Towards a large corpus of spoken dialogue in French that will be freely available: the "*Parole Publique*"[1] project and its first realisations

Pascale Nicolas[1], Sabine Letellier-Zarshenas[1], Igor Schadle[1], Jean-Yves Antoine[1], Jean Caelen[2]

[1] VALORIA, University of South Brittany, r. Y. Mainguy, F-56000 Vannes, France
{Jean-Yves.Antoine,Sabine.Letellier}@univ-ubs.fr

[2] CLIPS-IMAG, BP 53, F-38041 Grenoble Cedex 9, France
Jean.Caelen@imag.fr

Abstract

This paper presents two corpora (*OTG* et *ECOLE_MASSY*) which are the first delivery of the *Parole_Publique* (in English : *Public Speech*) project held by the VALORIA laboratory. This project aims at the achievement of a large corpus (orthographic transcription and morpho-syntactic annotation) of spoken French dialogues. It is primarily intended for researches on man-machine communication and will gather various types (human-human, Wizard of Oz, man-machine) of dialogues restricted to several specific tasks. The *Parole Publique* corpus will be freely distributed on the WWW.

1. Introduction

During the last decade, speech processing and more generally human language technologies have reached a degree of maturity which has resulted in the development of many marketed applications. This success owes much to the generalization of empirical corpus-based methods. In particular, speech corpora take on an essential importance in the development of spoken dialog systems, whether they are used for probabilistic language modelling, dialogue modelling or for evaluation purposes. Furthermore, the recent development of linguistic corpus shows also an increasing interest of number of linguists for experimental researches based on corpus studies (Véronis, 2000).

As a result, one of the core priorities of human language technologies lies at the moment in the realisation and in the distribution of large-scale linguistic resources.

Noticeable efforts have already been made to that effect during the 90s (*British National Corpus*, *Penn Treebank*,…). Nevertheless, the likelihood is that there will be still much more to do that what was already made! In particular, two limitations in the current development of language resources are susceptible to put a brake on further researches on spoken man-machine dialogue:

- On the one hand, the bulk of the spoken corpora that are currently available are mainly dedicated to language modelling for speech recognition systems. On the contrary, advanced corpora that concern speech understanding or spoken dialogue management are lacking. One should therefore wonder whether the lack of large-scale corpora of spoken dialogue should not limit future researches on interactive voice systems. Actually, most of spoken dialogue corpora only concern airline transport (ATIS) or railway information systems. This restriction should obviously add to the problems of genericity spoken dialogue systems are encountering (Hirschman, 1998).
- On the other hand, the availability of spoken language resources is highly different from one language to another. Aside from the large amount of data available in English — see for instance the *British National Corpus* (Leech, 1994) — there are only a few languages that are benefiting from a noticeable coverage. In particular, the backwardness of French in this domain is more and more obvious (Véronis, 2000): whereas the spoken part of the *BNC* has already exceeded a size of 10 millions words, the larger French corpus of transcribed speech consists of around one million words. This corpus, which has been collected for years by the DELIC (previously GARS) laboratory, presents moreover several limitations :
  - it is compound of monologues or interviews rather than really interactive dialogues,
  - it is not completely computerised,
  - is not freely distributed.

Whatever the considered French corpus, it suffers one of these drawbacks (Romary, 2000). This lack of available corpora is likely to penalize the development of French-speaking spoken language applications. The importance of this question is highlighted by the aim of the U.S. Linguistic Data Consortium (LDC) to constitute an American National Corpus (Ide and Macleod 2001) in addition to the British National Corpus (Leech 1994) in spite of the proximity of British and American English.

Thus, there is a real need for large French-speaking corpora of spoken dialogue. In this paper, we present two corpora (*OTG* and *ECOLE_MASSY*) which are the first delivery of the "*Parole Publique*" ("*Public Speech*") project of the VALORIA laboratory. This project aims at the collection of a large corpus of spoken French dialogues. It is primarily intended for researches on man-machine communication. This is why it will only gather dialogues restricted to several specific tasks (tourist information for instance). The whole of the collected corpora will be freely distributed on the WWW for any academic use.

At first, the *OTG* and *ECOLE_MASSY* corpora are described into details. Then, we present the motivations of the project as well as its practical and technical realisation. In conclusion, we outline the future developments of this project, which is by nature largely opened to any interested research centre. In particular, we discuss the connection of the "*Parole Publique*" project with complementary initiatives such as the ASILA initiative of the French CNRS research agency.

