

# For a repository of NLP tools

Stéphane Chaudiron\*, Khalid Choukri<sup>†</sup>, Audrey Mance<sup>†</sup>, Valérie Mapelli<sup>†</sup>

\*Ministère de la Recherche & Université de Paris 10 - CRIS  
200, avenue de la République 92001 Nanterre cedex, France  
stephane.chaudiron@u-paris10.fr

<sup>†</sup>ELRA/ELDA  
55-57, rue Brillat-Savarin, 75013 Paris, France  
{choukri, mance, mapelli}@elda.fr

## Abstract

In this paper, we assume that the perspective which consists of identifying the NLP supply according to its different uses gives a general and efficient framework to understand the existing technological and industrial offer in a user-oriented approach. The main feature of this approach is to analyse how a specific technical product is really used by the users and not only to highlight how the developers expect the product to be used. To achieve this goal with NLP products, we first need to have a clear and quasi-exhaustive picture of the technical and industrial supply. During the 1998-1999 period, the European Language Resources Association (ELRA) conducted a study funded by the French Ministry of Research and Higher Education to produce a directory of language engineering tools and resources for French. In this paper, we present the main results of the study. The first part gives some information on the methodology adopted to conduct the study, the second part presents the main characteristics of the classification and the third part gives an overview of the applications which have been identified.

## 1. Introduction

Usual classifications of natural language processing products and services break down the field according to application domains. Different established directories (Ofil, 1994; Mlis, 1999), market studies (Bossard, 1988; Ink, 1989; Owil, 1990; Ovum, 1991), or study reports for national governments and the European commission (Ovum, 1993; Euromap, 1998) distinguish 5 or more major applications which may benefit to users for increasing productivity in information processing.

For example, (Bossard, 1988) identified 6 categories of products while (Ovum, 1991) pointed out five major applications of NLP which are Text editing, Database interfaces, Machine translation, Content scanning and Talkwriters.

These studies are mainly conducted with an underlying philosophy which aims at filling the gap between research and industry, increasing economic efficiency within the European Community and identifying the measures required at government or supra-government levels to achieve these potential economic benefits.

Some of these studies try to identify the different perspectives, the user demand, the industrial supply and the technological offer, but none of them addresses the problem in terms of *contexts of use*. During the 1998-1999 period, the European Language Resources Association (ELRA) conducted a study funded by the French Ministry of Research and Higher Education to produce a directory of language engineering tools and resources for French.

Within this study, we assume that such a perspective, which consists in identifying the NLP supply according to its different uses, gives a general and efficient framework to better understand the existing technological and industrial offer in a user-oriented approach, even if we devote some sections to list some of the tools and the corresponding typology identified during the survey.

The main feature of the user-oriented approach is to analyse how a specific technical product is really used by the users and not only to point out how the developers expect the product to be used. But, to

achieve this goal with NLP products, we first need to have a clear and complete inventory of the technical and industrial supply.

We can therefore define the term "context of use" as referring to the social and individual appropriation of a technical object, Perriault (1989) and Harvey (1995) give some examples of this approach. Concerning NLP applications, this methodological approach gives results which are the first step for a future user-oriented evaluation.

In this paper, we present the main results of the study<sup>1</sup> according to a classification based on the contexts of use. The first part gives some information on the methodology adopted to conduct the study; the second part presents the main characteristics of the classification and the third part gives an overview of the applications which have been identified.

## 2. Methodological features of the study

The objective of the ELRA study was to identify the information processing tools that were elaborated by industry or research laboratories for the French language. Possible extensions to other languages are under discussion with some partners. In order to achieve this study, ELRA followed the following steps:

- Definition of a classification for the tool typology;
- Identification of potential producers and tools;
- Designing a tool description form;
- Drafting a questionnaire;
- Carrying out the survey;
- Analysis of the information collected;
- Structuring the final information set as a database.

### 2.1 Typology

This first task consisted of listing the different categories of tools for the Natural Language Processing (NLP) field, also trying to determine the relations between these categories. The main focus point was to define whether a tool indicated as an NLP tool did or

---

<sup>1</sup> A complete published version is under press and will be available soon.

did not include a language component. Therefore, we considered to be NLP tools either computerised tools which process language (e.g. analysis or translation systems) or tools that use language knowledge to process information.

