ATILF's computerized linguistic resources involved in cooperative projects

Jean-Marie Pierrel, Laurent Romary, Zina Tucsnak

ATILF's Computerized Linguistic Resources involved in Cooperative Projects

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

Laurent Romary
LORIA/INRIA
Campus Scientifique
BP 239
54506 Vandoeuvre-les-Nancy Cedex
Laurent.Romary@loria.fr

Zina Tucsnak
ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118 CNRS/Université de Nancy2)
BP 30687
54063 Nancy Cedex, France
Zina.Tucsnak@inalf.fr

Abstract

This paper presents some of the computerized linguistic resources of the Research Laboratory ATILF (Analyse et Traitement Informatique de la Langue Française) available on the Web. This considerable amount of resources concerning French language consists in a set of more than 3650 literary works grouped together in the textual database Frantext, plus a number of dictionaries, lexis and other databases. First, we try to place the resources in a wider perspective, next to identify special data representation and access problems. ATILF has great interest to stimulate re–usage of its resources by the NLP community. These computerized resources are involved in several national and international projects as Normalangue, INTERA, OLAC and ENABLER. The use of metadata to describe the available resources is a way of making those resources locatable and accessible. The metadata's visibility through some international harvesters will encourage collaborations and provide help for the industry, research and education. ATILF expects that the metadata description of the computerized resources will facilitate the access to local French language resources and increase the usage of the resources already available.

1 Introduction

In our Research Laboratory ATILF we started looking for a way to organize and describe the computerized resources already existent. All the ATILF repositories are accessible via the Web at http://www.inalf.fr/atilf for search and retrieval. Now we need to implement a metadata vocabulary and structure to enables others to locate our specific French language resources. The aim of our project is to improve the existent tagging toward an international standard as ISO/TC 37/SC 4 specifies. The challenge lies in the unification process of the variability of the existent tagging. In order to overcome all that diversity, our work relies on normalised XML formats. A future perspective: normalised interfaces for representation and access. In this article, we present some of the ATILF's textual resources and their metadata implementation for cooperative project as OLAC and INTERA. All tools and the meta-description files are available over the Internet via an HTTP server.

2 Metadata initiatives for ATILF's resources

Each textual database (cf § 4), in standalone, have his own graphical interface and search tool. But the amount of effort required by a user to locate our databases, understand how they are organized and perform a query is still very important. By using standard metadata ("data about data") we want to facilitate this process and open our resources to users other than their creators, colleagues and others already familiar with them. The answer to resource discovery is given by various initiatives that are currently developing standards for resource description, including the Dublin Core Metadata initiative, the Open Language Archives Community initiative and the International Standards in Language Engineering metadata initiative.

The OLAC linguistics specific metadata set and the harvesting protocol are very suitable to ATILF repositories. Trough a flat structure and a wide coverage of language resource type, the OLACMS facilitate the metadata description.

OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

ATILF joined OLAC in 2002. The first step was to build a local server http://olac.inalf.fr in order to implement the metadata OAI (Open Archive Protocol) harvesting protocol. Even if the resources keeps the original format, a standard Dublin Core and OLAC metadata description is added in order to have two big archives.

The first archive contain the metadata description of the Frantext textual database and the other one all the dictionaries end encyclopedia, especially the French computerized dictionary, TLFi. ATILF archives are harvested and registered for OLAC according OAI (Open Archive Protocol).

One can query ATILF archives by language, linguistic type, or linguistic field: http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search-advanced.html

Another inventory of existing metadata practice

in the language resource domain is the INTERA Metadata initiative. The IMDI metadata set is highly structured, have very detailed descriptions and accept OAI/OLAC interoperability. The TEI-OLAC-IMDI mapping is one of our challenges.

The INTERA (Integrated European language data Repository Area) projects has essentially two pillars: (1) It wants to build an integrated European language resource area by connecting international, national and regional data centers and (2) it wants to produce new multilingual language resources. The first goal implies the integration of a critical mass of language resources of different types with the help of metadata descriptions such that users can directly start suitable tools on the resources found. The second goal addresses the aspect that especially for lesser spoken languages there are no good multilingual resources available. ATILF joined INTERA as a member.

FRANTEXT's Bibliography	OLAC	IMDI
<EDI>	<publisher>	Session.Resources. Source.Access
<SIE>	<coverage>	Session.Location. Country
<PUB>	<rights>	Session.Resources. Source.Access
<TIT>	<title>	Session.Project.Title
<DAT>	<date>	Session.Resources. Annotation.Unit.Date

Figure 1

In figure 1 we present a mapping sample describing FRANTEXT bibliography (cf § 4.1) in OLAC and IMDI metadata.

