# MHATLex : Lexical Resources for Modelling the French Pronunciation 

Guy Pérennou, Martine de Calmès

Institut de Recherche en Informatique, Université Paul Sabatier<br>118,route de Narbonne, 31062 Toulouse Cedex France<br>(perennou, decalmes)@irit.fr


#### Abstract

The aim of this paper is to introduce the lexical resources and environment, called MHATLex, and intended for speech and text processing. A particular attention is paid to a pronunciation modelling which can be used in automatic speech processing as well as in phonological/phonetic description of languages. In our paper we will introduce a pronunciation model, the MHAT model (Markovian Harmonic Adaptation and Transduction), which copes with free and context-dependent variants. At the same time, we will present the MHATLex resources. They include 500,000 inflected forms and tools allowing the generation of various lexicons through phonological tables. Finally, some illustrations of the use of MHATLex in ASR will be shown.


## 1 Introduction

Lexical resources including a pronunciation component play an important role in automatic speech processing as well as in the phonetic description of languages.

In matters of automatic speech recognition a crucial problem is the extremely high variability of the pronunciation. One part of this variability can be taken into account through a careful training of the acousticphonetic units from a large amount of data. Another part of variability must be modelled in the lexicon as pronunciation variants.

During the last decade various experiments have proved that introducing pronunciation variants in automatic speech recognition improve the error rate, for example (Aubert \& Dugast, 1995), (Gauvain \& coll., 1993). Our experiments have focused on the contextdependant pronunciation variants and proved the error rate can be reduced by their introduction (Pérennou \& Pousse, 1998), (Pousse et Pérennou, 1999).

This was our motivation for developing lexical resources for spoken and written French. BDLex is our first generation of lexical material (Pérennou \& de Calmès, 1987), (de Calmès \& Pérennou, 1998). In BDLex each lexical entry includes orthographic form, various morpho-syntactic information and a phonological representation.

Pronunciation variants are generated thanks to a set of phonological rules, which take into account free- and context-dependant variations, and, in addition, special variation in borrowed foreign words.

It is a fact that the BDLex user may have to develop some tool for generating pronunciation-variants. In spite of the simplicity of the phonological component this may be a drawback.

This is one of the reasons why we have developed a second-generation lexical material, called MHATLex. Another reason has been the need for an appropriate lexicon for ASR in a language such as French where context-dependent pronunciations play a significant role. This aspect of pronunciation is also important for a text to speech synthesis of good quality.

In our paper we have a double purpose. For one thing, it aims to introduce the MHAT model of pronunciation including free and context-dependent variability
(Pérennou, 1995, 1996), (Pérennou \& Pousse, 1998). The role and the structure of the lexicon will be defined within this framework.

At the same time, within the above context, we will present the MHATLex materials which include 500000 inflected forms at the phonological level. Particular attention will be paid to the concept of flexibility, that is the various lexicons which can be derived from the basic materials and their use in ASR and the adaptation of phonological and the phonetic components according to the precision needed by specific application.

## 2 The phonological MHAT model

### 2.1 General framework

Four levels of representation play an important role for modelling the pronunciation:

- the syntactic level (S), where the representations consist of strings of references to inflected words and syntactic boundaries-NB: for the sake of legibility, at $\mathbf{S}$ level we refer to a word through its spelling, for example the French word mardi (Angl.: Tuesday), rather than its abstract address in the lexicon.
- the phonological word level (W), where the representations consist of strings of phonological units and phonological boundaries,
- the phonetic level ( $\mathbf{P}$ ), where the representations consist of strings of phonetic units
- the acoustic level (A), where the representation consist of stings of acoustical vectors.
The abstract morpho-phonological level(s) can be ignored as long as only the post-lexical phonological transformations are involved. This is generally the case for pronunciation dictionaries used in automatic speech processing or in various phonetic domains.

In the MHAT model a sentence is also represented at these levels. Moreover each level $\mathbf{N}$ has its alphabets of N -units and its structural constraints, which must be satisfied by any well-formed representation.

