# A French Phonetic Lexicon with Variants for Speech and Language Processing

P. Boula de Mareüil<sup>1</sup>, C. d'Alessandro LIMSI-CNRS BP 133 F-91403 Orsay cedex {mareuil,cda}@limsi.fr **F. Yvon** ENST – Dept. Informatique et Réseaux 46 rue Barrault F-75634 Paris cedex 13 yvon@enst.fr

#### Abstract

This paper reports on a project aiming at the semi-automatic development of a large orthographic-phonetic lexicon for French, based on the Multext dictionary. It details the various stages of the project, with an emphasis on the methodological and design aspects. Information regarding the lexicon's content is also given, together with a description of interface tools which should facilitate its exploitation.

#### 1. Introduction

This article describes the semi-automatic development of a wide-coverage phonetic lexicon for the French language. The main goal of the project reported here is to take advantage of data and expertise available in this field, for studying and assessing automatic grapheme-tophoneme (GP) transcription in French, within the framework of various applications in speech processing (synthesis and recognition) or natural language processing (e.g. spelling checking). The dictionary, which was set up on the basis of the Multext programme (Ide & Véronis, 1994), will be put at the disposal of the scientific and industrial community. It is also designed for the purpose of linguistic description of the French language — particularly of its phonology, dialectology, computerassisted learning, didactic and historical phonetics.

Depending on domains and applications, the problems of GP transcription and its evaluation are not exactly the same. Text-to-speech synthesis systems generally yield a single referential pronunciation, whereas in speech recognition, it is necessary to represent a set of possible pronunciation variants, for each orthographic entry. Therefore, the development of resources for GP transcription in a varied context raises a number of questions, especially: how fine should be the phonetic alphabet? which transcription norm should be chosen? which place should be given to variants? Indeed, the discussion led under the ægis of the French-speaking Francil network (Yvon et al., 1998), on GP conversion evaluation, showed the many difficulties of this task, as well as the lack of freely and directly available resources, except BDLEX (Pérennou et al., 1992). It was thus decided to build a dictionary, based on Multext and completed with proper names and acronyms, explicitly integrating parts-of-speech and phonetic variants.

The originality of this work lies on the use of semiautomatic methods, which enabled the transcription validation of large quantities of data (almost 350,000 orthographic forms). Indeed, our consortium disposes of V. Aubergé ICP – Univ. Stendhal 1180 av. Centrale BP25 F-38040 Grenoble cedex 9 auberge@icp.inpg.fr J. Vaissière, A. Amelot ILPGA – Univ. Paris III 19 rue des Bernardins F-75 005 Paris jvaiss@msh-paris.fr

three automatic GP converters, the reliability of which was diagnosed during the AUPELF-UREF Francil campaign, in terms of accuracy and precision. It is also worthwhile to notice that the three participating systems used here produced "normative" outputs. By matching these outputs, it is possible to detect the words which are transcribed differently, the transcription of which must then be imperatively controlled by an expert phonetician.

A major preoccupation has presided over the design of this lexicon: the re-usability of the resources. This finds expression in three constraints:

- taking into account all the potential types of use of these resources, which led to a reflection concerning the representation of variability;
- adhering to available or emerging standards;
- anticipating possible extensions of these resources.

Relying on three systems, a process of semi-automatic production of phonetic transcriptions was designed, enabling the development of a wide coverage lexicon, within a restricted time frame. This was done in three main steps, which are detailed in the following sections: lexicon design, addressing methodological issues and including precise specifications for the nature and the format of the data; production of the phonetic transcriptions; at least, development of tools enabling a better exploitation of the results achieved.

# 2. Development of the Lexicon

# 2.1. Original Lexicon

The original lexicon is one of the versions of the French dictionary developed within the framework of the Multext project. This version is the one which was used in the GRACE project, on the assessment of morpho-syntactic taggers for the French language (Paroubek et al., 1997). It is composed of 310,332 inflected orthographic forms, corresponding to 27,873 different lemmas. Each entry of the original lexicon comprises three fields: graphic form, associated lemma and morpho-syntactic description. The latter specifies the part-of-speech, as well as various additional information (e.g. gender and number for noun forms; tense, person, number and mode for verb forms, etc.). This way, the information associated to a graphic form enables the disambiguation of heterophonous homographs and an expansion of abbreviations. Table 1 provides a sample of this lexicon.

