I., Lopez Diaz, M.: De l'opacité des séquences figées comme exception sémantique. In: Actes-Exception-2003 (2003) <http://www.cavi.univ-paris3.fr/ilpga/Actes-Exception-2003/Auteurs/lopezdiaz-gonzalezrey/texte.pdf>
27. Gross, G.: Définition des noms composés dans un lexique-grammaire. In: Langue Française, septembre 1990, vol. 87 (1990)
28. Gross, M., Courtois, B.: Dictionnaire électronique DELAC: les noms composés ternaires et à plus de trois constituants, LADL (Décembre 1998)
29. Habert, B.B., Jacquemin, Ch.: Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques. In: traitement automatique des langues, vol. 2, Traitement automatique de la composition nominale (1993)
30. Habert, B., Jacquemin, C.: Constructions nominales à contraintes fortes et grammaires d'unifications. *linguisticae Investigationes* XIX-2, 401–427 (1995)
31. Jacquemin, Ch.: Variation terminologique et acquisition automatique de termes et leurs variantes en corpus. Habilitation à diriger des recherches en informatique, IRIN, université de Nantes (1997)
32. Jung, R.: Remarques sur la constitution du lexique des noms composés. In: Langue française, Paris, vol. 87 (1990)

33. Lemay, C.: Utilisation de corpus de référence pour dégager la terminologie d'un corpus ». In: Ateliers sur les corpus spécialisés en terminologie, Downloadable presentation. (2003), <http://www.ling.umontreal.ca/lhomme/cgi-bin/CL2.zip>
34. Levi, J.: The syntax and semantics of complex nominals. Academic Press, London, New York (1978)
35. L'homme, M.C.: Évaluation de logiciels d'extraction de terminologie : examen de quelques critères ». In: JIAMCATT, Réunion Inter institutions sur la terminologie et la traduction assistées par ordinateur, Office des Nations Unies, Vienne (Autriche) (2000)
36. L'homme, M.C.: Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. L'Impact des nouvelles technologies sur la gestion terminologique, (2001) (Downloadable text)
37. Mathieu Colas, M.: Typologie des noms composés. In: Rapport de recherches n° 7 du LLI, Université de Paris 13 (1989)
38. Meilland, J.C., Bellot, P.: Extraction de terminologie à partir de libellés textuels courts. In: P.U.R., (Presses Universitaires de Rennes), Linguistique de corpus (2003)
39. Monceaux, A.: La formation des noms composés de structure Nom Adj, élaboration d'un lexique électronique, Doctorat thesis, université de MLV, Noisy le Grand (1994)
40. Morin, E.: Extraction de liens sémantiques entre termes à partir de corpus de textes techniques, Doctorat thesis, IRIN, Nantes (1999)
41. Orliac, B.: Acquisition de collocations verbe + nom à partir de représentations syntaxiques. In: Ateliers sur les corpus spécialisés en terminologie (2003) Downloadable version at <http://www.ling.umontreal.ca/lhomme/cgi-bin/BO.zip>
42. Oueslati, R.: Aide à l'acquisition de connaissances à partir de corpus. Thèse de doctorat, Université Louis Pasteur Strasbourg (1999)
43. Poibeau, T., Dutoit, D., Bizouar, S.: Evaluating Resource Acquisition Tools for Information Extraction. In: Proceeding of the International language Resource and Evaluation Conference, (LRERC 2002), Las Palmas, Les canaries (2002)
44. Reymond, D.: La co-occurrence en T.A.L.: dis moi qui tu fréquentes et je te dirai qui tu es ». In: L'Ecriture dans tous ses états. Approche en sciences cognitives. Colloque 20 et 21 mai 2003, Université d'Aix en Provence (2003)
45. Rochibeau, B.: Extraction de collocations fondée sur les méthodes statistiques, in Ateliers sur les corpus spécialisés en terminologie (2003) downloadable version at <http://www.ling.umontreal.ca/lhomme/cgi-bin/BR.zip>
46. Savary, A.: Recensement et description des mots composés- Méthodes et applications. Doctorat thesis, université de Paris7 (2000)
47. Silberztein, M.: Le dictionnaire électronique des mots composés ». In: Langue Française, septembre 1990, vol. 87 (1990)
48. Silberztein, M.: Dictionnaires électroniques et analyse automatique de textes: le système INTEX, Edition Masson, Paris (1993)
49. Silberztein, M.: Les groupes nominaux productifs et les noms composés lexicalisés. In: Lingvisticae Investigationes XVII: 2, John Benjamins B.V, Amsterdam (1993)
50. Silberztein, M.: NOOJ's dictionaries. In: The Proceedings of the LTC (2005)
51. Silberztein, M.: NOOJ' manual, (2006), <http://www.nooj4nlp.net>
52. Smadja, F.: XTRACT: an overview. Computers and Humanities 26, 399–413 (1993)
53. Smadja, F., McKeown, K., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics 22(1), 1–38 (1996)
54. Term: Constitution de corpus à partir du Web pour l'acquisition automatique de terminologie: une expérience, (2003) www.poleia.lip6.fr/slodzian/sberland/

Appendix: INFO_COMP Dictionary Extract for the Term “Accès”

accès accidentel	accès conflictuel
accès accordé	accès d'arrivée
accès aléatoire	accès d'objet
accès aléatoire à la mémoire	accès de base
accès anonyme	accès de départ
accès arbitraire	accès de recherche
accès arborescent	accès de test
accès au destinataire	accès désiré
accès au support technique	accès direct glossaire
accès autorisé	accès direct à la mémoire
accès aux bases de données	accès direct à la mémoire évolué
accès aux données	accès direct au programme
accès aux services supplémentaires	accès direct aux mémoires
accès califourchon	accès direct en mémoire
accès complet	accès direct mémoire
accès complet au réseau	accès discret
accès concurrentiel	accès discu