In this study, a distinction had to be made between "language resources", such as corpora (speech or text), electronic dictionaries, glossaries, grammars, etc., and natural language processing tools, which allow to analyse, generate, understand, evaluate, extract, translate, etc. all kinds of information.

Finally, we established a list allowing to distinguish the different categories of NLP tools, including a special part concerning language resources. We chose not to make a hierarchical list of tools since it was too difficult to settle on the relationship between the tools. Therefore, we opted for a linear presentation of the tool categories. The main top categories that could be distinguished are the following: language resources, language analysis, automatic generation, automatic translation, automatic summarisation, language understanding systems, terminology management, speech processing, information management and retrieval, computer-aided authoring tools, optical character recognition, computer-aided learning, system and resource building, NLP systems evaluation.

## 2.2 Prospect list

Preparing the list of prospects was not an easy task. They were namely extracted from the ELRA contact database, which consists of more than 1200 contacts all over the world and over 300 for France; the list of organisations members of the Aupelf-Uref, now known as AUF (Agence Universitaire de la Francophonie) (1998); a few directories were provided by the French Ministry of Research and Higher Education; *The Language Engineering Directory* (Hearn, 1996) was also used.

More information was also extracted directly from several Web sites.

The use of these different sources allowed us to collect a good amount of information about contact persons, available tools, organisation's profile, etc. That also helped to carry out a targeted survey.

## 2.3 Questionnaire

In order to contact the different players in the NLP field, we decided to draft a specific questionnaire. This questionnaire consisted of five parts:

1. Tool identification: this part includes the minimum information required to identify a tool, in particular the tool name, type of tool (is the tool a language resource, an application, or a software?), tool category (according to the typology that we defined and that was given as an annex to the questionnaire), usage (potential users), availability, language(s).

2. Provider identification: this part includes all necessary information concerning the provider, i.e. organisation's name, contact, address, etc.

3. Technical and commercial information: this section requires information about the medium, the size of the data, the workstation, documentation available, constraints for distribution, etc.

4. Detailed description: this blank section (limited to a maximum of 3 complete pages) helps provide extra technical and linguistic information.

5. Free description: a blank field was added to help the prospects add more details, such as related information, bibliography, information sources, other existing tools that they were aware of.

## 2.4 Survey

Once the contact list had been completed, we decided to directly contact each player that we had identified by sending them the description form. We only completed the contact section so that the contacted person could freely fill in each part of the form.

## 2.5 A catalogue of tools

As soon as the description forms were collected (either under Word format or as hardcopies), all information was gathered in a specific database under Access-97. Beyond the description forms that we managed to collect, we decided to add other tools that were found during the study. This directory of language processing tools for the French language may be completed later on with a list of tools from French speaking countries that will be carried out by the FRANCIL network (AUF). Other studies might also be carried out in the framework of the European HLT programmes with the co-operation of the DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz, University of Saarbrücken).

## 2.6 Similar studies

Several studies were carried out in the field of language processing tools. These can be divided into two different areas:

### 2.6.1 French speaking organisations' studies

Other interesting information can be found at four main French speaking organisations: the OFIL (Office Français des Industries de la Langue, France) which published the *Guide des produits et services d'ingénierie linguistique* (language engineering products and services guide); the DGLF (Délégation Générale à la Langue Française, France) has a Web site that includes a directory of French firms and research centres in the language engineering field (<http://www.culture.fr>); the RIOFIL (Réseau International des Observatoires Francophones des Industries de la Langue, Québec) which enquired about the language resources offer and the needs from the natural language processing players' point of view; the FRANCIL network (Réseau FRANCophone de l'Ingénierie de la Langue, AUF) which is working to expand the ELRA tool directory, that focused on French organisations, to the French speaking countries.