So, at each level $\mathbf{N}$, a sentence has two representations: the input representation (called also $N$ representation) $\mathrm{Y}=\mathrm{y}_{1} \ldots \mathrm{y}_{\mathrm{m}}$ and the output representation (called also $\mathrm{N}^{\prime}$-representation) $\mathrm{Y}^{\prime}=\mathrm{y}^{\prime}{ }_{1} \ldots \mathrm{y}_{\mathrm{n}}$. $\mathrm{Y}^{\prime}$ results from the adaptation of Y -the harmonic adaptation component $\mathrm{C}\left(\mathrm{N}, \mathrm{N}^{\prime}\right)-$. Consequently $\mathrm{Y}^{\prime}$ is a well-formed representation of the level $\mathbf{N}$, but this is not generally the
case for Y (except if $\mathrm{Y}=\mathrm{Y}^{\prime}$ ).
The distinction between input- and outputrepresentation structures each level $\mathbf{N}$ as a twofold representation domain including two (sub-)levels $N$ and $N^{\prime}$

The input representation Y results from an output $M^{\prime}$ representation $\mathrm{X}^{\prime}=\mathrm{x}_{1}{ }_{1} \ldots \mathrm{x}_{\mathrm{m}}$ 列 by a transduction. In the MHAT model, such a transduction is a one-step noncontextual transformation, that is a transformation based on rules of the form $x_{j}^{\prime} \rightarrow y_{k} \ldots y_{k+1}$. These rules are stored in a dictionary $\operatorname{Dic}\left(\mathrm{M}^{\prime}, \mathrm{N}\right)$.

So, at the level $S^{\prime}$, the units are inflected words, represented in $\operatorname{Dic}\left(S^{\prime}, W\right)$ as a string of $\mathbf{W}$-units, that is phonological units. For example the entry (mardi ; maRdi) means that mardi (Tuesday) is represented by /maRdi/one word reference represented by a string of 5 phonological units (the slash being the classical metasymbol used as phonological delimiter at level $\mathbf{W}$ )--


Fig. 1 - Structure of MHAT model.

### 2.2 The syntactic level

An $S$-representation consists of a word string and a syntactic structure (which is not considered here). C(S,S') inserts phonological boundaries according to the syntactic context; for example (1a) is represented by (1b) at the $S^{\prime}$ level:
je voudrais me rendre à Paris
(I would like to go to Paris)
§ je \# voudrais \#(\#) aller) \#\# à \# Paris §
Principles for the boundary adaptation are given in SPE (Chomsky \& Halle, 1968) and, for French, in (Selkirk, 1974), (Dell, 1985). We introduce a version more adapted for our purpose - for a more complete discussion see (Pérennou, 1995, 1996, 1998).

### 2.2.1 Boundary insertion

The major post-lexical phonological rules of French can be formalised using the following boundaries:

1. \# between two words having a strong cohesion, as between a proclitic and a verb, a determiner and a noun, etc;
2. \#(\#) between two words, if a liaison is optional whenever it is phonetically possible, for example between a finite verb and its complement;
3. \#\# between two words where the syntactic structure prohibits a liaison, for example between two proposition phrases;
4. $\S$ at the beginning and at the end of a phonological group (phrase).

### 2.2.2 Allomorphic adaptations depending on the phonological context

Certain words have spelling and phonological (allomorphic) variants-refered here as SPV-words-, depending on their phonological context in the sentence.

This is the case for several French determiners, including adjectives like beau/bel (beautifull). These words have two allomorphs, one before a non-consonant (that is the right context is: __\#[-cons]), the second one in other contexts. For example the demonstrative adjective ce/cet has two contextual entries:

$$
\begin{array}{ll}
(\text { cet } ; \mathrm{sEt}) / & \overline{\#[-c o n s]} \\
(\mathrm{ce} ; \mathrm{s} @) & \text { ELSE }
\end{array}
$$

where the slash stands in this case for "in the context".
NB - In this paper, all the phonetic transcriptions use the SAMPA code completed by special conventions of MHATLex. Ex.
cet horaire (this timetable) VS. ce train (this train)
Words as ce/cet, which have two possible forms in texts, have two context-dependent entries in $\operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$.