[1] "*PAROLE PUBLIQUE*" : public speech.

2. Corpus collection

The standardization of linguistic resources is a crucial preoccupation in terms of reusability. This is why the *OTG* and *ECOLE_MASSY* corpora follow a common methodology of collection, transcription and encoding that will be respected throughout the *Parole Publique* project. In this section, we first describe the common characteristics of these corpora. Section 3 will present specifically each corpus.

2.1.1. Recording : audio files

The recording conditions vary from one corpus to another (clandestine vs. visible recording, unique vs. multiple tracks recording, analogic vs digital recording). Each corpus gathers a collection of dialogues. According to the recording procedure (single or multiple tracks recording), each dialogue corresponds to one or several audio files (*wav* format).

2.1.2. Orthographic transcription

Each dialogue corresponds with one transcription file that gathers the whole speech-turns between the speakers and should include morpho-syntactic annotations. The transcription files are encoded in the XML format (Bray, Paoli and Sperberg-McQueen, 1998).

The corpora respect common guidelines of orthographic transcription. The latter are highly based on the guidelines defined for spoken French by the DELIC (Blanche-Benveniste and Jeanjean 1987). The main aim of these guidelines is to respect the syntactic structure of spoken utterances without lingering over anecdotal alternative pronunciations. In particular, these guidelines forbid the use of reduced word forms (Gibbon, Moore and Winski 1997 : 155) to represent slight pronunciation variations. For instance, the French relative *qui est*[2] should be pronounced either /kiɛ/, /kjɛ/ or /kɛ/. We represent these alternative pronunciations by the only transcription (*qui est*) and not to several reduced forms (respectively *qui est, ki-est* and *k-est*).

We have slightly modified the DELIC guidelines in respect with some recommendations of the SPEECHDAT project (Gibbon, Moore and Winski 1997). These changes concern mispronunciations, unintelligible words and word fragments.

For the moment being, number sequences have not been spelled out since we have not found alternative spelling. However, we have decided, mainly for standardization considerations, to spell out these sequences in a final version of the corpora.

The transcription was carried out with the free software *Transcriber* (Barred and al. 1998). The output XML format respect the DTD defined by *Transcriber*. An additional header, based on the TEI recommendations, describe more precisely the dialogue situation.

2.1.3. Morpho-syntactic tagging

The morpho-syntactic annotation of our corpora is in progress for the moment being. We have decided to use the labels set defined during the GRACE evaluation campaign (Paroubek *and al*. 1998) of French-speaking morpho-syntactic taggers. The GRACE labels set should be considered *de facto* as a standard for the morpho-syntactic annotation of French.

The morpho-syntactic tagging of our corpora is carried out with *Cordial Analyseur* from the *Synapse* company. The studies of (Valli and Véronis 1999) showed indeed that this software — used as a morpho-syntactic tagger — presents a noticeable robustness on spontaneous speech. A translation table has been defined from the Cordial to the GRACE labels set.

A final stage of checking of the annotation by an expert remains obviously unavoidable.

2.1.4. Encoding : distributed corpora

The transcription files are encoded in the XML format. Transcriptions use the Unicode alphabet encoded on 8 bit (UFT-8).

These transcription files are distributed according to three output formats that are directed to different potential uses:

- **initial encoding** (XML). The figure 1 presents an example of orthographic transcription provided by *Transcriber*, without any morpho-syntactic annotation. Speech turns with eventual simultaneous speech are directly represented in the XML structure defined by the DTD.
- **text format encoding** (ASCII) of the orthographic transcription (figure 2) and possibly a morpho-syntactic annotation. Speech turns and simultaneous segments are represented in this format as well. On the opposite, the temporal alignment of the speech turns is not kept in this format.
- **encoding in an unique file** (Postscript or Acrobat PDF formats) that gathers the previous ASCII files.

These corpora can be freely downloaded on the WWW after the signature of a use agreement (see section 4).

3. The OTG and ECOLE_MASSY corpora

This section describes the *OTG* and *ECOLE_MASSY* corpora. In particular, it details the application domains of the collected dialogues, the recording procedures and finally the complete contents of the distributed corpus.