### 2.6.2 European studies

The European Language resources Distribution Agency (ELDA) offers on its Web site (<http://www.elda.fr>) a catalogue of language resources including different types of language resources: speech and written corpora, monolingual and multilingual lexica and terminology databases. Language and

Technology (Spain) constituted a language engineering directory, which consists of a thousand language tools and resources and over 600 language engineering organisations, on behalf of the European Community within the framework of the MLIS programme (Multilingual Information Society - CE- DGXIII). This directory is also available on the Web (<http://www2.echo.lu/mlis/fr/direct/home.html>). The DFKI is currently working on a tool directory, namely the *Natural Language Software Registry* (<http://corp-200.dfki.uni-sb.de/lt/registry>).

### 3. Three Classes of NLP applications

In this paper, we use the term "application" as a generic term designing all kinds of automatic processing, including NLP techniques or technologies. A NLP product aims at extracting information from linguistic data (as input of the system). The output of the system may be either the result of a linguistic transformation (translation, filtering, information retrieval,...), or a state transformation of a complex system (NL interface with a security system for nuclear central, voice control of a fighter plane,...).

Considering the use of NLP for professional information managing, we can distinguish three different classes of products and services. The low level, or *basic tools* which produce weak added-value to the information processing; the *linguistic agents*, or *lingaware*, which both address linguistic and informational tasks and the *integrated applications* which use linguistic technologies at different steps of the information processing.

#### 3.1.1 Basic tools

This first class concerns the basic components which implement very precise and limited linguistic function as lemmatisation tools, morpho-syntactic and/or semantic analysers, terminology extractors, etc. The linguistic transformations may use complex technologies but these modules do not create any informational added-value, or very little if any.

In this sense, the output of the system is not the result of information processing but is the result of a linguistic data processing. These components are developer-oriented and are used by software engineers.

#### 3.1.2 Linguistic agents

*Linguistic agents* are off-the-shelf software or modules which achieve complex informational tasks (help to text-editing, text translation, query translation, filtering,...). Off-the-shelf products are user-oriented whereas modules are developer-oriented.

By their own or integrated in an information processing platform, they not only achieve linguistic data processing, but a process of information management. For example, translation is not only considered as a single process of homothetic transfer between phrases but fits into a context of information exchange. Information retrieval or information filtering and routing may use linguistic technologies and, if they do, create a high added-value in the context of information processing.

At the present time, linguistic agents are the most important technological and commercial supply source.

### 3.1.3 Integrated applications

The third class of applications concerns what we call *integrated applications*. These applications do not only achieve specific linguistic and informational functions as linguistic agents but process more complex tasks for information processing, or even knowledge management. It concerns for example a process of strategic information intelligence in a multilingual environment using several linguistic agents to perform the different tasks at the various steps: search engine, data extraction, translation tool, information filtering and routing for example. Integrated applications process the "information content" using language knowledge.

## 4. Study outcome: a directory of language processing tools

### 4.1 Some figures

Most of the contacts were from commercial organisations (about 62%) compared with research laboratories (38% of responses). Only 83 out of 253 organisations answered by returning the questionnaire duly completed (about 33% response rate), of which there were 47 commercial organisations and 36 research organisations. A total of 161 questionnaires were collected, out of which 143 that described tools and 18 that described language resources. Among the commercial organisations that answered the survey, we could distinguish between SMEs, large French groups, French subsidiaries of international groups, from different activity fields, such as computational linguistics, research and development, information management, speech technologies, software distribution. As for the academic organisations, those were mainly from the written field.

The table below gives the number of tools identified within the study according to their top category. A total of 261 tools and resources were identified within the survey (240 tools and 21 language resources). In addition, ELRA collected information on tools for which no questionnaire was completed. These appear in our survey under the label "Tools identified by ELRA". Moreover, some tools belong to several domains of the table but are counted as one single tool in our directory (such as COATIS which is both a terminology consolidation tool and a semantic analyser). Table 1 summarises the findings and results in a total of 283 tools.