<sup>&</sup>lt;sup>1</sup> Now with Elan TTS, Toulouse (mareuil@elan.fr)

| Form      | Lemma     | Tag     |
|-----------|-----------|---------|
| couvent   | couvent   | Ncms    |
| couvent   | couver    | Vmip3p- |
| mat       | mat       | Afpms   |
| mat       | mat       | Ncms    |
| as        | as        | Ncms    |
| as        | avoir     | Vaip2s- |
| MM.       | messieurs | Ncmp    |
| Messieurs | monsieur  | Ncmp    |

**Table 1**: examples of lexical entries.

This lexicon was then completed with about 10,000 proper names: place names, first and last names, company names, etc. Most of them have been automatically extracted from newspaper corpora. These extractions were realised thanks to the morpho-syntactic tagger developed at LIA (Spriet *et al.*, 1996), and the lists obtained were manually checked. Note the Multext tagset has been extended marginally, since its "semantic type" feature only distinguishes two kinds of proper names: town or country names (e.g. *Londres*, Np-s-c) and company names (e.g. *IBM*, Np-s-s). The following values were additionally introduced for semantic categorisation: f for first names (e.g. *Jean*, Npms-f) and 1 for last names (e.g. *Bailly*, Np-s-1). These additional distinctions were also made in the original Multext lexicon.

### 2.2. Transcription Guidelines

The first step of the specification work consisted of defining the level of linguistic description encoded in the transcriptions. The development of a phonetic lexicon raises the problem of the description of *variability*: any orthographic form may be pronounced a number of ways. We adopted a system enabling the representation of the various possible phonetisations of a form, in a single field. This approach, which is also that of other pronunciation dictionaries such as BDLEX, seemed to be more consistent with Multext than the alternative, consisting of listing the different variants — which is done, for instance, in CELEX (Burnage, 1990). However, this approach requires a format be specified; and some procedures need be associated to this representation, in order to yield all the possible variants.

Following Laporte (1989) and Lacheret-Dujour (1990), in particular, several types of phonological variants may be distinguished:

- systemic variability, conditioned by the linguistic environment — illustrated by the three different realisations of the word *six* in the following utterances
  - Luc a six (/si/) billes,
  - Luc a six (/siz/) ans,
  - *Luc en a six* (/sis/);
- contextual or stylistic variability, that is conditioned by the realisation of the units (idiolectal or sociolectal variants, variants linked to prosody or co-articulation, in the optional realisation of liaisons, schwas, diæreses or assimilations).

Given the constraints (to use existing transcription systems), a notation scheme was specified on the basis of "broad" phonetics, widely stemming from IPA and SAMPA for its machine-readable representation. SAMPA<sup>2</sup>, which is a well-established reference (Gibbon *et al.*, 1997), is enriched with a series of diacritics and meta-symbols which enable a compact encoding of pronunciation variants (see Tables 2 and 3). Diacritics specify the phoneme which precedes (like ~ already present in SAMPA), so as to distinguish: optionally realised phonemes which correspond to floating segments (i.e. present in the underlying form, e.g. sonant geminates, terminations of numbers such as *cinq*), latent phonemes, neutralisation phenomena (particularly of mid vowels, in unstressed position) and voicing assimilations (reserved to consonants other than sonants). By "underlying form", we mean the quality represented in the lexical representation of the word, here used as the starting point of the derivation of assimilation and neutralisation rules.

Whereas BDLEX uses archiphonemes for mid vowels when the quality opposition is neutralised, the diacritics < and > enable a straightforward assessment of the precision of a transcription in a traditional alphabet. It is the same with the mute e (/@-/) and the phonological behaviour of finals. The SAMPA recommendations for French define a special role for the forward slash / (ASCII 47), namely as a marker of certain vowels archiphonemes or indeterminacies: e.g. *maison* /mE/zo~/. However, as this symbol is widely used as a delimiter of phonemic transcriptions, the diacritics < and > were preferred in our project. Likewise, x-SAMPA<sup>3</sup> suggests the figure zero be reserved for use as a diacritic meaning voiceless (IPA under-ring), a convention we have decided to follow.