3.1. Human-human dialogues

Although the *Parole Publique* project concerns various kinds of finalized dialogue (human-human, Wizard of Oz, man-machine: see section 3), the corpora *OTG* and *ECOLE_MASSY* are what (Caelen and al. 1997) have called “pilot” corpora. A pilot corpus gathers natural, reliable dialogues between people involved in a specific task. In the ATIS domain, it should be for instance a telephone conversation between a passenger and the receptionist of an airline company. Such real interactions highlight linguistic and dialogical phenomena that are inherent to the considered task (Caelen and al.. 1997). The interest of pilot corpora lies in this observation of real uses and real needs, even if a natural interaction should only be considered as an idealization of man-machine communication. From a “best practices” point of view, a linguistic analysis of usages based on a pilot corpus — or on a Wizard of Oz corpus — should be helpful for the design of dialogue systems as well as for the preparation of their evaluation (Gibbon, Moore and Winski 1997 : 578-580).

[2] qui est : *which is* or *who is*

```
<?xml version="1.0" encoding="UTF-8"?> <!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Nicolas" audio_filename="1ag0365" version="1" version_date="011008">
<Speakers>
<Speaker id="spk1" name="hôtesse" check="no" type="female" dialect="native" accent="" scope="local"/>
<Speaker id="spk2" name="client" check="no" type="female" dialect="native" accent="" scope="local"/>
</Speakers>
<Topics>
<Topic id="to1" desc="1ag0365"/>
</Topics>
<Episode>
<Section type="report" startTime="0" endTime="5.980" topic="to1">
<Turn startTime="0" endTime="0.629" speaker="spk1">
<Sync time="0"/>
bonjour madame
</Turn>
<Turn speaker="spk2" startTime="0.629" endTime="3.420">
<Sync time="0.629"/>
bonjour est ce que vous avez le programme de oui e e je
</Turn>
<Turn speaker="spk1 spk2" startTime="3.420" endTime="3.856">
<Sync time="3.420"/>
<Who nb="1"/>
oui
<Who nb="2"/>
Connaissance
</Turn>
<Turn speaker="spk2" startTime="3.856" endTime="4.24">
<Sync time="3.856"/>
du monde
</Turn>
```

Figure 1— Example of XML encoding : orthographic transcription without morpho-syntactic annotation (extract of the *OTG* corpus).

fichier audio : 1ag0365

<001> hôtesse
h: **bonjour madame**
<002> client
c: **bonjour est ce que vous avez le programme de oui e e je**
<003> hôtesse+client
h: **oui**
c: **Connaissance**
<004> client
c: **du monde**

Figure 2: Example of text format encoding : orthographic transcription without morpho-syntactic annotation (extract of the *OTG* corpus identical to figure 2).

3.2. The *OTG* corpus

The collection of the *OTG* corpus was founded by the AUF French-speaking agency ("Spoken Dialogue" research action). It was recorded at Grenoble (France) by the CLIPS-IMAG laboratory and transcribed by the VALORIA laboratory. It is a pilot corpus which gathers real conversations between ordinary people and a tourist office.

3.2.1. Recording procedure

The corpus was recorded at the Tourist Office of Grenoble. It gathers conversations between a receptionist and one or several tourists. The recording followed a semi-clandestine procedure : unlike the receptionists, the tourists were not aware of the fact that they were being recorded. In order to guarantee a maximal naturalness of the dialogue, the receptionist were subjected to no particular instruction. The speech of the receptionist and the tourists were collected separately by two hidden unidirectional microphones. The counterpart of this "ecological" recording procedure is a rather high ambient noise level, since this tourist office is usually very busy.

Speech signals were recorded on two separated tracks by a digital recorder (DAT). Every dialogue provides

therefore two audio files, one for the receptionist and another one for the tourist(s).

A selection of five hours of recording was preserved for the constitution of the speech corpus. This initial resource has already been distributed to the participants of the "Spoken Dialogue" research action of the AUF.

3.2.2. Orthographic transcription

Since it was recorded on location in a noisy environment, this corpus presents a significant number of transactions with a low signal-to-noise ratio. The transcription of the noisiest dialogues proved to be difficult indeed even impossible: the transcribers did not manage to agree on many extracts of these dialogues. The DELIC guidelines suggest to represent in parallel the corresponding alternative transcriptions. In view of the significant number of conflicting speech turns in these dialogues, we chose on the contrary to keep dialogues that were presenting no listening ambiguity for the transcriber. This is why we transcribed only dialogues with an excellent or a good sound quality (table 1). Some dialogues with a good sound quality should however present an inaudible part. These dialogues have not been transcribed for the moment being. Likewise, about thirty dialogues that correspond to trilogues (interaction between a receptionist and two or more tourists) were not transcribed. In this case, it proved to be difficult to make a sure distinction between the productions of the different tourists.