3 Implementing metadata on French language resources

Exploring multiple ATILF resources in correct, effective and efficient use is a challenging task. We are using various terminology, notation, organisation of data and search commands. The unified representation of the metadata allows speedy searches, using finally the existing tools. The core component of our metadata project is a database server that acts as intermediary between the metadata representation and the virtual harvester. All the computerised resources reside in separate, remote servers. We use XML encoding and TEI specifications.

The resources presented below are searchable and browsable French repositories across metadata specification, interlinked with existing relevant search tool in order to facilitate users' access .

3.1 First example: FRANTEXT

Frantext is a corpus of written literary French texts from 16th to 20th centuries: 3665 literary works. The textual database is covering a period dating from 1507 to 1998. In this corpus, subject to regular updating and enhancing, the proposed texts are errors less and the captured data editions are very accurate. On a subset of the corpus, morphologically annotated (1940 texts), you can do interrogation on the graphic forms, and/or on morphological tags, either independently, or in the same request. There are multiple applications to various areas of linguistic research: sub-lexicon extraction, morphological studies, local syntax and recurrent syntactic patterns, semantic and stylistic

studies, corpus tagging and evaluation procedures.

Historically, the base was constituted in order to provide samples to be used in the elaboration of the TLF (Trésor de la Langue Française). It was started in the 60's.

At the present time, 3665 literary works are grouped together in Frantext. They cover a period dating from 1507 to 1998, plus 240 texts dating from 842 to 1502. This corpus is subject to regular updating and enhancing, with several objectives: good text and edition quality , enlargement of the database with new texts, in order to restore the balance between dates, or genres, or to facilitate some special operations in the domain of linguistic research or teaching. These different texts use equivalent or near-equivalent tagging. It is thus necessarily to establish tagging guidelines and to normalize the output toward TEI (Text Encoding Initiative) standards. In figure 2 one can see some mapping examples from the existing mapping to the TEI standard.

FRANTEXT	TEI
>R> >R/> >L01> >L01/>	<sourceDesc></sourceDesc>
>JP> >JP/>	<pb n="x"/>
>J> >J/>	 <head>x</head> <head>x</head>
>JN> >JN/>	<speaker> </speaker>
>JC> >JC/>	<cit> <quote>… </quote> <bibl>…</bibl> </cit>
>Z>XIV	<num>XIV</num>
>...>	<gap/>

Figure 2

We create a new ATILF production line for XML texts (see figure 3). The original paper material is digitised by efficient means. Next, two OCR engines recognize and export texts in ASCII format. After that, special automated methods are applied. Finally, a human correction, a check with the original book, a spellchecking and a final XML tagging is performed. The error average is less than: 1error for 20000 characters for modern editions.

Any user can access the two versions of

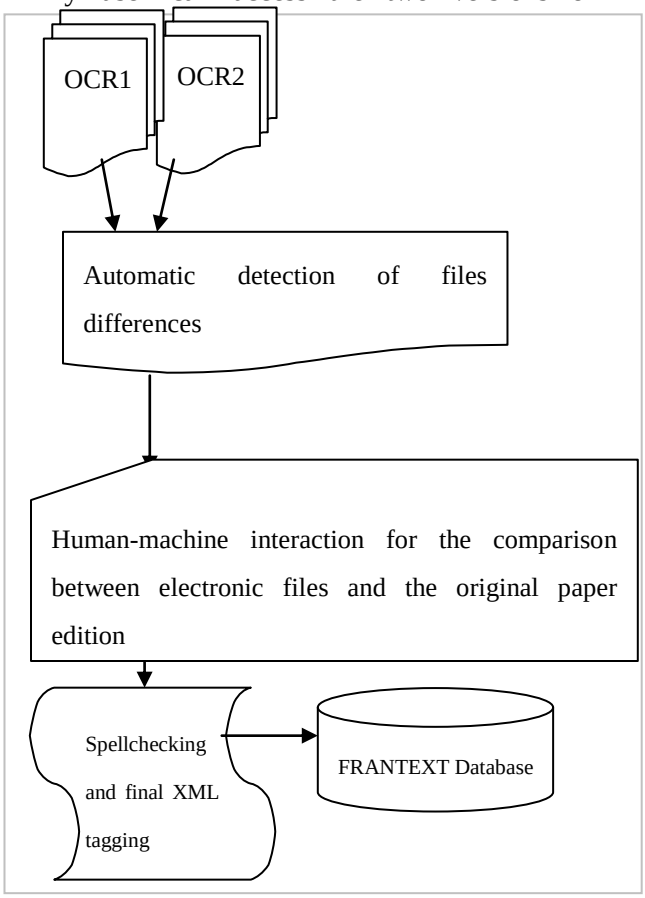

Figure 3

Frantext (http://atilf.inalf.fr/frantext.htm) through the search engine called STELLA (Dendien, 1991) (an annual subscription of 305€ is needed)