A second extension to SAMPA was required when a simple sequence of atomic units was no more sufficient. In particular, parentheses surround symbols to be substituted (e.g. *charter* /(|t)SaRt(9|E)R/: the first pronunciation being assumed to be the preferred one). Curly brackets are used for liaison-related variants — the notation scheme is the following:

{TranscriptionWithoutLiaison|TranscriptionWithLiaison\_}

The liaison — this specificity of French, absent from traditional dictionaries, is a major problem for speech recognition — is the realisation of a normally mute final consonant, in the context of a following word which begins with a vowel, a mute h, or sometimes a glide.

| Diacritics          | Meta-symbols         |
|---------------------|----------------------|
| - optionality       | ( beginning of group |
| . voicing           | ) end of group       |
| 0 unvoicing         | { beginning of final |
| < opening           | } end of final       |
| > closing           | alternative          |
| _ liaison consonant |                      |

 Table 2: symbols added to SAMPA.

To finish with, let us mention another symbol, used to note the disjunctive *h*: the star, like in *Le Robert* or BDLEX, which precedes the phonetic transcription (e.g. *hic* /\*ik/). The \* may also be used before a phonological glide (e.g. *ouistiti* /\*wistiti/) or between the transcriptions of the letters composing spelled acronyms (e.g. *HLM* /\*aS\*El\*Em/).

<sup>&</sup>lt;sup>2</sup> http://www.phon.ucl.ac.uk/home/sampa/

<sup>&</sup>lt;sup>3</sup> http://pitch.phon.ucl.ac.uk/home/sampa/x-sampa.htm

Guidelines detailing the norm adopted were compiled, in order to enable a coherent control over the transcriptions. They will be joined to the final distribution of the project.

The variants described are lexical, with a couple of noticeable exceptions: the liaison and the mute *e*. All the (other) contextual, stylistic variants, a fine description of co-articulation phenomena, etc. are excluded. However, the latter should be deduced automatically. This is an important point for the applications which will use this lexicon. For instance, the typical and well-known cases of *médecin, maintenant* and the drop of the liquid as in *quatre* can be computed a posteriori, without having resort to the written string. Likewise, the place of possible epentheses (e.g. *ours blanc, film*) are to be post-processed.

| 1: /                           | 1 + 1 = 10 + 0             |
|--------------------------------|----------------------------|
| samedi /sam@-di/               | anecdote /anEk.dOt@-/      |
| <i>vendredi</i> /va~dR@di/     | absolu /ab0sO>lu/          |
| <i>quatre</i> /katR@-/         | ananas /anana( s)/         |
| <i>petite</i> /p@-tit@-/       | grand /gRa~{ t_}/          |
| villégiature /vil-leZjatyR@-/  | grands /gRa~{ z_}/         |
| <i>prompte</i> /pRo~p-t@-/     | <i>divin</i> /div{e~ in_}/ |
| irréaliste /iR-Realist@-/      | $six/si{s- z_}/$           |
| $\acute{e}tait / e < tE{ t_}/$ | $plus/ply{s- z_}/$         |

Table 3: examples of phonetic transcriptions.

Liaison consonants (/n/, /p/, /R/, /t/, /z/) are added after nouns in the plural and after all the other parts-of-speech. In return, we did not indicate whether the liaison is more or less likely (e.g. in the context of words in -r + latent consonant, as in *court*) — on this topic, see Adda-Decker *et al.* (1999).

Optional and obligatory realisations of schwas are distinguished. With a few exceptions such as *pelage* (/p@laZ@-/), which is in phonological opposition with *plage* (/plaZ@-/) a kind of "three-consonant law" is applied: final schwas and non-final schwas, in the context (V#)CeCV are marked as optional. It is the same within compound words formed with *garde-* and *porte-* when the second element is at least disyllabic: e.g. *garde-malade* (/gaRd@-malad@-/), *porte-bonheur* (/pORt@-bOn9R/).

