

Adaptation of a computerized dictionary for language learning: The “*Trésor de la Langue Française informatisé*” and French language

CHRYSTA PELISSIER

ATILF & LIRDEF

17 quai du Port Neuf F-34500 Béziers

chrysta.pelissier@iutbeziers.univ-montp2.fr

CLAIRE BECKER

ATILF & CRAPEL

44, avenue de la Libération F-54063 Nancy cedex

claire.becker@atilf.fr

Abstract:

The objective of this article is to present the first results of a research related to the design of a linguistic resource dedicated to help pupils in their learning process. We intend to show how a dictionary for language specialists can be adapted in order to help children to learn French language.

By "linguistic resource" we mean any document (paper and computerized), which presents us with information related to language. Among these documents, we can quote language dictionaries, encyclopaedias, lexicons and glossaries, etc. The use of these documents is recommended for the preparation of the baccalaureate, the exam taken by pupils before leaving high school. Indeed, French teachers are asked by the French Department of Education to integrate dictionaries, encyclopaedias and databases in their courses, as well as documents related to the press [Ministère de l'Education Nationale 2002].

This paper first presents the main computerized dictionary called the TLFi (*Trésor de la langue française informatisé*, which can be translated as “Computerized Treasury of the French Language”) created by the research laboratory ATILF (Analyse et Traitement Informatique de la Langue Française). Then, we will present various uses of this dictionary for French learning in France. We will specifically focus on three didactic situations during which the dictionary is used.

1 General presentation of The TLFi

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. The *Trésor de la langue française* (or TLFi) [on the Internet at <http://atilf.atilf.fr/tlf.htm>] is the most important electronic dictionary on French language. It first existed as a paper version and groups the vocabulary of the 19th and 20th centuries, in sixteen 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 database Frantext. Frantext is a textual database, which gives access to more than 4 000 texts (the access to Frantext depends on a subscription). This resource is used by linguists, researchers and others specialists of French literature. Frantext gives the possibility to consult parts of text (which contain a word or an expression) and to consult a list of authors or a list of books (which were written by a particular author at a particular time). The computerized version of the dictionary, the TLFi, contains the same data as in the paper version; with its 350 million characters, its articles are structured according to the notion of textual objects. Thanks to its software Stella, it can be seen as a finely lexical structured database.

1.1. The specific data

Its originality is based, firstly, on its wordlist, which is rich of about 100 000 entries, present either in the funds or in dictionaries in the ATILF laboratory. The TLF was a pioneer in the treatment of morphemes or in the treatment of structures of specific vocabularies. Then we can say that its originality lies in the richness of the number of examples (about 430 000) and syntagms (about 165 000), quoted throughout its 16 volumes. Besides, its list of meta-textual objects such as headwords, definitions, indications of domains, semantic and stylistic indicators, and examples with their sources, fixed phrases is exceptional (about 40 different meta-textual objects). The data are proposed in different sections: synchrony, etymology, history, pronunciation, and bibliography. 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 eliminating noise in the requests. To allow this, the whole dictionary has been tagged into an XML document, with special delimiters for each type of its textual objects. Tags in the TLFi introduce possibilities to do queries.

1.2. Different level of queries

Three levels of queries are possible depending on the user's need. The first level is called "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 French accents (acute, grave or circumflex) for instance. All kinds of mistakes are allowed as long as the right pronunciation is correct. 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. The second level is called "aided requests". At that level you have the possibility of using the dictionary as a textual knowledge database and to make queries throughout the sixteen volumes in one click of mouse. One can make requests on graphic forms, on inflected forms as well, on sequences of words, etc. The third level is called "complex requests". The user, at that level, can make requests using regular expressions, lists of words or even more complex requests crossing criteria and playing with embedded structure or related textual objects. One can extract all the conjugated forms of the French verb "mourir" contained in the core of an example taken from Balzac. It is possible to make lists of words and to use them in requests. For example, one can extract all the words ending with suffix *-able*, and then extract from this list all words, which are not adjectives. The fine structure contained in the TLFi, allied to a very friendly user's interface, with help on line, allows pertinent results when making very complex requests.

1.3. Hyper-navigation

Hyper-navigation 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 between another article of the TLFi, Frantext if he needs more examples, a lexical database which gives information on the grammatical category of the word, the Academy dictionaries (8e and 9e editions), and the historical database for French language, called DDL. This function can help the user to construct his own knowledge. He can then navigate like he wants, to go deeper in a particular notion or a word in different resources.