Duration	excellent sound quality	good sound quality
< 30 s	159	135
30 s - 1mn	35	42
1 mn - 2mn	12	24
2 mn - 3 mn	0	2
> 3 mn	0	0

Table 1: Distribution of the dialogues of *OTG* corpus according to their duration and their sound quality.

All things considered, 315 dialogues were transcribed. This corpus corresponds approximately to 2 hours of recording and gathers dialogues between five different receptionists and more than 300 tourists (table 2). The corpus has a total size of around 26 000 transcribed words. We plan to integrate in the future dialogues of lower sound quality but whose transcription remains possible, in order to reach a critical size of 40 000 transcribed words.

recording duration	117 minutes
number of dialogues	315
number of speakers	5 receptionists / 315 tourists
number of words	25 695

Table 2: Synthetic description of the *OTG* corpus.

3.3. The *ECOLE_MASSY* Corpus

The *ECOLE_MASSY*[3] corpus was recorded in the Roux primary school of Massy (France). It was transcribed by the VALORIA laboratory. It consists of real conversations between scholars of 7 years old and their teacher. Two scenarios were used in order to restrict the dialogues respectively on a task of leisure planning in the tourist attractions around Paris. It is therefore a pilot corpus on a specific task that concerns tourist information. This corpus answers a precise scientific motivation which is the differential study of linguistic variabilities according to the age of the user (young people in this case). Because of its specific aim, this corpus is not directly usable for the design of dialogue systems between a (adult) human and a machine. It should however interest linguists and researchers in educational sciences. Moreover, one should note that many works are currently concerning the adaptation of dialogue systems for specific populations such as the elderly (Privat 2000). The adaptation to a specific kind of users will certainly represent in the future a significant research area in Man-Machine Communication. The design of adapted dialogue systems should require the observations of corpora such as the *ECOLE_MASSY* one.

3.3.1. Recording

The corpus was recorded in the classroom of a primary school of Massy (Paris area) and gathers dialogues between the teacher and her scholars on leisure activities. The recording followed a visible procedure: all the speakers were aware of the fact that they were being recorded.

The instructions provided to the children before every recording were very limited. They concerned only the objective of the transaction: the choice of a film show or the planning of leisure activities around Paris. The teacher, who played the role of the receptionist of a tourist office, received furthermore the instruction to simulate a relatively guided dialogue. Our aim was indeed to study guided dialogue strategies rather than mixed-initiative ones. In order to preserve a certain linguistic naturalness, the interactions were based on the real possibilities of leisure offered at the time of the recording.

The dialogues between the teacher and the children were collected by one visible omnidirectional microphone. Since the speakers overlap rarely in these guided dialogues, the use of a unique omnidirectional microphone is not a problem (Gibbon, Moore and Winski 1997 : 134). The speech signal — one track for the two speakers — was recorded on an analog device (audio tape). Although these recordings were made in a natural environment, the noise level is moderate.

Since the children were under the responsibility of their teacher, it was not possible for an observer of our laboratory to intervene during the recordings. This resulted in a certain loss of spontaneity of the spoken productions of the children, who faced their teacher. This is the expression of a linguistic adaptation that can be compared with others studies on Man-Machine Communication (Spérandio and Létang-Fogeac 1986)

From a semantic point of view, the productions of the children remain on the contrary very free, without going

[3] *ECOLE MASSY* : *Massy school*

beyond the perimeter of the task. The corpus is therefore representative and could be usefully compared with that of adult users. On the whole, the corpus gathers 45 minutes of recorded speech.

3.3.2. Orthographic transcription: distributed corpus.

Since all of the recorded dialogues present a quite acceptable signal-to-noise ratio, it is possible to transcribe completely this corpus. This transcription gathers 31 dialogues (table 3) between the teacher and 19 different children (table 4). The corpus has a total size of around 5 300 transcribed words.

Duration	Task : film show planning	Task : leisure activity planning
< 30 s	2	0
30 s - 1mn	6	0
1 mn - 2mn	6	10
2 mn - 3 mn	0	7
> 3 mn	0	0

Table 3. Distribution of the dialogues of the *ECOLE MASSY* corpus according to their task and their duration.

A first observation of the corpus shows a very low rate of overlapping speech turns. As stated before, this is more the expression of a controlled speech rather than those of a spontaneous speech. It would be interesting to compare this oral gender with those defined by (Biber 1988).

recording duration	45 minutes
number of dialogues	31
number of speakers	1 teacher / 19 scholars
number of words	5 300

Table 4: Synthetic description of the *ECOLE_MASSY* corpus.