To decide, for instance, between /O>/ and /o</, from the underlying transcription, two rules (and only two) are stated to produce quality assimilations of the pairs /o|O/, /e|E/, /2|9/. They are applied sequentially:

- 1. An open vowel becomes closed in open syllable, except in ferm final syllable (i.e. when a final schwa is optional, as in *alcalose*, the preceding vowel keeps its quality). Hence the transcriptions of *têtu* (/tE>ty/) facing *tête* (/tEt@-/).
- 2. There is regressive assimilation of aperture (i.e. modification underr the influence of the following phoneme), by vowel harmony: in a sequence mid-closed mid-open (resp. mid-open mid-closed) vowels, the first one gets open (resp. closed). In the sequences /E...e, E...o, O...e, O...o/, whatever the complexity of the consonant cluster /.../, the glides are transparent: they do not stop the propagation.

This leads to:

fêté: /fE>te/ (by 1) testé: /tE>ste/ (by 2) microphone: /mikRo<f0n@-/

(by 1, which opens the closed /o/ of the prefix)

*jeunesse*: /Z9nEs@-/ (by 1, which would give

/Z9>nEs@-/, then 2, which opens the /9>/).

The underlying form of the graphic 'o' elsewhere than in ferm final syllable is /O/, except before a lengthening consonant (/z/), and except in the prefixes *auto-*, *psycho-*, etc., where the target phoneme, possibly accompanied by an opening diacritic, is /o/ (Walter, 1976). For the verb endings in *-ai*, of the 1<sup>st</sup> person of the singular in the plural and in the preterit, the notation we adopted is (e|E): e.g. *dirai* /diR(e|E)/. For the 'e' before double consonant, the <u>default</u> rule is as follows: /e/ if the 'e' is initial and if the consonant is not 'r' (e.g. *effort* /e<fOR/, *ecchymose* /ekimoz@-/); /E/ in the other cases — more often than not.

# 2.3. Automatic Transcription

This project gathers four teams: ENST, LIMSI-CNRS, ICP (which provided and adapted their own devices to the specific needs) and ILPGA, whose contribution focussed on the definition of transcription conventions, as well as on the manual verification of the pronunciations. For further information concerning the architectures of the three GP systems used, refer to Aubergé (1991), Yvon (1996) and Boula de Mareüil (1997). Each transcriber yielded a phonetisation for all the entries of Multext, independently. The transcriptions this way obtained were then aligned, so as to establish and weight a list of forms for which there was disagreement among the three laboratories. We used an algorithm implementing a tolerant editing distance computation between strings, based on a weighted grammar of "graphonic" correspondences.

A first list of about 35,000 forms was manually corrected, corresponding to the forms for which the three systems gave different transcriptions. This very time-consuming work required the involvement of three experts during several months, as well as the development of an ergonomic interface dedicated to this task. Then, the cases of disagreement between two systems were analysed, and gave rise to new corrections, on about 12,000 forms. On the whole, one can consider that about 85 % of the transcriptions were automatically validated by matching the results of the different systems; the remaining 15 % being hand-crafted by human expertise.

Besides and independently, ENST contributed to the construction of lists of phonetised proper names, which were automatically transcribed and entirely verified, according to the procedure described Section 2.1.

# 3. Exploitation of the Lexicon

# **3.1.** Formats and Visualisation

The final result of the project consists of orthographicphonetic data (with the corresponding lemmas and Multext morpho-syntactic tags), and of a set of tools enabling a visualisation of the data as well as extractions or format conversions. The original Multext lexicon is represented under the form of a text file, whose fields are separated by tabulations. We preferred to make our pronunciation lexicon available under the form of an XML database, which looked more suited. We thus specified an XML DTD, enabling a structured representation of the information. This operation led us to tick off the original limitations of this representation, which does not enable us to associate a form to the corresponding lemma directly. This correspondence was established by associating a unique numerical key to each entry, and by calculating the associations between forms and lemmas.

An interface allowing the consultation of the lexicon through the Internet was developed, including research functions (for a form, a lemma, a transcription). It is also possible, thanks to hypertext links, to reach a lemma associated to a given form directly. This interface will be distributed with the whole of the resources.