One or more entry per word? The question becomes more perplexing for words as le and les (the singular resp. plural) also sensitive to right contexts __\# [-cons]. The former $l e / l l^{\prime}$ can be assimilated to a SPV-word and has two lexical entries

$$
\begin{array}{ll}
\left(l^{\prime} ; 1\right) & / \\
(\mathrm{le} ; 1 @) & \text { ELSE }
\end{array}
$$

Ex.: l'horaire /l \# ORER@/ (the timetable) VS. le train /1@\#tRe~/ (the train)

The second has only one spelling form, but two phonological variants ( $\mathrm{lE} / \mathrm{lEz}$ ) (such words are refered here as PV-words):

$$
\begin{array}{ll}
(\text { les ; lEz) } / & \#[-c o n s] \\
(\mathrm{les} ; 1 \mathrm{l}) & \text { ELSE }
\end{array}
$$

Ex.: les horaires /lEz \# ORER@/ VS. les trains /lE \# tRe~/

From a phonological point of view, these words have two phonological variants and are sensitive to the same context. However the word les has a single entry in $\operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$. This is motivated by the fact that only one form appears in texts.

Of course, it would be possible to consider a dictionary where such words have two entries, for example $l e s_{1} / \mathrm{IEz} /$ and $l e s_{2} / \mathrm{IE} /$.

This question will come up in connection with the definition of the phonotypical dictionary $\operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}^{\prime}\right)$ introduced below.

Note that the English articles a/an and the present similar alternations, and can be compared respectively to $l e / l^{\prime}$ and les -see also, for example, the Italian articles un/uno, il/lo/l' where the contexts involved are slightly different.

### 2.3 The phonological word level

The $W$-representations consist of phonological word representations separated by boundaries. This is illustrated in (2):
Voici les horaires (here are the timetables)
/§ vwasi \#(\#) lE<+z"\#> \# ORER @ §/
(MHATLex conventions used here are explained in §2.4).
The contextual adaptation component $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$ applies
the juncture (Sandhi) phonological rules. Given the phonological context, it transforms each W-representation into a W'-representation, also called phonotypical representation.

We recall below some rules that play an important role in ASR for French (a more introduction can be found in (Pérennou, 1995, 1996, 1998). ASR literature mentions often the liaison as THE important phonological alternation of French. As a matter of fact it is ONE among others.

### 2.3.1 Liaison of latent consonants

A latent consonant (notation $\mathrm{C}^{\prime \prime}$ ) is an ending consonant, which may be pronounced in a liaison context, more precisely:
$\mathrm{C}^{\prime \prime} \rightarrow\left\{\begin{array}{ll}\mathrm{C} & \text { /__\#[-cons] } \\ (\mathrm{C}) & \text { /_\#(\#) [-cons] } \\ () & \text { lelse }\end{array} \quad\right.$ (Liaison)
where (C) stands for an optional consonant and () for the empty string.

So a latent consonant $\mathrm{C}^{\prime \prime}$ is (almost) always pronounced before the boundary \# followed by a nonconsonant (that is a vowel or a glide). Ex.:

$$
\begin{align*}
& \text { deux heures (two o'clock) } \\
& \text { /d2z"\#9R@z"/ } \rightarrow \text { [d2z 9R] } \tag{3}
\end{align*}
$$

A latent consonant $C^{\prime \prime}$ may be pronounced before the boundary \#(\#) followed by a vowel or a glide. Ex.:

$$
\begin{align*}
& \text { voudrais avoir (would have): } \\
& \text { /vudREz"\#(\#)avwaR/ } \rightarrow\left\{\begin{array}{l}
{[\text { vudREz avwaR }]} \\
{[\text { vudRE avwaR }]}
\end{array}\right. \tag{4}
\end{align*}
$$

