The PAROLE Program

Georges Vignaux

Institut National de la Langue Française et Laboratoire Communication et Politique, CNRS, Paris
vignaux@poleia.lip6.fr

Abstract

The PAROLE project (Contract LE2-4017) was launched in May 1996 by the Commission of the European Communities, at the initiative of the DG XIII (Telecommunications, Information Market and Exploitation of Research). PAROLE is not just a project for gathering and creating a corpus. We are creating a true architectural model whose riches and quality will constitute strategic assets for European linguistic studies. This two-level architecture will link together two major morphological and syntactical component.

1. Introduction: Language Processing Systems

The market for products and services related to language processing systems is continually growing and reaching increasingly large numbers of people. Most recently, this includes households starting to buy microcomputers as new tools to help their children with their studies. This popularity means that effective natural language processing systems must be developed, which cannot be done without language resources. We are seeing an inevitable expansion in linguistic engineering's responsibilities and prospects. In France, the market for linguistic engineering is currently estimated to be worth 200 million francs. Upwards of 80% of this figure is made up of indexing and documentation research, while about 15% of the market is for for computer-assisted translation. (P. Oudart, in: *Rapport du Conseil consultatif sur le traitement informatique du langage*, 1995-1997, p.88). This pales in comparison to the market for multimedia in the United States, but we can expect, in the years to come, an explosion in the availability of various types of products, such as terminology and knowledge bases, general and specialized electronic dictionaries, text bases, and electronic technical documents.

Before any of these products can be created, basic studies must be done in the areas of pure linguistics (such as extensive description of the language and its evolution and research on formal structures) and in computational linguistics (models and control algorithms). This type of study relies on large corpora, both raw (corpora for training and validation of hypotheses and research models) and tagged (i.e., annotated).

2. The PAROLE Project

The PAROLE project (Contract LE2-4017) was launched in May 1996 by the Commission of the European Communities, at the initiative of the DG XIII (Telecommunications, Information Market and Exploitation of Research). The PAROLE project brings together fifteen partners. The European partners are: Consorzio Pisa Ricerche (Italy, project coordinator), Centro di Linguistica da Universidade de Lisboa (Portugal), Det Danske Sprog - og Litteraturselskab (Denmark), Fundacion Bosch Gimpera Universitat de Barcelona (Spain), Goeteborgs Universitet, Dept of Swedisch Sprakdata (Sweden), Institiuid Teangeolaiochta Eireann (Ireland), Institut d'Estudis Catalans (Spain), Institut Sprache (Germany), Institute for Language and Speech Processing (Greece), Instituut voor Nederlandse Lexicologie (Netherlands) University of Birmingham (United Kingdom) University of Helsinki (Finland), Université de Liège (Belgium). The French partners are two: GSI-ERLI and l'Institut National de la Langue Française (INaLF: coordinators: G. Vignaux, DR CNRS, and B. Habert, MC-ENS Fontenay).

Each partner is responsible for creating in its own language a corpus of at least 20 million words, divided up as follows:

- 60 % from newspapers: 12 million
- 20% from books: 6 million
- 10 % from periodicals: 2 million
- 10 % from other sources: 2 million.

No more than 20% of these occurrences can date from prior de 1980!

This corpus of data structured according to current standard formats (SGML) is to be completed by two sub-corpora: one consists of 250,000 forms tagged with a morphosyntactic description (according to the part of speech) plus another 50,000, for which all syntactic and semantic positions will be verified. When completed, all this information is to be made available on an Internet site permitting researchers to use different processing methods. A contract for the preparation and distribution of these resources is being signed between the NaLF and the European Language Resources - Distribution Agency (ELDA), a key partner of the PAROLE project.

The stakes are high: the corpus will be one of the largest French language corpora; it will be not only recent and diversified but also available for study without the usual copyright barriers. With respect to the data from the press, an agreement was signed in December of 1997 between the InaLF-CNRS and the European Language Resources - Distribution Agency (ELDA) by way of which ELDA will make available to the InaLF for a specific period of time a corpus of five years of the newspaper *Le Monde*, subject to use solely for linguistic purposes and within the strict framework of the PAROLE project. This is the first time that a large national daily has been associated with a project of this magnitude.

3. General Scope of PAROLE

The general scope of PAROLE is highly significant. We have seen that it consists of corpora representing thirteen major European languages based on a common generic lexicon model. In the twelve European countries involved, substantial amounts of resources will be made available. The objective is to increase European industrial competitiveness in the large language industry sectors and, by so doing, to eliminate the danger of linguistic exclusion of any of the Member States. The success of European integration largely depends on our ability to master the problems of multilingual communication. Preservation of the the cultural diversity of the European nations is a legitimate and necessary aim; the survival of their national languages is of the utmost importance. If the linguistic resources conducive to research, experimentation, and evaluation are not brought together, we will not be able to put into place viable systems for acquiring and transferring knowledge. This would have grave consequences for European industry and public services.

So that this aim may be accomplished, PAROLE is not just a project for gathering and creating a corpus. We are creating a true architectural model whose riches and quality will constitute strategic assets for European linguistic studies. This two-level architecture will link together two major morphological and syntactical components. At the morphological level, various types of relations underlie the lexical entries: spelling, derivations, inflexions, affixes, internal components. The syntactical level takes into account the major morphosyntactical categories: verbs, categories and sub-categories, objects, pronouns, ordering constraints.