Tool category	Completed forms	Tools identified by ELRA	Total
1. Language resources	19	3	22
2. Language analysis	19	9	28
3. Automatic generation	4	5	9
4. Machine translation	12	10	22
5. Automatic summarisation	2	2	4
6. Language understanding	3	0	3
7. Terminology management	16	4	20
8. Speech processing	23	18	41
9. Information management and	39	6	45

retrieval			
10. Computer-aided authoring	9	8	17
11. Optical character recognition	7	2	9
12. Computer-aided language learning (CALL)	12	26	38
13. System and resource building	15	3	18
14. NLP systems evaluation	1	0	1
15. Other tools	4	2	6
TOTAL	185	98	283

Table 1. Number of identified tools

In order to offer an as exhaustive directory as possible, ELRA completed the information sent by the contacted persons with some extra information mainly gathered from the Web.

## 4.2 List of tools

ELRA's study led to the creation of a tool directory where tools were classified according to the typology defined within the survey. Examples of identified tools, ranked according to their top category, are given below.

### 4.2.1 Language analysers

Analysers are basic modules for most language processing systems. The identified tools were ranked according to different levels of analysis. Among existing morphological and morpho-syntactical analysers we can quote the *CRISTAL morphological analyser* (GRESEC, Grenoble 3), the *AMFLEX lemmatiser* (IRIT, Toulouse 3), analysers from CIMOS and Xerox, *MAUD* (LORIA, Nancy), *Labelgram* (CRLT), *EtiWeb* (LIA, Avignon). As for syntactical analysers, these come mainly from CORA, Xerox, LIMSI. Identified semantic or pragmatic analysers come also from LIMSI and Xerox and, notably, the Tropes semantic analysis software from Acetic.

### 4.2.2 Automatic generation systems

These systems allow the production of textual data in a natural language form. They are mainly used in translation and summarisation systems. We opted for two categories of generation tools: morphological generation and text generation. Examples of morphological generation tools are *Lemma le fléchisseur* (CORA), an inflected form generation tool (IGM, Marne-la-Vallée), and conjugation systems from CIMOS. Autonomous text generation systems are few and are often integrated into other systems like translation systems. These are *Flaubert* (CORA), *CRISTAL generator* (GRESEC).

### 4.2.3 Machine translation systems

Machine translation can be used in various applications: translation of technical documents, multilingual information processing, information retrieval, etc. These systems can be classified into two main categories: (i) machine translation tools, that automatically generate a text in a target language

(*Systran* or *Reverse Pro* (Softissimo), *Power Translator*<sup>®</sup> or *iTranslator* (Lernout & Hauspie), *LIDIA* and *C-Star II* (GETA), *TACT* (CRLT, Franche-Comté)); (ii) computer-aided translation systems, such as *An-Nakel Al-Arabi* and *European Translator* (CIMOS), *TRANSIT* (Star), *Translator's Workbench* (Trados).

### 4.2.4 Automatic summarisation systems

Today, no summarisation system is reliable enough to answer general needs. Existing systems are designed to process homogeneous corpora for very precise tasks but cannot deal with heterogeneous corpora.

Known summarisation systems are *SAFIR* (Cams-Lalic and EDF), *Cicéron* (CORA), *SummarizerTM* (Inxight - Xerox), *RAFI* (Landisco, Nancy).

### 4.2.5 Language understanding systems

Understanding systems are composed of different NLP modules, such as analysers and generators. They are used as a basis to man-machine dialogue systems, translation, summarisation, knowledge retrieval systems, etc. and have been developed for either written or spoken data. For instance, *ILLICO* (LIM, Marseille), a *prototype for NLP multi-agent system* (GRESEC, Grenoble), and *ItiSACT* (IASC, ENST Bretagne) were designed for written tasks, whereas *Openvox SLS* (Vecsys), *DictaMed* (GREYC, Caen), *C-Star II* (GETA, Grenoble 1) were designed for speech understanding, recognition and synthesis.

### 4.2.6 Computer-aided authoring systems

Computer-aided authoring systems can be classified into two different categories, computer-aided checking and computer-aided authoring systems. Many spell and grammar checkers are available on the market. Among them can be found *Voltaire* (CORA), *Hugo Plus* (Softissimo), *Pro Lexis* (Editions Diagonal), *Sans-Faute/Grammaire* (Bcdl - Hexacom), *Cordial* and *Lexical* (Synapse), *Vortex* (IRIT, Toulouse 3), *OrthoNet* (CILF), *ADICO*<sup>®</sup> *médical* (IES). As for computer-aided writing systems, these are in particular *Vitipi* (IRIT, Toulouse 3), *Unitype* (Softissimo), *NTK.FOCUS* (Nemesia), *Euro-Letter Professional* (distributed by Apsydoc).