Otherwise the latent consonant is deleted.
Remark. The phonological unit $h$ of French is not pronounced (It is why we prefer the notation /*/) but it prohibits the liaison. Consequently it has the feature [+cons]. In spelling h and $/ * /$ are not in a one to one relationship (compare (3) and (5) for example). Ex.:

> deux héros (two heros):
> /d2z"\#*eRoz"/ $\rightarrow$ [d2 eRo]

### 2.3.2 Schwa elision (deletion) at the word ending

At the phonological level the representations may contain the schwa symbol @. This abstract unit can generally be elided; if not, it is pronounced [6] the central and neutral vowel of French (close to the IPA schwa). A simplified version of the @-elision rule at word ending position is as follow:

$$
@ \rightarrow \begin{cases}() & \text { I_ } \$[\text {-cons }] \\ (6) & \text { I二 }\{\# \# ; \S ; \$[+ \text { cons }]\} \\ 6 & 1![-\mathrm{cons}] \ldots \# *\end{cases}
$$

(@fin-Elis)
where $\$$ stands for \# or \#(\#) and "!" for any boundary.
The first rule means that the ending schwa is almost never pronounced before $\$$ followed by a vowel or a glide. Ex.:
prendre à Pierre(to take from Peter): /pRãdR@ \#(\#)a\#pjER@/ $\rightarrow$ [pRãdR a pjER]

The second one is a synthetic and simplified rule for the optional elision of an ending schwa -a careful formulation must, for example, introduce a specific rule for monosyllabic words like the article le. Ex.:
horaire du bus (bus timetable)

## /ORER@\#(\#)dy\#bys/ $\rightarrow\left\{\begin{array}{l}\text { [ORER6 dy bys }] \\ \text { [ORER dy bys] }\end{array}\right.$

The schwa elision probability depends on various factors.

The third rule applies for the monosyllabic words /C@/ before the /*/ unit.Ex.:
le huit (the eight)
/l@ \# *Hit/ $\rightarrow$ [16 Hit]

### 2.3.3 Liquid elision

A liquid (/l/ or $/ \mathrm{R} /$ ) in ending position is always elided

$$
\mathrm{L} \rightarrow() \quad / \mathrm{C} \_\$ \quad \text { (LiqElis) }
$$

This situation occurs generally after a @-elision. Ex.:

$$
\begin{align*}
& \text { votre billet (your ticket) }  \tag{9b}\\
& / v O t R @ \# \text { tikE } / \rightarrow\left\{\begin{array}{l}
{[v O t R 6 \text { tikE] }} \\
{[\text { vOt tikE] }}
\end{array}\right. \tag{9a}
\end{align*}
$$

Two optional pronunciations are possible: (9a) if schwa is pronounced, (9b) if schwa is elided; in this case $/ R /$ is also elided.

### 2.3.4 Nasal assimilation of stop consonants

These rules apply optionally and participate to the contextual adaptation. The regressive (resp. progressive) nasal assimilation (PNA) (rep. RNA) applies optionally after the elision of an ending schwa and possibly of a liquid.

PNA needs a nasal vowel in the left context. Ex.:
prendre son temps (take his time)
/pRãdR@ \#(\#)sõn"\#tã/ $\rightarrow\left\{\begin{array}{l}{[\text { [pRãdR6 sõ tã }]} \\ {[\text { pRãn sõ tã] }}\end{array}\right.$
RNA needs a nasal consonant in the right context. Ex.:

> regarde-moi (look at me)
$/ \mathrm{R} @$ gaRd@\#mwa/ $\rightarrow\left\{\begin{array}{l}{[\mathrm{R}(6) \mathrm{gaRd6} \mathrm{mwa}]} \\ {[\mathrm{R}(6) \mathrm{gaRn} \mathrm{mwa}]}\end{array}\right.$
A more complete presentation can be found in (Pérennou, 1997).
Other ending juncture rules. Voice assimilation of glides and consonants, reduction of geminated units ... may be considered for an exhaustive specification of $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$.

