An Electronic Dictionary of Computer Science Terminology

Farida AUOGLHIS (1), Elisabeth METAIS (2)
(1) Université Mouloud Mammeri, 15000 - Tizi Ouzou, Algérie
(2) CEDRIC / CNAM: 292 rue Saint Martin - 75003 Paris, France

Abstract: Automatic text analysis systems can lexically recognize a word only if it already exists in the electronic dictionary. The same thing is true for the NOOJ system analysis programs. One understands here by electronic dictionaries the lexical databases where all information is explicit because they are intended for computer programs use. These bases aim at the modelling of the language, which distinguishes them from the electronic lexicons created for particular applications needs. For technical languages or of speciality, work remains to be made to build dictionaries. Our work applies to the NOOJ system, with for immediate objective the installation of an electronic dictionary of French computer science terms (compound words), with an aim of analysing, automatically indexing texts. The linguistic aspects of the terminology are retained.

Key-Words: Terminology, terminology extraction, electronic dictionary, compound words, NOOJ.

1 Introduction
The exchanges internationalization, the always increasing production of documents under electronic format, the stakes related Internet and networks Intranet development create needs increasingly significant and varied in terminological resources. They can be [5] 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. [5]

The construction of a terminology [2] depends on the application in which one wants to use it. The selected terms and their degree of description are different according to whether one wants to build a reference terminology for a drafting assistance system or a lexical network to improve information retrieval on Internet.

The computer science compound words dictionary INFO_COMP that we build will allow the analysis, the automatic computer science texts indexing, using NOOJ system developed by Silberztein M. For the construction of our dictionary for NOOJ, the extraction of terminology is manual and we also extract terms automatically with NOOJ. This paper presents INFO_COMP, the electronic computer science dictionary of terminology for NOOJ. In section 2, we present the context of the works. Section 3 describes related works about terminology extraction. The section 4 presents the INFO_COMP dictionary, its entries, the inflexional descriptions, the NOOJ grammars and an example of automatic extraction of terms from a text, with NOOJ. We find our remarks and future works in section 5. An extract of the dictionary is given in Annex 1 for the term card.

2 Context of the works
The environment of development is NOOJ [21] which is a new linguistic development environment, issued of author Silberztein M. 12 years experience, as the INTEX user and designer at the LADL [18]. The LADL was the principal research center on the electronic dictionaries of French. NOOJ is free and can be downloaded at [21].

3 Related Works about terminology extraction
One will find in [14] works an examination of the techniques of extraction of the terminological data and their impact on the terminologist work. A census of the various existing systems is made in [10].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. The majority of the authors [10] 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.

3.1 Tools using the linguistic methods
They are also called symbolic methods. They are based on a syntactic analysis of the texts. A preliminary morpho-syntactic tagging locates the noun syntagms, to analyze them in order to extract the terms which are candidates. The candidate terms will be submitted to the expert to retain the compound words that are terms. Four systems are listed in [5], we have TERMINO [9], LEXTER [4], XTERM by Cerbah, F. developed in 1999, and FASTER [13]. The system LEXPRO [17] uses dictionaries DELAS and DELAC of English as well as system INTEX [18] to acquire English computer science terms. With NOOJ [21] we can extract terms.

3.2 Statistical methods

3.3 Mixed methods
System like ACABIT, developed by Daille B. [8] allows to cure the problems of noise which arises in the statistical methods with the introduction of linguistic owners and the use of statistical filters, with XTRACT, Smadja F. [22] gives a generic tool for location of collocations and not only of the terms.

4 The electronic dictionary of computer science terms: INF0_COMP for NOOJ
NOOJ developed by Silberztein M. is a development environment [21] used to construct large-coverage formalized descriptions of natural languages, and apply them to large corpora, in real time.

4.1 The compound words, the term
Our aim is to build an electronic dictionary of computer terms for NOOJ. 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).

We studied the shape of an input and the compound nouns gender and number, the possible determinants, and the inflection of the compound nouns. In [18], we find the plural of the compound nouns for the principal classes. In [1] we inventoried the different classes of compounds.

4.2 NOOJ dictionaries
The description of natural languages are formalized as electronic dictionaries, and grammars represented by organized sets of graphs.

NOOJ dictionaries [20] are a great enhancement over INTEX DELA-type dictionaries [6] as well as lexicon grammars. NOOJ dictionaries are similar to DELAS-DELAC dictionaries and can represent spelling and terminological variants. In INTEX [18], DELAS is the electronic dictionary of simple words DELAC [7] is the dictionary of the compound nouns.