2 Various uses of the TLFi

At the beginning, the TLF in its paper version was intended to a public of linguists and language specialists. The computerization of its data, the friendly interface with its help on line and the capacities of the software Stella give a new life to the TLF. The idea is to propose the TLFi as a tool for different pupils (pupils of primary, secondary and high schools). Thus, a group of researchers in the ATILF laboratory gathered together in September 2003 and created the team called EDUC'ATILF (www.atilf.fr/pedagogie). The objective of this team is now to show how important resources like the TLFi and Frantext can be adapted to different types of pupils. A particular methodology has then been used for our research [Pélissier & al. 2004]. First, we identify the knowledge concerning French language learning in primary schools. Then, we precisely describe the resource with its data and their organisation. In addition, we characterize some didactic activities using the TLFi. Indeed, we only have made a study of Frantext for high schools pupils (15-18 years old) since some texts in this database are studied within the framework of the French baccalaureate. We have set out three different uses of the two main linguistic resources of the ATILF laboratory: Frantext and the TLFi. These uses will be described later on in this article. First, we will explain the use of the TLFi for primary school pupils, then we will cope with the use of the TLFi in secondary schools and finally, we will deal with the new tool made up of Frantext and the TLFi for teachers and pupils in high schools.

2.1 Uses in primary schools

In France, the primary school begins for 5 year-old-children and ends when they are 12. For this public, some dictionaries are proposed.

2.1.1 The position of dictionaries in learning French language

Since 1972, French educational official texts give dictionaries a real place in the classroom. Nowadays, teachers advise parents to buy a dictionary for their children. This is the reason why these books nowadays have a place in French family and school. Indeed, 8 students out of 10 own a dictionary and about 30 000 dictionaries are sold every year [Gross 1989].

2.1.2 Different dictionaries for children

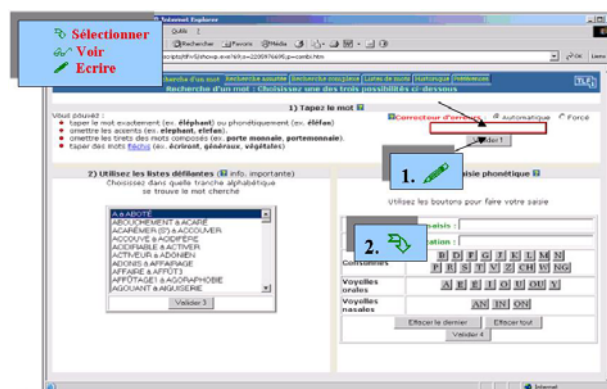
A lot of dictionaries for children are available in the market. Pupils, their parents and teachers encounter some difficulties in choosing the most adapted dictionary since it exists a lot of paper dictionaries, electronic dictionaries or

language and encyclopaedic ones, associated to different cycles of primary school. The repartition of dictionaries is made according to the age and the classroom of the child ([Lehmann 2000]; [Buzon 1983]).

2.1.3 Problems of the dictionaries in use

The majority of electronics dictionaries generally propose only one possibility of access words. For example, in *Mon Premier dictionnaire Super Génial* [Nathan 98], only the selection of the first letter of the unknown word is given. But, for 5-year-old child, it is impossible to isolate, identify and select the first letter of word. Thus, the majority of electronic dictionaries does not offer this possibility of accessing to words when spelling is unknown by the user. 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 and adverbs for instance), syntactical classes (in studying constructions), etymological classes (verbs borrowed from English), stylistic classes (ironical or metaphorical uses) or morphological classes (words ending with a suffix or beginning with a prefix). The use of the TLFi can also be foreseen for learning and teaching French language. We are now thinking of the pedagogic possibilities given by that computerized dictionary such as it is available at the moment. We also work on the modifications we could make in order to encourage its use in education. In order to use the TLFi as it is available nowadays, we have defined a methodology whose aim is to grant pupils an access to this important linguistic resource. In collaboration with a teacher we have prepared and tried out a set of pedagogical activities for children in the CE1 class (7-year-old children).

To illustrate our topic, here is an extract of a didactic activity made for CE1 pupils and presented to them afterwards.



Picture 1: Pupil's sheet for the use of the TLFi

2004 with two groups of 12-year-old children. These experiments have been made in order to define whether the graphical user interface of the TLFi allows usage in autonomy by children of this age without being helped by a technical sheet or by the teacher. Children had to look for the word “*oiseau lyre*” in two dictionaries (the TLFi and the TV5 dictionary available on the site <http://dictionnaire.tv5.org/>) in order to compare the definitions. They also had to give their impressions concerning the handling of these two tools. We can point out that pupils have found that the TLFi is very powerful and offers a big amount of information in its definitions; however, on the practical point of view, they had preferred a more easy to handle electronic dictionary (more rapid and less austere) to the TLFi. Furthermore, we have created a pedagogical activity for 14-year-old pupils based on the construction of French compound nouns and their plural. Pupils had to find the right spelling of compound nouns, to understand the sense of affixes and their behaviour. The last exercise for them was to identify grammatical rules and to write them down. However, the experiment has not taken place yet.