### 2.3.5 Schwa elision in a word initial-syllable

In words like record $/ \mathrm{R} @ \mathrm{kOR} /$ (record), after a closed syllable (refered as +CS) the schwa is always pronounced as in (12a).

In other cases it may be elided as in (12b).Ex.:
Quel record? / Un record.
/kEl R@kOR/ $\rightarrow$ [kEl R6kOR]
/9~n" R@kOR/ $\rightarrow$ [9~R(6)kOR]
As an approximation, the phonological rule underlying such examples can be schematised as follow:

$$
@ \rightarrow\left\{\begin{array}{l}
6 \quad /+\mathrm{CS} \$[+ \text { cons }] \\
(6) / \text { else }
\end{array}\right.
$$

(@init-Elis)

For the French language, this is the only initial juncture rule to be considered in ASR systems. In this matter, certain languages may exhibit more important initial phonological alternations, for example Celtic and Italic languages.

### 2.4 Contextual phonological group and multipronunciation group

### 2.4.1 Contextual phonological group

In a W -representation a sub-string $\mathrm{x}_{1} \mathrm{x}_{2} \ldots$ of phonemes, object of interdependent phonological adaptations, are called contextual phonological group (CPG) and are represented by $\left\langle x_{1} x_{2} \ldots\right\rangle$ or, if necessary, $\left\langle\operatorname{Lx}_{1} x_{2} \ldots R\right\rangle$ where L (resp. R ) is a diacritic related to the left (resp. right) word context; for example: cadre (frame) represented by $/ \mathrm{ka}<\mathrm{dR} @>/$, prendre (to take) represented by /pRã<~dR@>/ where $L=\sim$ involve a possible progressive nasal assimilation. So <dR@> and <~dR@> have not the same pronunciations in all contexts (see Table 1 and 2).

A CPG may also consist of a single unit; for example: pendant (during) represented by /pãdã<t">/ where $R=$ " denotes a latent (liaison) consonant, les (the) represented by $/ \mathrm{IE}\langle+\mathrm{z} " \#>/$ where $L=+$ denotes an ending boundary and $R=$ "\# denotes a latent consonant in a proclitic (in this case the liaison is always required except before a consonant).

### 2.4.2 Multi-pronunciation group

In a W'-representation a sub-string $\mathrm{x}_{1} \mathrm{X}_{2} \ldots$ of phonemes having interdependent and multiple pronunciations are called multi-pronunciation group (MPG) and are represented as ( $\mathrm{x}_{1} \mathrm{x}_{2} \ldots$ ) or, if necessary, ( $\operatorname{Lx}_{1} \mathrm{x}_{2} \ldots \mathrm{R}$ ). For example:

$$
(\sim \mathrm{dR} @) \rightarrow \mathrm{dR} 6 ; \mathrm{n}
$$

A probability law can be associated to such a MPGrule; which can then be learned from corpora and adapted for each particular application. A MPG-rule can also be simply used as a set of possible pronunciations of a phonological unit in a given context.

### 2.4.3 The $\mathbf{C}\left(\mathbf{W}, \mathbf{W}^{\prime}\right)$ adaptation

From this definition, $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$ can be considered as a one-step phonological transformation. For example <n"\#>, <dR@>, <~dR@> are subject to the following context-dependent and one-step transformations:

$$
\begin{aligned}
& \left\langle n " \#>\rightarrow \begin{cases}n & / \ldots \#[- \text { cons }] \\
() & / \text { else }\end{cases} \right. \\
& \langle\mathrm{dR} @\rangle \rightarrow \begin{cases}\mathrm{dR} & /[\$[-\mathrm{cons}] \\
(\mathrm{dR} @ \S) & /[\ldots \S \\
(\mathrm{dR} @) & / \text { else }\end{cases} \\
& <\sim \mathrm{dR} @>\rightarrow \begin{cases}\mathrm{dR} & / \ldots \$[\text { cons }] \\
(\sim \mathrm{dR} @ \S) & / \ldots \ldots \$ \\
(\sim \mathrm{dR@}) & / \text { else }\end{cases}
\end{aligned}
$$