4. Technical Specifications of the Project

The work carried out by the team coordinated by G. Vignaux (CNRS/InaLF) and B. Habert (ENS Fontenay-St-Cloud) included:

- recovering in electronic form all the primary data mentioned above;

- transferring these very different documents (various media: diskette, CD, ZIP disk; different operating systems: Mac, Windows, MS-DOS; different data entry software: various versions of Word, WordPerfect, DTP software such as Quark Xpress) onto a single architecture;

- developing a strategy for recovering information on the physical layout of the source document in order to transform it into logical information (text structure, etc.);

- formatting PAROLE;

- adding identifying information to each document (for the primary document and the changes made);

- using SGML verification to check that each document complies with the PAROLE format;

- organizing the complete corpus as a coherent whole and providing a CD archive.

The objective at the outset was to create in one’s own language a corpus of at least 20 million words, divided up as follows:

- 60 % from newspapers: 12 million
- 10% from various sources: 2 million
- 10 % from periodicals: 2 million
- 20 % from books: 4 million.

The actual figures integrated from the various categories are as follows:

Various : Data from ELRA (CRATER, MLCC Multilingual, MLCC Parallel)	2 025 964 words
Books (CNRS Editions)	3 267 409 words
Periodicals (CNRS Info, Hermès)	942 963 words
Press (*Le Monde*, from ELRA)	13 856 763 words
Total	20 093 099 words

4.1. Information from the Press: 14 million words and 250,000 tagged words:

14 million words randomly selected from complete issues of the newspaper *Le Monde* for the years 1987, 1989, 1991, 1993(?), and 1995 make up the *Press* section of the corpus produced within the framework of the project. Benoît Habert and I will call this corpus LeMondeNu.

241,484 words, taken from seven issues of Le Monde from September of 1987, were extracted from LeMondeNu, automatically tagged and manually corrected for the part of speech. We will call this sub-corpus LeMondePOS.

In LeMondeNu and LeMondePOS, each article forms a header (Dunlop D. (1995), “Practical considerations in the use of TEI headers in large corpora”, Computers and the Humanities, 29, 85-98—following the proposals of the TEI (Text Encoding Initiative). The identifying fields provided by the documentation from Le Monde have been changed into classification categories in the headers. This way it is possible to extract articles from different headings or categories; it is the consistency of the main headings that is examined. The table below indicates the average size of the articles for each of the categories used.

Headings	*articles*		*words*		*average*	
ETR = foreign	5,464	149	2,366,055	77,347	433	579
ECO(nomy)	3,478	108	1,443,923	38,540	415	356
POL(itics)	2,305	83	1,326,576	36,703	575	442
ING(gen. info)	0,838	80	364,590	34,284	435	427
ART(media)	2,261	76	1,080,620	37,220	477	489
EMS (Educ, Med, Soc)	1,092	42	457,626	17,390	419	414
TOTAL	15,438	538	7,039,390	241,484	455	448

4.2. Data from Periodicals and Books

The aim at the outset was to form a corpus of data:

- taken from periodicals for a total of 800,000 words;

- taken from books for a total of 3.2 million words.

4.2.1. Periodicals

Periodicals come from two sources: from the CNRS magazine “*Hermès*” and from the electronic magazine *CNRS-Infos.*

HERMES

Source: Seven issues of the magazine were made available to us: issues 15 to 22. The data was provided on diskette, one Word format file per article (for a total of 134 files).

Working format: HTML (Hypertext Markup Language) was chosen as the working format because it is the most structured format accessible via export from Word. We had two possibilities:

- direct conversion into HTML
- conversion into RTF (Rich Text Format) then, via a converter, to HTML.

We opted for the second solution since the result is “cleaner”, that is, easier to process.

Conversion: The conversion of HTML files to the PAROLE format was carried out using programs written with flex. Since the structure of the various files was relatively stable, the different versions (seven in all) essentially dealt with title structure.

This result of this procedure for each article:

- a header file containing information on the author and identifying the article
- a body file containing the article itself in PAROLE format

Finally a perl script creates the final file from the header and body files.

CNRS-Infos

Source: The data extracted from CNRS-Infos come from the magazine’s web site (http://www.cnrs.fr/Cnrspresse/cnrsinfo.html). From our studies, it would appear that the HTML data from the CNRS server conforms to nine different structures.

Working format: The data was directly usable because it was in HTML format.

Conversion: The first step was to recover the various articles from the web site.

Then each file was processed as follows:

-cleanup of the HTML header
- extraction of the summary
- cleanup of the HTML markups
- conversion to the PAROLE format
- creation of the header and body files (as for Hermès)

As with Hermès, a *perl* script creates the final file from the header and body files.

For periodicals, the average number of words is 942,963.

4.2.2. Books

Source: All the books we processed were originally provided on CD-ROM in Xpress format. Each book has its own structure!

Working format: Xpress allows conversion to “Xpress markup” format. Although this is a non-standard format, it becomes a good working base after standardization and cleanup. This step is the same for all the data. With the “Xpress markup” format, if the Xpress file has been correctly laid out (which is not always the case), the different structures of a book can be detected.

Conversion: For each book, the structure had to be studied in order to come up with the *perl* script allowing conversion to the PAROLE format.

For books, the total number of words was 3,267,409.

Verification: NSGMLS tools were used to ensure conformity with the PAROLE format. The errors discovered during verification were corrected by hand in the file.

The task was complicated by the fact that, practically speaking, each document presented unique features that prevented us from defining hard and fast rules for standard transformation rules.