### 4.2.7 Speech processing systems

We decided to rank speech processing tools in four main categories: (i) analysis systems, for example *SNORRI* (LORIA, Nancy), *WaveEdit* (GEOD, Grenoble 1), *Phonedit* (LPL, SQLab); (ii) recognition systems, namely *CK10.5* (Parrot SA), *DragonDictate V3* and *Dragon Naturally Speaking* (Dragon), *ViaVoice* (IBM), *Voice Xpress Profesional* (Lernout & Hauspie); (iii) synthesis systems, including *Syntaix* (LPL, Provence), *Lia\_phon* (LIA, Avignon), *KALI* (Elsap, Caen), *ELAN Text to Speech* (Elan Informatique); and (iv) dialogue systems, such as *Openvox SLS* (Vecsys), *DictaMed* (GREYC, Caen), *C-Star II* (GETA, Grenoble 1). Well-known tools from LIMSI are not described in this first release of the inventory.

### 4.2.8 Computer-aided language learning (CALL) systems

These systems are used for a variety of learning tasks and are consequently numerous. Some of them use written and speech processing modules. They were classified into five different categories: teaching French as a foreign language, French language learning, professional training, computer-aided communication, authoring systems. Among them, we can quote *Tell me More* and *Talk to me* (Auralog), *Alexia* (LIB, Franche-Comté), *Alfy* or *Orthogram* (Chrysis), *Copie Double* (Goto), *French Connexions* (Vektor), *Medmed multimédia* (CRIM, INALCO), *ALEX* (IASC, ENST Bretagne), *Kombe* (Prologia), *Speaker Auteur* (Neuroconcept), *Amical* (LRL, Clermont 2).

#### 4.2.9 Terminology management systems

Automatic processing of terminology (terms and semantic relations between terms) can have various possible applications, like elaboration or enrichment of knowledge bases, information and knowledge retrieval, text indexing, translation of technical texts, etc. Some tools can allow terminology data acquisition from existing databases: *LEXTER* and *COATIS* (EDF), *ACABIT* (IRIN, Nantes), *RCFilter* (Linguanet), *IOTA* (MRIM, Grenoble), *Seek* (IDIST-CREDO, Lille 3). Other tools are dealing with knowledge and semantic relations extraction: *LexiTrack* and *LexiBuild* (LexiQuest), *Prométhée* (IRIN, Nantes), *STK* (GREYC, Caen). There also are more complete tools such as *SPI-Grappe* (CEA), *Dixit* (Terminotics), *MultiTerm'95 Plus!* (Trados), *Termstar* (Star), *Ztermino* (Lilla, Nice), *ERACLES* (INIST) and *Lexp* (LCI).

#### 4.2.10 Optical Character Recognition systems

An increasing need is expressed to computerise information (in particular forms of old manuscripts). Scanners and optical character recognition software allow to transform these hardcopies into electronic documents. Two types of OCR were identified: typed character recognition where identified tools are *ICR Suite Pro*<sup>®</sup> (SWT), *EasyReader Elite* (Mimetics), *IrisPen* and *ReadIris* (IRIS) *TextBridge Pro 98* (ScanSoft), *OmniPage Pro 9.0* (Caere) and hand writing recognition from which only one tool was identified, namely *Solare* (LERISS, Paris 12).

#### 4.2.11 Information management and retrieval

Information processing is of a paramount importance in terms of strategy and economy. More precisely, having a quick access to a pertinent information, especially intranet or internet information is required by all professional sectors. Different types of tools were identified in this field: indexing tools, information retrieval and filtering, natural language query systems (including Internet search engines), text mining. Many products are offered to the users, from which *Search'97* (Verity), *Intuition* and *Darwin* (CORA), *Alchemy* (Bva), *eXtense* (Echo), *ANAGRAM* (Triel), *DigOut4U* and *Class4U* (Arisem), *SPIRIT* (Tgid), *LexiQuest* (LexiQuest), *WINDEX* (Multimédia Solutions), *ECILA* (Ecila), *Lokace* (Forlog), *Voila* (France Telecom), *Halpin* (GEO, Grenoble 1), *Ulysse* (GREYC, Caen), *Gargantua* (Siatel), *Docubase Entreprise* (Docubase Systems), *CinDoc* (Cincom).