First, the total base, containing at present 3900 texts (raw text, not annotated, for a total amount of more than 220 million occurrences). Texts can be interrogated on the graphic forms of the words (all texts at the same time, or one by one). Second, a sub-base of the corpus, in "modern" spelling, containing 1940 texts (about 127 million occurrences), called "Frantext catégorisé". This corpus is morphologically annotated with Part-of-Speech labels, with a specific ATILF categorizer. Interrogation is possible on the graphic forms, and/or on the morphological tags, either independently, or in the same request.

Once the work corpus (all the texts at the same time, or by author, or by dates, etc.) is defined, several requests are possible from FRANTEXT interface: searching simple occurrences (words, tags), co-occurrences, or sequences (possibly including optional terms), using simple graphic forms, word lists, or grammars.

3.2 Second example : TLFi (Trésor de la Langue Française Informatisé)

The TLFi (obvious result of the TLF started in the 60's) can be seen as a lexical database (Dendien, 1996). It contains: about 100000 words with 270000 definitions, special sections concerning their history, formation, etymology, more than 430000 examples from the last two centuries literature. The TLFi allow full text requests throughout its whole content; it contains also a hierarchy between the textual objects, using a special internal tag set and a special control grammar. It is operated by the software STELLA (Dendien, 1991), a toolbox for the textual resources advance search. The whole dictionary has been transformed into an XML document. One can count almost 36 millions XML tags. A new user's interface allows very pertinent results.

The TLFi can be freely consulted via internet, at http://www.inalf.fr/tlfi. A user can simply consult the dictionary, article after article, putting or not into evidence such or such type of information (a definition, an author, etc.). He has a possibility to use a formulary of "aided request", i.e. consult the dictionary in a simple way (asking for a *definition*, a *domain* or another proposed "*object*"), or in a transverse way (crossing the criteria , for example : a *definition* in the *domain* of…). A third way of consulting the TLFi is to use a more complex request crossing several criteria and taking into account the hierarchical structure of the textual objects. This request can be single- or multi-objects. It is possible to make and use word lists.

In the TLFi, a user can make requests according to other types of criteria: domains, grammatical tags (POS), and other types of textual objects: definitions, examples, etc. For example, he can extract lists of proverbs, lists of words of a specific domain, etc. Other type of studies:

- Morphological studies: By the way of word lists, a user can make requests about morphological phenomena, in Frantext as well as in the TLFi. Examples of derivation or composition phenomena, of multi-words containing a hyphen can be found in both resources. For example, in the TLFi, he can obtain the list of all transitive verbs ending with *–er* or *–oir*, or all adjectives ending with *–esque* or all verbs beginning with prefix *re-* or *dé-*.

- Semantic and Stylistic studies: If a user chooses to ask for the list of all adjective ending with *–esque,* present in the TLFi, it is perhaps because he wants to know which ones are *pejorative* or not…The request can be established, using this kind of criteria. He can also make requests specific to an author of the many examples cited in the TLFi with their reference, and possibly ask which examples from Balzac contain a conjugated form of the verb *aimer*, not in the source of the example, but in the core of the example. This is only possible because of the hierarchical structure of the textual objects in the article. Frantext can be used by teachers and students in order to detect the nuances of a word, by looking at its environment. It is also possible to concentrate on the evolution of particular semantic fields, on the evolution of a word between its first attestation in the base and the last one, on problems linked to synonymy: are "synonyms" really "synonyms" (for example *suspicion* and *soupcon*, or *espoir* and *espérance*, etc.), and hence, what is "synonymy"?

3.3 Last example: ancient dictionaries and Encyclopaedia

The following resources are all integrated in OLAC archives and in INTERA project. They are searchable and browsable through the search engine Philologic (Olsen, 1999).

The ***Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres*** was published under the direction of Diderot, with 17 volumes of text and 11 volumes of plates between 1751 and 1772. Containing 72,000 articles written by more than 140 contributors, the **Encyclopédie** was a massive reference work for the arts and sciences, as well as a **machine de guerre** which served to propagate Enlightened ideas. The impact of the **Encyclopédie** was enormous. Through its attempt to classify learning and to open all domains of human activity to its readers, the **Encyclopédie** gave expression to many of the most important intellectual and social developments of its time. The current electronic version implements a fast, flexible search engine. It is a cooperative project of ATILF and ARTFL (American and French Research of the French Language). The **Encyclopédie** (http://encyclopedie.inalf.fr) database contains 20.8 million words, 400000 unique forms, 18000 pages of text, 17 volumes of articles, and 11 volumes of plate legends. One can do: searches for articles or groups of articles (by : article headword, author, category of knowledge, part of speech), full text searches (for : words, phrase, co-occurrences of words, word or phrase frequency per article) , cross-references from one article to another, links from plate legends to plate images, links to digitized images of each page of text. The resource have a restricted access (an institution can subscribe to a database package: FRANTEXT and Encyclopédie Project).