Tables 1 and 2 (line $1,2,3$ ) show examples of one-step C(W,W') adaptations

### 2.5 The phonetic level

The pronunciation variants are generated at level $\mathbf{P}$ using $\operatorname{Dic}\left(\mathrm{W}^{\prime}, \mathrm{P}\right)$. These pronunciations are subject to coarticulation effects and can be adapted through the $\mathrm{C}\left(\mathrm{P}, \mathrm{P}^{\prime}\right)$ component. In ASR systems it is frequently based on contextual allophones. This last component, which is not included in the lexical resources, is not discussed here.

### 2.5.1 Transduction through $\operatorname{Dic}\left(\mathbf{W}^{\prime}, \mathbf{P}\right)$

The transduction of W'-representation into Prepresentation is based on such MPG rules stored in
$\operatorname{Dic}\left(\mathrm{W}^{\prime}, \mathrm{P}\right)$. Each rule lists the possible pronunciations of a MPG and is context-free (in W'-representations the context has already be compiled using $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$ ). Ex.:

$$
\begin{aligned}
& (\mathrm{dR@} \text { @ }) \rightarrow \mathrm{dR} 6 ; \mathrm{d} \\
& (\mathrm{dR@}) \rightarrow \mathrm{dR6} ; \mathrm{dR} \\
& (\sim \mathrm{dR@}) \rightarrow \mathrm{dR} ; \mathrm{dR} ; \mathrm{n} \\
& (\sim \mathrm{dR@}) \rightarrow \mathrm{dR} 6 ; \mathrm{n}
\end{aligned}
$$

Tables 1 and 2 (line 4) give examples of such transductions.

|  | prendre <br> (to take) |  |  |  | un \# <br> (a) |  | cadre(frame) $\quad \S$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| W | p | R | ã | <~dR@> | 9~ | <n"\#> | k | a | <dR@> |
| W' | p | R | ã | dR | 9~ |  | k | a | (dR@) |
| P | p | R | ã | dR | 9~ |  | k | a | $\begin{array}{\|l} \hline \mathrm{dR6} \\ \mathrm{dR} \\ \hline \end{array}$ |

Table 1
In the example of Table 1 the liaison is prohibited for the article un. Only the last word cadre has pronunciation variants (two variants: [kadR6] and [kadR]).

|  | prendre(to take) $\quad$ (\#) |  |  |  | $\begin{aligned} & \text { son } \\ & \text { (his) } \end{aligned}$ |  |  | avis §(opinion) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| W | p | R | ã | <~dR@> | S | O | <n"\#> | a | V | i |
| $\mathrm{W}^{\prime}$ | p | R | ã | ( $\sim$ dR@) | S | O | n | a | V | 1 |
| P | p | R | ã | $\begin{aligned} & \mathrm{dR} 6 \\ & \mathrm{n} \end{aligned}$ | S | O | n | a | V | 1 |

Table2
In the example of Table 2 the liaison is require for $u n$. Only the first word prendre has pronunciation variants (here there are two variants: [pRãdR6] and [pRãn], but, of course, more sophisticated list of pronunciations may be considered).

## 3 MHAT lexicons

### 3.1 Lexicons within MHAT model

We call lexicon every list of word entries represented by given attribute(s), for example spelling, phonological representation ...


Fig. 2 - Derived lexicons.
Several kind of lexicon may be defined within MHAT model as shown in figure 2 .