In NOOJ [20], the INTEX dictionaries are represented in one unique format the full description of the inflexion and derivation is encoded inside NOOJ dictionaries for the entries. NOOJ dictionaries are used to represent, describe and recognize simple and compound words. Dictionaries are .nod files that are compiled from editable .dic source files.

4.3 Entry format of the dictionary INF0_COMP
The dictionary contains all the lemmas of the language and with the lemma we have a morpho-syntactical code, possible semantic and syntactic codes, inflectional and derivational paradigms. The fig.1 shows some entries. For the term *algorithme de gestion mémoire* :
  - The term: *algorithme de gestion mémoire*;
  - The category: N+NPNPN;
  - Info: informatique;
- FLX: algogemem is the flexional model name for this term;
- Hild98: text source of extraction.

4.4 Inflection of the compound words
NOOJ’s inflexion module is triggered by adding the special property “+FLX” to a lexical entry. NOOJ provides two equivalent tools to describe the inflexional paradigms:
- Inflection description files with .flx extension are set of rewriting rules,
- Inflectional grammars with .nof extension are special kinds of morphological grammars.

4.5 NOOJ grammars
A NOOJ local grammar (with file extension .nog) makes it possible to gather the terms by family, a grammar can contain several graphs.

4.6 Automatic extraction of compound terms with NOOJ
After the linguistic analysis, we can use “locate a pattern” to extract compound nouns from texts or corpus, here we want to locate <N> de <N> terms in the text:

The concordance located with the <N>de<N> pattern for the text infotest.not is in the Fig.5:
5 Conclusion and Future Works

These newspapers like "Le Monde Informatique" made it possible to collect manually terms but not much. Specialized books of computer science, dictionaries were taken as corpus for manual extracting and the number of terms clearly increased with the Hildebert [12] dictionary.

Actually, more than 10,000 terms were extracted and listed manually, and codified and added as entries of the to INFO_COMP dictionary. Others informations (conceptual, syntactic, synonymous, English translation) can be added at the entry of the dictionary according to the use that one will make of it.

30,000 terms and more are found and will be codified. The collection of the terms is manual; it is long and needs to read of big corpus.

We are building a big corpus of computer science with different texts of computer science from PDF, HTML, WORD and PS files. The corpus is in the Unicode text format. We meet difficulties to find French electronic corpus of computer science terminology. NOOJ can analyze the corpus and we can automatically locate new terms.

The elaborate dictionary INFO_COMP will make it possible to analyze computer science corpus, texts. With NOOJ, one will be able to treat computer science technical texts and to use them in various applicability such are the automatic indexing, the information retrieval, the automatic analysis of texts, and machine translation. It remains to finalize the coding of others terms of the dictionary, to set up all the other grammars. Tests are then designed to analyze the big corpus elaborated and computer science texts.

We plan to compare NOOJ terms acquisition with statistic and mixed methods.

Acknowledgements:
I would like to thank Max Silberztein, from the LASELDI laboratory from the Franche-Comté University for his help and all his remarks.

References:

Annexe.1
carte à 80 colonnes
carte à 90 colonnes
carte à bande
carte à carte
carte de chargement
carte à fenêtre
carte à mémoire flash
carte à mémoire
carte accélétratrice
carte binaire
carte binaire par colonne
carte binaire par rangée
carte chercheuse
carte circuit
carte circuit imprimé
carte courte
carte Datavoice So
carte de chargement
carte de contrôle
carte de décompression
carte de fin
carte de jeu électronique
carte de lancement
carte de police de caractères
carte de télécopie
carte demi-format
carte d'expansion mémoire
carte d'extension
carte d'extension
carte d'extension
carte d'interface
carte dorsale
carte fille
carte graphique 3D
carte graphique Hercules
carte graphique
carte intelligente
carte local bus PCI
carte magnétique
carte mère
carte mère Mercury PCI
carte mère NA H83
carte mère Pentium PCI
carte mère Pentium PRO
carte multi-fonctions
carte non peuplée
carte nue
carte PCMCIA
carte Pentium
carte perforée
carte perforée
carte pilote
carte programme
carte RAM
carte ROM
carte Scii Telecom Expres SOISA
carte son H
carte système
carte vidéo
carte video graphique
carte video local bus PCI
carte vierge