A computerized dictionary : Le Trésor de la langue française informatisé (TLFi)

Pascale BERNARD
ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118-CNRS/Université de Nancy2)
44 Avenue de la Libération, BP 30687
F-54063-Nancy Cedex, France
Pascale.Bernard@atilf.fr

Jacques DENDIEN
ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118-CNRS/Université Nancy2)
44 Avenue de la Libération, BP 30687
F-54063-Nancy Cedex, France
Jacques.Dendien@atilf.fr

Jean-Marie PIERREL
ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118-CNRS/Université de Nancy2)
44 Avenue de la Libération, BP 30687,
F-54063-Nancy Cedex, France
Jean-Marie.Pierrel@atilf.fr

Abstract

This paper presents one of the main computerized resources of the research Laboratory ATILF (Analyse et traitement informatique de la langue française) available via the Web: the computerized dictionary TLFi (*Trésor de la langue française informatisé*). Its highly detailed XML structure (over 3,6 million tags) is powered by Stella: the extended capacities and potentialities of this software allows high level queries as well as hypernavigation through and between different databases.

1 Introduction

A dictionary, in its electronic form, is a textual database to be used in any natural language processing system. The size and the contents of the existing dictionaries vary a lot according to the target of the attended public and the cost of their collected resources.

In the article, we present one of our textual resources, the TLFi which is the most important computerized dictionary on French language. It is accessible via the web at http://www.atilf.fr/tlfi.

2 Presentation

Le *Trésor de la langue française*, The Treasury of the French language, existed first as a paper version. It is a dictionary of the 19th and 20th century vocabulary, in 16 volumes. The first volume was published in 1971 and the last one in 1993. It contains about 100 000 head words with their etymology and history, that means 270 000 definitions, 430 000 examples with their source, the majority of them are extracted from the Frantext database.

The computerized version of the dictionary, the TLFi (*Trésor de la langue française informatisé*), contains the same data as the paper version, with its 350 million characters. With the help of very sophisticated automata, we have been able to insert in the text a very complex set of XML tags in such a way that every textual object is clearly identified and that the hierarchy containing these objects is clearly designed. With this tag set and thanks to its software Stella, it can be seen as a lexical finely structured database.

2.1 Stella: the Toolbox for the TLFi exploitation

As well as all the textual resources of the laboratory, the textual database FRANTEXT (www.atilf.fr/frantext), the 8th and 9th editions of the dictionary of Académie française and several others lexical database (Bernard et col, 2001 et 2002), the TLFi runs on its own specially software program STELLA, written in the laboratory,

Stella allows a compact data storage (with a mathematically demonstrable optimality) of structured texts (Dendien, 1996). Above all, it offers, to developpers, very powerful tools for the access and handling of textual data including several XML hierarchical taggings.

Stella offers the users:
- An environment to make the requests. The interface is very friendly, with a lot of help on line. It offers fine-grained request possibilities, allowing precision in the results.
- An optimal response time to all requests.
- A good quality of service: Stella contains a linguistic "knowledge" (flexions, categorized databases) which allows a user to make complex requests.
- A powerful capacity of interrogation: a user can write parametrable grammars to be used and re-used in different contexts.
- A possibility of hypernavigation throughout all the databases interconnected under Stella.

2.2 Specificities of the TLFi

2.2.1 The data:

Its originality is based, firstly, on its wordlist, which is rich of about 100 000 entries, present either in our funds or in other dictionaries. The TLF was a pioneer in the treatment of morphemes or in the treatment of structures of specific vocabularies.

Secondly, it is original too, by the richness of the number of examples, about 430000, and syntagms, about 165 000, quoted throughout its 16 volumes.

Besides, its list of metatextual objects such as headwords, definitions, indications of domains, semantic and stylistic indicators, examples with their sources, fixed phrases is exceptional. About 40 different metatextual objects.

The data are proposed in different sections: synchrony, etymology, history, pronunciation, bibliography.

2.2.2 The structure:

One of the main advantages of a computerized dictionary is to consider it as a knowledge database in which one can extract any items contained in any textual object. In order to get the most relevant answers when querying the database, the XML tag set is divided into two subsets : the tags of the first subset are used as delimiters and identifiers of the different kinds of textual objects used in the TLF ; the tags of the second subset represent the hierarchical organisation of the articles, in a way similar to the block structure of modern programming languages. This introduces very useful notions:

The **scope** of an object (from a pragmatic point of view it is the paragraph in which the object appears plus all the subparagraphs of this paragraph) is the smallest block containing this object.