4. The PAROLE PUBLIQUE projet

The two first corpora represent a relatively limited linguistic resource. Actually, this work must be related to the *PAROLE PUBLIQUE* project, which aims at the collection of a large corpus of spoken dialogues.

The recording, the transcription and the annotation of spoken dialogues constitute a heavy activity that requires a strong human investment. In order to build a representative linguistic resource, we decided to focus specifically this project on Man-Machine Communication. As a result, our corpora will concern exclusively spoken dialogues restricted to some precise tasks. It is however reasonable to think that these corpora should be used for various scientific or technological applications (adaptation of language models for speech recognition systems, design of speech understanding systems, dialogue modeling, corpus linguistic studies...).

4.1. Motivations

This project answers several motivations that should favours the representativeness of the corpora in the framework of spoken Man-Machine Communication.

4.1.1. Various types of dialogue

The project will concern different kinds of dialogue : human-human conversations as well as human-machine dialogues (Wizard of Oz simulation or real interaction with a computer). Every kind of dialogue should fulfil a specific need for interactive systems design (Gibbon, Moore and Winski 1997 : 573-594). Apart from design purposes, a comparative study of these various types of dialogue should be instructive, as former works on user adaptation showed (Morel and al. 1985; Spérandio and Létang-Figeac 1986).

4.1.2. Various types of dialogue

Genericity of spoken dialogue systems is a important problem which remains an open issue (Hirschman, 1998). As a result, the *PAROLE PUBLIQUE* project aims at considering a large variety of application domains in order to constitute a generic referent in spoken Man-Machine Communication:

- tourist information,
- hotel reservation,
- administrative inquiries,
- air control,
- computer-aided navigation …

The corpora will concern face to face interaction as well as telephone conversations. Besides, we are envisaging the collection of multi-modal corpora in collaboration with other research centers.

4.1.3. Various types of users

As we already raised, the adaptation of spoken dialogue systems to the type of user is a form of genericity which should be studied more and more in the future. Several types of users can be distinguished according to various sociologic dimensions: age (children, teenagers, adults, elderly people), handicap (disabled persons), professional vs. untrained people…

The *PAROLE PUBLIQUE* project tackles this important question from the unique point of view of the age of the speakers. The *ECOLE_MASSY* corpus was for instance intentionally restricted to young speakers. More generally, we plan to collect diversified corpora that should authorize comparative analyses with respect to this dimension.

4.2. Distribution

The whole of the corpora collected within the framework of the *PAROLE PUBLIQUE* project will be freely distributed for any academic use[4]. Practically, two ways of distribution are offered:

- On-line distribution of the transcription of any available speech corpus (orthographic transcription eventually completed by morpho-syntactic annotations provided they are available). The corresponding files (XML, TXT, Postscript or Acrobat PDF formats) can

[4] A free distribution should be examinated for certain industrial uses. In such cases, llease contact Jean-Yves Antoine (Email : Jean-Yves.Antoine@univ-ubs.fr).

be downloaded directly from the WWW pages of the *PAROLE PUBLIQUE* project (figure 3) :

www.univ-ubs.fr/valoria/antoine/parole_publique

- Off-line distribution of the transcribed corpus with the corresponding speech corpus (audio files at the *wav* format). In view of their storage size, these corpora are distributed on CD support. The distribution is free, but you are asked to pay a contribution towards postage and packing. These corpora can be order on the WWW pages of the project too.

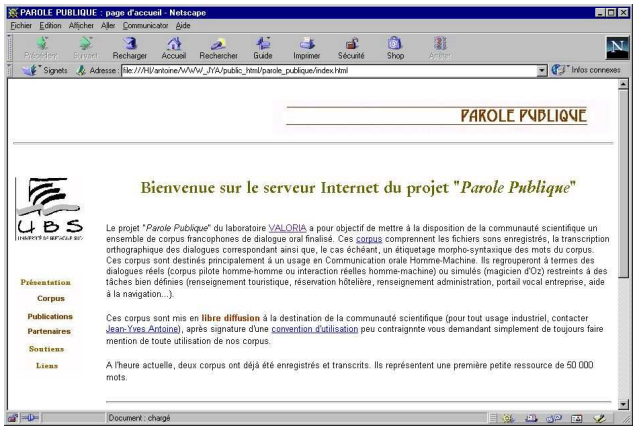

Figure 3 — Web site of the *PAROLE PUBLIQUE* project

The WWW pages of the *PAROLE PUBLIQUE* project provide a detailed description and a significant extract of every available corpus. Whatever the chosen way of distribution, you will be asked to accept a use agreement before any order. This use agreement is few restricting: you are only asked to mention any eventual use of our corpora (reference to a publication that presents the corpus, quotation of the WWW pages of the project) in your publications and WWW pages.