### **3.2. Processing Tools**

Moreover, we integrated the possibility of deriving a number of additional and useful information from the representation of the phonetic transcriptions. For instance, a function enables the explicit generation of the primary variants of a form, expressed in the basic SAMPA. This way, the word *annoté*, transcribed /an-nO>te/, gives rise to four primary variants: /anOte/, /annOte/, /anote/ and /annote/. It is also possible, for each transcription, to calculate the decomposition in syllables, a skeleton in CV cohorts... Eventually, by applying a set of rewrite rules which may be parameterised, one can generate pronunciations corresponding to a more relaxed realisation of the form (e.g. [me~nna~] for *maintenant*). These calculations are made by a set of functions written in Perl, which are distributed with the lexicon, too.

# 4. Conclusion

A large phonetic, Multext-compliant lexicon of the French language spoken in France was created, including proper names and acronyms, as well as a number of pronunciation variants: it will be distributed under the form of an XML database. One of the most salient characteristics of this project was the semi-automatic approach for the development of linguistic resources: the matching of the transcriptions produced by different systems happened to point at phonetisation errors with a great precision. For the laboratories involved in the project, an important repercussion was an informal evaluation of their system's performance against lexical data.

For facilitating the development and the assessment of GP transcription systems, this lexicon would improve on being extended to a certain number of very frequent idioms: their pronunciation must be finely modelled, since it is extremely variable according to the speech rate. On the other hand, it would be essential to add phonetised running texts and recordings to this corpus, to handle heterophonous homographs, to foresee the realisation of liaisons, to capture strategies of coherent pronunciations (for instance in relation to the so-called schwa, and to different speech rates).

To us, the experience of this project was useful, and is directly usable. We hope that its operational exploitation, which we wish as wide as possible, will yield a feedback which will enable us to improve this quite exhaustive lexicon.

### Acknowledgement

This work was funded by a CNRS "Language engineering" program. Thanks to F. Sannier, O. Corbin and C. Rustin, who verified the phonetic transcriptions, together with A. Amelot. We also thank F. Béchet for providing us with tagged lists of proper names.

### References

- Adda-Decker, M., Boula de Mareüil, P. & Lamel, L. (1999). "Pronunciation variants in French: schwa & liaison", *ICPhS*, San Francisco (pp. 2239-2242).
- Aubergé, V. (1991). La synthèse de la parole : « des règles aux lexiques », PhD thesis, Université Pierre Mendès-France (Grenoble).
- Boula de Mareüil, P. (1997). Étude linguistique appliquée à la synthèse de la parole à partir du texte, PhD thesis, Université de Paris XI (Orsay).
- Burnage, G. (1990). *CELEX: a guide for users*, Technical Report, University of Nijmegen, Center for Lexical Information, Nijmegen.
- Gibbon, D., Moore, R. & Winski, R. (Eds.) (1997). Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Berlin.
- Ide, N. & Véronis, J. (1994). "Multext: Multilingual text tools and corpora", *COLING*, Kyoto (pp. 588-592).
- Lacheret-Dujour, A. (1990). Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur, PhD thesis, Université de Paris VII (Jussieu).
- Laporte, É. (1988). *Méthodes algorithmiques et lexicales de phonétisation de textes. Applications au français,* PhD thesis, Université de Paris VII (Jussieu).
- Paroubek, P., Adda, G., Mariani, J., Rajman, M. (1997). "Les procédures de mesure automatique de l'action GRACE des assignateurs de parties du discours pour le français", *Premières Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF*, Avignon (pp. 245-252).
- Pérennou, G., de Calmès, M., Ferrané, I. & Pécatte, J.-M. (1992). "Le projet BDLEX de bases de données lexicales du français écrit et parlé", Séminaire lexique, Toulouse (pp. 41-56).
- Spriet, T., Béchet, F., El-Bèze, M., de Loupy, C. & Khouri, L. (1996). "Traitement automatique des mots inconnus", *TALN*, Marseille.
- Walter, H. (1976). La dynamique des phonèmes dans le lexique français contemporain, France-Expansion, Paris.
- Yvon, F. (1996). *Prononcer par analogie : motivations, formalisation et évaluation*, PhD thesis, ENST, (Télécom Paris).
- Yvon, F., Boula de Mareüil, P., d'Alessandro, C., Aubergé., V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Goldman, J.-P., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., Zellner, B. (1998). "Objective evaluation of grapheme-to-phoneme conversion for text-to-speech synthesis in French", *Computer Speech and Language*, Special Issue on Evaluation, **12**:3 (pp. 393-410).