The comparison of the scopes defines a binary relation between objects. Let us take the example of an article devoted to the headword W, containing the stylistic indicator "popular" and an example of Zola. Does this example illustrate a popular sense of W? The answer is yes if the scope of the indicator contains the scope of the example, otherwise the answer is no.

2.2.3 The different level of queries :

Three levels of queries are possible depending on the users's need.

2.2.3.1. First level : Simple visualization of an article.

You can read an article dedicated to a specific headword by three means.

Firstly, you have the possibility to write the word with mistakes if you do not know the right spelling of the word. That is very useful for the users who do not remember the right accents (acute, grave or circumflex) for instance. All kinds of mistakes (like bad or omitted accents, simple/double consonants, missing hyphens) are allowed. What is more, any mistake is possible as long as the right pronunciation is correct. For instance, if you write "ornitorink", the article "ornithorynque" will be found. It is also possible to enter an inflected form of a verb (ex. *danseront*) or of an adjective or noun (ex. *généraux*), or even a phonetic equivalent of such a form (ex. jénéro), even with bad accents (ex. *jenero*).

Secondly, you can use the possibility of seeing the list of the main articles contained in the TLFi, this allows the user to discover unknown words, just as if he was turning the pages of the paper version of the dictionary.

Thirdly, the user can find an article thanks to selecting sounds and not alphabetical characters.

At that level of consulting, you read the dictionary article by article, yet with easy ways of searching a word.

2.2.3.2. Second level : aided requests.

At that level you have the possibility of using the dictionary as a textual knowledge database and to make queries throughout the 16 volumes in one click of mouse. One can make requests on graphic forms, on inflected forms as well, on sequences of words ...

The requests can be mono-criterion or multi-criteria. Examples of mono-criterion requests : all the words borrowed from Spanish language, all the words of a specific domain, all the onomatopoeias, all the metaphors and so on. By specifying several criteria, one can extract from the dictionary all the nouns of a specific domain, all the verbs which are used with a stylistic indicator (for instance "popular") and which have been used by an author (for instance Victor Hugo). Other example : one can extract also all the definitions which contain a word (for instance *instrument*) and which at the same time do not contain the word *measure*, and which are found in the domain of *optics*, and so on.

2.2.3.3. Third level : complex requests.

The user, at that level, can seek for a set of textual objects $\{O_1, O_2,..., O_n\}$ imposing them to be conformant to a set of constraints combining the **type**, the **contents** and the **relations** between objects.

Type and contents specifications are sometimes possible in other computerized dictionaries query systems. They allow to find articles talking about architecture or containing an example contained from Zola. Suppose now that we are looking for articles where an example of Zola is related to the

domain architecture. The simple fact to combine the two criteria (must talk about architecture **and** contain examples of Zola), is not enough : may be some part of the article **is not** devoted to architecture. If the example of Zola is in such a part, the article is not relevant.
In the TLFi, the problem is solved with a new kind of constraint, by using the scope of objects. We will state that the example must be **hierarchically inferior** to the domain. Thus, all the articles where the example is not in the scope of the domain will be shifted.
This feature is nothing but the simple reflect of the XML tags representing the hierarchy. It gives the TLFi query system an incredible accuracy. Strangely enough, it seems to be ignored in other computarized dictionaries, with the consequence of very poor quality results.

This powerful feature, plus many other ones, such as the possibility to make list of words in many ways (manually, by automatical generation of the inflected forms of a given lemma, or by high level regular expressions selection) and reuse them in requests, allied to the rich content of the TLFi and a very friendly user's interface, with help on line, allow very complex querying, with pertinent results.

3 Hypernavigation

Hypernavigation throughout all the databases interconnected under Stella is possible. For example, when consulting the TLFi and by simply clicking on any word in an article, the user can navigate towards another article of the TLFi, towards a lexical database giving information on the grammatical category of the word, towards the 8th edition of the dictionary of the French Academy (1932-1935), towards the 9th edition (started in 1990), towards the historical database for French language, called DDL, giving new datations for words or phrases, and also towards the database FRANTEXT, which enables the user to discover other or more examples on a given word. Historically, the base derives from the texts assembled for providing examples to be used in the elaboration of the *Trésor de la langue française* (TLF). This first raison d'être turned to a new one: the desire to offer the scientific community a vast corpus of texts linked to an efficient query tool. Nowadays, FRANTEXT is a textual database which currently includes 3737 French texts covering a period from 1505 to the present day, amounting in all to more than 217 millions of words.