The third edition (1552) of Robert Estienne's *Dictionarium latinogallicum* marks the culmination of his work on the Latin-French dictionary. Estienne, the father of modern Classical Latin and French lexicography (cf. Starnes 1954), had established for Classical Latin and 16th-century French a tripartite series of dictionaries: for Latin a monolingual Thesaurus intended for scholars, and two Latin-French dictionaries, one, the *Dictionarium latinogallicum*, for advanced students, and an abridged version, the *Dictionariolum puerorum latinogallicum*, for beginners.

Jean Nicot's *Thresor de la langue françoyse* is

the key to the development of French lexicography. The sum of four editions of Robert Estienne's bilingual *Dictionaire françois-latin*, the ***Threysor*** assumed, through the contributions of Nicot, the nature of a monolingual French dictionary.

Pierre Bayle's *Dictionnaire historique et critique* stands as the supreme achievement of one of the seventeenth century's most prominent men of letters. The database contains a facsimile version of the 1740 edition.

Theses resources are access free at http://dictionnaire.inalf.fr/dictionnaires . Users can handle joint query on all that computerised dictionaries.

The ***Académie française*** was founded by Cardinal Richelieu in 1635 with the primary goal of creating a French dictionary. A total of eight editions have been published in the years since its foundation, from the first edition in 1694 up until the eighth edition in 1935. The ninth edition is currently in progress. Users can freely search either on the electronic versions of the first (1694), 5th (1798), and 6th (1835) editions or on the 8th (1932-35) and the 9th.

French Academy dictionary 8th edition and **French Academy dictionary 9th edition (1st volume, 1992—from A to Mappemonde)** support hyper navigation from Frantext or the TLFi.

4 Future work

Future during next steps of INTERA project and RNIL project should include:

- Defining simple schemas for specific metadata elements and vocabulary.
- Develop a complete mapping between IMDI and OLAC metadata standards with respect to ATILF resources and TEI standards.
- Promote development of standards in language resource management.

5 Conclusion

Frantext, the TLFi as well as all computerized ATILF's resources, do not compose a closed set of textual resources. They can be independently consulted and interrogated, and can be a starting point to a number of linguistic projects. The objective of the laboratory ATILF is to let the community know that such resources exist for French Language. These resources have been initiated a long time ago and are today available at ATILF, where work is still in progress. In order not to exclude anyone of the process of distributing these tools, we propose a mutualisation of these resources to the benefit of the entire community. Networks and infrastructure projects as Norma langue (in France), INTERA or OLAC are the best ways to inform a large group of language resource users of the potential of the ATILF resources. The general policy of our laboratory is to welcome and give the research and teaching world the widest access to all our resources.

Another objective is to implement several International standards and other Technical Reports that cover computer-assisted lexicography and language engineering trough specific annotated information (morphology and syntax).

References

Andreev, L., M. Olsen, 1999. Conception de systèmes hypermedia à grande échelle pour les sciences humaines ; Présentation de Philologic : le logiciel d' ARTFL, in *67th Congrès de l'ACFAS (Association canadienne française pour l'avancement des sciences), May 11, 1999, University of Ottawa*.

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

Bonhomme, P., 2000. Codage et normalisation de ressources textuelles, in (*Pierrel, 2000*).

Dendien, J., 1991. Access to information in a textual database : access functions and optimal indexes, in *Research in Humanities Computing, Papers from the 1989 ACH-ALLC*, Oxford: Clarendon Press.

Dendien, J., 1996. Le projet d'informatisation du TLF, in *Lexicographie et Informatique*, pp. 25--34.

Dublin core metadata initaitive. Web site, http://dublincore.org/

INTERA (Integrated European language data Repository Area). Web site, http://www.mpi.nl/world/ISLE/index.html

ISO/TC 37/SC4, Web site, http://www.tc37sc4.org/

Open language archives community (OLAC). Web site,

http://www.language-archives.org/

Pierrel, J.M., 2000. *Ingéniérie des Langues*, Paris: Editions Hermès.

TEI (Text Encoding Intiative), Web site : http://www.tei-c.org/ .

Véronis, J., 2000. Annotation automatique de corpus : panorama et état de la technique, in (*Pierrel, 2000*).