2.3 Uses for High school students

2.3.1 General overview

The environment LyText (Lycée + Textes) is now a part of the regional project called “e-Lorraine”. The purpose of “e-Lorraine” is to give French teachers and high-school pupils a linguistic resource adapted to their needs. The resource, named LyText, will be available on the Internet, on the “e-Lorraine” website (www.e-lorraine.net). LyText has two main goals: helping pupils to prepare the French baccalaureate and giving a tool for the teachers, helping them to prepare their lessons as well as the exams taken by the pupils.

2.3.2 Various types of modules

Since it is dedicated to language teaching and learning, LyText is made up of three modules called “Texts visualization”, “Texts preparation” and “Training”. The module “Texts visualization” shows extracts of books associated to an informational model. These extracts can be the subject of a particular work in class such as a preparation for the oral test or the constitution of the corpus for the written test. In these two tests, the pupil has to answer one (oral exam) or several (written exam) questions about one (or several) extract. The pupil has to understand the text to answer these questions. He also has to interpret it. So as to facilitate his work, this module will propose him, for each text, to visualize an informational model made up of various pieces of information given by the text. For example, these elements are related to the textual structure, the lexicon or the

stylistic devices in the text. The module “Texts preparation” makes it possible for the teacher to choose books belonging to the same period (a century, a particular date), to the same literary movement (romanticism, classicism, humanism, realism...) or to select books having the same genre (poetry, plays, autobiography...) or belonging to the same literary topic [Bouty, 1992]. He can then select the particular extracts on which he wants his pupils to work. Thus, the teacher will be able to define new groups of texts or to select new extracts of books studied for the oral test. This module also gives the teacher the possibility to modify (to remove or to add) information associated to each text in the informational model. For instance, with LyText, the teacher will be able to remove the presentation of a stylistic device like the zeugmas or the metaphors, to modify the lexical fields found in the text (to remove a word or a group of words belonging to a particular field). Eventually, the module “Training” gives the pupil the possibility to train himself to determine which pieces of information are relevant for his own analysis of the text. These pieces of information are those of the informational model presented in the module “Texts visualization”. Thus, the pupil can choose to work on one or more extracts of books (which he already knows or not), to train himself to determine one or more types of information (lexical fields, stylistic devices, connectors...) that he chooses to seek inside this (or these) text(s).

2.3.3 Use of the TLFi in LyText

In LyText, the TLFi is used on two levels: the former is the user level, the latter is the designer level.

On the level user (in the module “Texts visualization”), the informational model gives the definition taken from the TLFi, for certain words of the text, which seems to be the most adapted. This definition helps the pupil understand the word in its context [Fayol, 2003]. In the module “Texts preparation”, the teacher can determine the words of each extract, which he thinks they might pose a problem in understanding. Then, he can choose to give for each of these words one or more definitions extracted from the TLFi and/or various other information such as the etymology of the word. Finally, in the module “Training”, the system will present the text and the possibility offered to the pupil to click on each word (or group of words). Then, he can pick the definition extracted from the TLFi, which seems to be the most adapted. The LyText environment can help the pupil in his decision-making. The system will be able to propose assistance to the pupil like brighter information in the TLFi (the field to which the definition belongs, the indicators of time [old, out-of-date], the register of language [popular, standard...], etc.

On the designer level, the concern is to use the TLFi within the framework of automatic processing. Two levels of processing are currently under development in the module “Texts visualization”. First, it is a question of automatically

determining in each text the definitions taken from the TLFi, which seem to be the most adapted according to the context (for certain words defined by the teachers like difficult to understand) [Véronis & Ide, 1990; Péliissier & Jacquy 2004]. Finally, in the module "Texts visualization", the informational model offers to visualize the various lexical fields of the text. We are currently determining a system that would automatically extract lexical fields (love, sadness, death...) using all pieces of information contained in each input of TLFi (especially indicators and fields).

In February 2005, we have experimented LyText in a French classroom in Lunéville (54), France. Two teachers have accepted to use LyText with their pupils. On the one hand, pupils studying literature have used LyText as a means to revise a text; on the other hand, for pupils studying science and technique, LyText has been used as a tool allowing the discovery of a new text, which eventually has led to a text commentary. We can point out that the construction of different interfaces for these two uses of LyText is compulsory. Indeed, the approach of a text commentary does not imply the same functionalities than for the revision of text analysis for the oral test.