4 Uses of the TLFi

At the start the TLF in its paper version was intended to a public of specialists in linguistics, erudite persons and scholars.

The computerization of its data, the friendly interface with its help on line, the capacities of the software Stella give a new life to the TLF.

As described above, all the possibilities of queries allow anyone, specialist or not, to extract any piece of information from that textual database.

The TLFi is directed at the widest audience with the possibility of making queries without knowing the right spelling of a word. The majority of electronic dictionaries do not offer this possibility of access to words when spelling is unknown by the user. For instance, in the French word "sculpture", the letter -p- is mute. Thanks to this first query of research, the user can ask the word by its spelling : "skultur". The right answer will be immediately found. Besides, another friendly way to read the computerized TLFi consists in putting in evidence one or several textual objects by colouring them.

Its interest for linguistic research is obvious for it is a powerful tool to help the user who wants to study grammatical classes (verbs, adjectives, adverbs, for instance), syntactical classes (in studying constructions of a verb, more precisely, the user can find the verbs which are followed by the french preposition "de"), etymological classes (verbs borrowed from a language, Latin or English), stylistic classes (ironical or metaphorical uses, for instance all the adverbs used in slang), morphological classes (the words ending with a suffix, or beginning with a prefix, for instance all the verbs ending with the suffix –age and which are an action of a verb). It is also possible to exclude the content of a textual object, for instance if the user wants to extract all the words ending with the suffix –able which are not adjectives.

The use of the TLFi can also be foreseen for *learning and teaching French language*. Actually we are thinking, on one hand, about the pedagogic possibilities given by that computerized dictionary such as it is available at the moment, and on the other hand, about the modifications we could make in order to encourage its use in education

In the use of the TLFi in teaching such as it is at present, we have defined a methodology whose aim is to allow the use of this important linguistic resource. In collaboration with a teacher we have prepared and experimented a set of pedagogical activities for children in the CE1 class (7 years old). The results show that young children, who are learning to read, are able, alone or in pairs, to use the TLFi to write a word in capital letters, to

tell its grammatical category, to find a noun phrase or verb phrase containing this word or to say whether it is monosemic or polysemic.

As regards the modifications we could implement in order to get the TLFi to adapt to the learners, we could mention different types of modifications according to the ages of the learners. Firstly, we could, in the articles, change the literary examples extracted from the textual database Frantext for examples coming from another textual database containing youth literature. Secondly, we intend to present the information in a slightly different way. For instance, the syntagms could be found at start of a new paragraph, instead of being at a place where it is not easy to detect them for a young learner. Thirdly, we intend to give specific helps and commentaries on line to guide the young learner.

5 Conclusion

Designed and produced first as a paper dictionary, the TLFi is nowadays a computerized dictionary in which each user, linguist or not, can make successfully simple or complex requests. We will produce a CD-Rom version in autum 2004 but yet every one can have a free access at: http://www.atilf.fr/tlfi.

References

CNRS (1976-1993) Trésor de la langue française, dictionnaire de la langue du 19e et du 20e siècle, CNRS, Gallimard, Paris.

P. Bernard, C. Bernet, J. Dendien, J.M. Pierrel, G. Souvay, Z. Tucsnak. 2001. *Un serveur de ressources informatisées via le Web*. Actes de TALN-2001, Tours, Juillet 2001, pages 333-338.

P. Bernard, J. Dendien, J. Lecomte, J.M. Pierrel. 2002. *Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella*. Actes des 8e Journées Internationales d'Analyse Statistique des Données Textuelles JADT 2002, Saint-Malo 2002, pages 137-149.

P. Bernard, J. Lecomte, J. Dendien, J.M. Pierrel. 2002. *Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis : TLFi, Frantext and the software Stella.* Actes de LREC-2002, Las Palmas (Canaries).

Dendien J., Acces to information in a textual database : acces function and optimal indexes, in *Research in Humanities Computing, Papers from the ACH-ALLC Conference*, Oxford, Clarendon Press, 1996

J. Dendien et J.M. Pierrel *Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence* in « Les dictionnaires électroniques », M. Zock et J. Carroll editors, Traitement Automtique des langues, Vol 44 – n° 2/2003, Hermès – Lavoisier, Paris, 2003, pp. 11-37

C. Pélissier, C. Jadelot, J.M. Pierrel. 2004. *Méthodologie liée à l'Utilisation de Grandes Ressources Linguistiques dans le Cadre de l'Apprentissage : le cas du TLFi en Français au Cycle 3*. EURALEX 2004, Lorient.