## 5. Maturity of language tools

The present study was completed by another survey conducted by ELDA aiming at analysing the maturity of language engineering tools within the French market. "Maturity" is a vague notion, since a language technology can be considered as mature or not depending on its application and on the targeted users. It is precisely the users and the way technology meets their needs that will define a tool or an application as mature or not. Some systems are not widely used, either because they have unsatisfactory performance or their application, if any, do not fulfil users' requirements.

However, faced with the huge amount of information available, users will need Natural Language systems and interfaces to access data in a more intuitive way. They will also need tools to structure and manage information, and to disseminate it.

Among the tools identified during the survey that would be adequate for successful technology transfers, we may quote:

- Text generation systems, which could be associated with translation, OCR or speech recognition systems, to facilitate the dissemination of multilingual information.
- Machine translation systems, associated with Internet search engines, will enable people to access information in any language. For professional translators, the development of controlled languages should improve the performance of MT systems.
- Text summarisation systems could be integrated in information management systems or search engines in order to help users select information.
- Language understanding systems should enable users to do searches in a more intuitive way, without being obliged to constantly rephrase their requests.
- The integration of spell/grammar checkers into OCR and machine translation systems should improve their performance.
- Speech recognition and voice synthesis systems could be combined with oral translation systems to make multilingual access to information possible.
- OCR systems should improve their performance by becoming "hybrid", that is processing both hand writing and typed characters. frequent as data input mode, data output being made by voice synthesis.

## 6. Conclusion

This survey allowed to identify a large number of NLP tools. Not all of them are or will be available for technology transfer that may turn them into useful and (best-)selling products. Many are still prototypes or research components. Nevertheless, a clear panorama may help streamline the efforts required to do so. No specific assessment has been conducted to evaluate the performance of these tools neither in terms of technology nor in terms of usefulness and usability.

An evaluation paradigm is still necessary to measure how mature such tools are before incorporating them into information processing systems. Here we should insist that "non-mature" NLP tools may be sufficient for a large number of applications but the level of maturity and performance should be measured and explained to the application integrators.

## 7. Acknowledgements

We would like to thank the French Ministry of Research and Higher Education that has allowed ELDA to carry out this valuable survey work. Our gratitude is also extended to Emilie Marquois and Métiyé Meydan for their contribution to the survey.

## 8. References

- Bossard Consultants, *Industrie de la langue : marché et perspectives à 5 ans*, décembre 1988.
- Euromap, *The Euromap Report : Challenge and Opportunity for Europe's Information Society*, CCE – DG XIII, Telematics Applications Programme, september 1998.
- Ink International and ECC, *Language industries survey*, 1989.
- Mlis, *Language Engineering Directory*, CCE – DG XIII, Multilingual Information Society Programme, <http://www2.echo.lu/mlis/en/direct/home.html>, updated 07/23/1999.
- Ofil, *Guide des produits et services d'ingénierie linguistique*, OFIL, 1994.
- Ovum, Engeliën, B., McBryde, Ronnie, *Natural Language Markets : Commercial Strategies*, London, OVUM, 1991.
- Ovum, Lewin, D., Lockwood, Rose, *Language Engineering 2000*, London, OVUM, 1993.
- Owil, Biérin, E., Moulin, A., Pichault, F., *Les Industries de la langue : un marché en devenir*, Liège, Observatoire Wallon des Industriels de la Langue, 1990.
- Perriault, J., *La Logique de l'usage : essai sur les machines à communiquer*, Paris, Flammarion, 1989.
- Harvey, P.-L., *Cyberespace et communautaire*, Laval, PUL, 1995.
- Hearn, Paul M., *The Language Engineering Directory, A resource Guide to Language Engineering Organisations, Products and Services*, Madrid, Language and Technology SL, 1996.