Conclusion

Designed and produced first as a paper dictionary, the TLFi is nowadays a computerized dictionary in which each user, linguist, and researcher or not, can find different pieces of information. Concerning primary schools, children can find information to understand words and expressions. In secondary schools, pupils could understand orthographic problems when looking for historical information. Then, for high schools pupils as we have seen, the TLFi is not quite adapted: the TLFi is included in the LyText environment. The latter totally uses the capacity of the TLFi. Indeed, to show the most appropriate definition of TLFi according to the context, the system must use all information (date, indicator of language and meta-language, etc.). This work is currently in progress.

References

- Bernard, P., Bernet, C., Dendien, J., Pierrel, J.-M., Souvay, G., Tucsnak, Z., 2001, "Un serveur de ressources informatisées via le Web", Actes de TALN-2001, Tours, Juillet 2001, pages 333-338.
- Bernard, P., Dendien, J., Lecomte, J., Pierrel, J.-M., 2002. "Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella", Actes des 8^e Journées Internationales d'Analyse Statistique des Données Textuelles JADT 2002, Saint-Malo 2002, pages 137-149.

- Bernard, P., Lecomte, J., Dendien, J., Pierrel, J.-M., 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).
- Bouty, M. (1992). *Dictionnaire des oeuvres et des thèmes de la littérature française*. Hachette-Education, 2002, Paris.
- Buzon, C., 1983, "Au sujet de quelques dictionnaires monolingues français en usage à l'école élémentaire", Christian Buzon, *Études de Linguistique Appliquée*, n°49, janvier-mars 1983, Didier Erudition, pp 147-173.
- CNRS (1976-1993), *Trésor de la langue française, dictionnaire de la langue du 19e et du 20e siècle*, CNRS, Gallimard, Paris.
- Fayol, M., Gaonac'h, D. (2003). "La compréhension, une approche de psychologie cognitive". In Gaonac'h, D., Fayol, M., coord., *Aider les élèves à comprendre, du texte au multimédia*, Profession Enseignant, Hachette, 2003, pp 5-72.
- Gross, G., 1986, "Reconnaissance des emplois à l'aide d'un dictionnaire électronique", *Études de Linguistique Appliquée*, Janvier-Juin 1992, n°85 - 86, Didier Erudition, pp 89-97.
- Gross, G. 1989, "Le dictionnaire et l'enseignement de la langue maternelle", Gaston Gross, *Ein internationales Handbuch zur Lexicographie / An International encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*, Hausmann Franz Josef, Reichmann Oskar, Wiegand Herbert Ernst et Zgusta Ladislav Hrsg., (1989-1991), « Wörterbücher / Dictionaries / Dictionnaires », volume 1, article n°22, Berlin / New York, Walter de Gruyter, 1989, pp 174-180.
- Lehmann, A., 2000, "Les dictionnaires pour enfants : diversité et uniformisation", *Le Français Aujourd'hui – Construire les compétences lexicales*, Alise Lehmann, revue trimestrielle, n° 131, revue de l'Association Française des Enseignants de Français, septembre 2000, pp 87-98.
- Ministère de l'Éducation Nationale, 2002, *Programmes Français, classe de seconde*, Collection Lycée, voie générale et technologique, CRDP, 2002, Paris.
- Nathan 1998, « Mon premier dictionnaire Super Génial Nathan », premiers apprentissages, plateformes Mac/PC, dès 3 ans, Pack « Nathan Benjamin Super Génial », Nathan, 1998.
- Pélissier, C., 2002 : "Analyse de Mon Premier Dictionnaire", revue ALSIC (Apprentissage des Langues et Systèmes d'Information et de Communication), rubrique Analyse de logiciel, décembre 2002, volume 5, numéro 2, pp 269-286. Internet : http://alsic.u-strasbg.fr/Num09/pelissier/alsic_n09-log1.htm.
- Pélissier, C., Jadelot, C., Pierrel, J.-M., 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.
- Pélissier, C., Jacquy, E., 2004, "Contribution d'un dictionnaire de référence informatisé dans un environnement didactique du français", Actes de la journée *TAL et apprentissage des langues*, Grenoble.
- Pierrel, J.-M., Dendien, J., Bernard, P., 2004, "Le TLFi ou Trésor de la langue française informatisé". EURALEX 2004, Lorient.
- Pruvost, J., 2001, "Les dictionnaires d'apprentissage monolingues et langue française (1856-1999) : problèmes et méthodes", Jean Pruvost, *Les dictionnaires de la langue française*, sous la direction de Jean Pruvost, Honoré Champion, Paris, 2001, pp 67-96.
- Véronis, J., Ide, N., 1990, "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries", Proceedings of the 14th International Conference on Computational Linguistics (COLING'90)