4.3. The ASILA research action

The whole of the corpora of the *PAROLE PUBLIQUE* project will also be distributed — under the same use agreement — in the framework of the ASILA research action[5]. It is a new initiative founded by the French CNRS research agency that aims at gathering together available corpora of French-speaking dialogues. The following laboratories are involved in this project : CORDIAL-IRISA, GREYC, LIMSI-CNRS, LIUM, LORIA.

5. Conclusion: an initiative for the free distribution of French spoken corpora

Our aim is to constitute in next three years a collection of 15 representative corpora. At the present time, the completion of this project remains however subject to budgetary constraints.

In spite of its ambitious nature, this project represents only 15 % of the annotated part of the British National Corpus. This is why this effort has to be continued and relayed by the whole of the French-speaking community on Man-Machine Communication. One objective of the *PAROLE PUBLIQUE* project is precisely to start a virtuous circle for the free distribution of large spoken French linguistic resources. Within the framework of this project, a partnership with the CLIPS-IMAG (Grenoble) and the CORDIAL-IRISA (Lannion) laboratories will enable the integration or the joint collection of various corpora of spoken dialogue. Our main wish is that other laboratories join this project or the ASILA initiative.

6. Acknowledgments

This work was partially founded by the AUF French-speaking agency ("Spoken Dialogue" research action) and the Regional Council of Brittany (PhD fellowship).

7. Références

Barras C. *and al.* 1998. Transcriber : a free tool for segmenting, labeling and transcribing speech. Proc. *LREC'98*, Granada, Spain, 1373-1376.

Biber D. 1988. Variation across speech and writing. Cambridge, Cambridge Univ. Press.

Bray T., Paoli J., Sperberg-McQueen C.M. 1998. Extensible Markup Language (XML) 1.0. W3C, http://www.w3.org/TR/REC-xml

Blanche-Benveniste C., Jeanjean C. 1987. Le français parlé : édition et transcription. Paris, Didier Erudition.

Caelen J. *and al.* 1997. Les corpus pour l'évaluation du dialogue homme-machine. Proc. *JST'97 FRANCIL*, Avignon, France, 215-222.

Gibbon D., Moore R., Winski R. (Eds.) 1997. Handbook of standards and ressources for spoken language systems. Berlin, Mouton de Gruyter, 825-834.

Hirschman L. 1998. Language understanding evaluations : lessons learned from MUC and ATIS. Proc. *LREC'98*, Granada, Spain, 117-122.

Ide N., Macleod C. 2001. The American National Corpus : a standardized resource for American English. Proc. *Corpus Linguistics 2001*. Lancaster, UK, 274-280.

Leech G. 1994. 100 million words of English : the British National Corpus. *English Today*, 9(1), 9-15.

Morel M.-A. *and al.* 1985. Analyse linguistique d'un corpus oral finalisé. Technical report, GRECO Communication Parlée, CNRS.

Paroubek P., Lecomte J., Adda G., Mariani J., Rajman M., The GRACE French Part-Of-Speech Tagging Evaluation Task. Proc. *LREC'1998*, Granada, Spain, 433-441.

Pierrel J.-M. 2000. Ingénierie des langues, Paris, Hermès.

Privat R. 2000. Interrogation multimodale de consultation de serveurs d'informations : application aux personnes âgées. Proc. *RJC-IHM'2000*, Berder island, France, 127-130.

Romary L. 2000. Outils d'accès à des ressources linguistiques. *In* (Pierrel 2000), 193-212.

Spérandio J.-C., Létang-Figeac C. 1986. Simulation expérimentale de dialogues oraux en communication homme-machine. Technical report GRECO Communication Parlée, CNRS.

Valli A., Véronis J. 1999. Etiquetage grammatical des corpus de parole, *Revue Française de de Linguistique Appliquée*, *RFLA*, 4(2), 113 -134. [English version : Grammatical tagging of spoken corpora: an experiment. Workshop on *New methods and formalisms for corpus linguistics*. Oct. 2000. Aix-en-Provence, France].

Véronis J. 2000. Annotation automatique de corpus, In (Pierrel 2000), 111-130.

[5] ASILA WWW pages : http://www.loria.fr/projets/asila/