### 3.1.1 The basic lexicon Lex( $\mathbf{S}^{\prime}$,W)

The basic lexicon of MHAT model is $\operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$-that is $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}\right) \equiv \operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$ - introduced below (Figure 1, §2.1) where each lexical entry is a pair (spelling, Wrepresentation), for example :
(les; lE<+z"\#>)
( $l^{\prime} ; 1$ )
(mardi ; maRdi)

### 3.1.2 The phonotypical lexicon $\operatorname{Lex}\left(\mathbf{S}^{\prime}, \mathbf{W}^{\prime}\right)$

This lexicon specifies the word phonological representations in each pertinent context, that is the word phonotypical representations. It is derived from $\operatorname{Dic}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$ through the $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$ adaptation, that is by applying CPGrules stored in $\operatorname{Dic}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$.

It consists of 4-uples (spell.;Idx;L-ctxt;W'-rep.;R-ctxt), where L- (resp. R-) ctxt specifies the left- (resp. right-) context to be satisfied; Idx is the number of the phonotypical variant. For example :

$$
\begin{array}{ll}
(\text { les } ; 1 ; ; \mathrm{IEz} ;-\mathrm{C}) & \quad(\text { record } ; 1 ;+\mathrm{CS} ; \text { r6kOR ;) } \\
(\text { mardi, } 1 ; ; \text { maRdi ;) } & (\text { record } ; 2 ;-\mathrm{CS} ; \mathrm{r}(6) \mathrm{kOR} ;)
\end{array}
$$

In these examples -C (resp. +CS, -CS) stands for nonconsonant (resp. closed syllable, non-closed syllable) in the R- or L-ctxt attributes.

An empty value means: no condition on the context. So the representation /maRdi/ is well formed at the level W' for any context; the W'-representation /lEz/ needs only that R-ctxt is -C and this results from the CPG-rule related to <+z"\#>.

In the above examples each W'-lexical entry, except the last one, has a single pronunciation, which is the replica of the W'-representation. So the first one has always the pronunciation [lEz] (so long as the R-context is a non-consonant).

In the last example, two pronunciations, $[\mathrm{RkOR}]$ and [R6kOR], can be derived (if L-context is not a closed syllable) by applying the MPG-rule : (6) $\rightarrow() ; 6$.

The index Idx is useful for transcribing usual texts into phonotypical texts where each word spelling is completed by its pertinent Idx, that is the Idx for which L- and R-ctx fit with the environment of the word in the text. So the example given in table 2 will be transcribed: ( $I l_{1} \quad v a_{1}$ ) prendre ${ }_{3}$ son $_{2}$ bus $_{1 .((\text { He is going) to take his bus). }}$
Property. A one to one relationship exists between a text and its phonotypical transcription. As a matter of fact when a word occurs in a text its legal Idx results without ambiguity from its context.

Remark. If we consider that already a few words are spelled with respect to their phonological contexts in usual text (word as le/l' as pointed out in §2.2.2), we might consider that usual texts are partly phonotypical; the use of Idx leads to a systematic phonotypical transcription.

### 3.1.3 Derived phonetic lexicons

Two lexicons play an important role in ASR.
Lex $\left(\mathrm{S}^{\prime}, \mathrm{W}^{\prime}, \mathrm{P}\right)$ is obtained by applying MPG-rules of $\operatorname{Dic}\left(\mathrm{W}^{\prime}, \mathrm{P}\right)$ on phonotypical representations. So each entry of $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}^{\prime}\right)$ generates as many entries in $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}^{\prime}, \mathrm{P}\right)$ as pronunciation variants of the phonotypical word (given the context). Lex $(\mathrm{S}, \mathrm{W}, \mathrm{P})$ is obtained in a similar way except for the context: each entry of $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}\right)$ generates as many entries in $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}, \mathrm{P}\right)$ as pronunciation variants of the word.

Table 3 shows the lexical entries of the word prendre (to take) in the four lexicons introduced above (the Lcontext, which is empty in this example, is omitted ).

| Lex | Spell. |  | W-rep. | R-ctxt | P-rep. |
| :---: | :---: | :---: | :---: | :---: | :---: |
| S'W | prendre |  | pRã<~dR@> |  |  |
| S'WP | prendre |  | pRã<~dR@> |  | pRãdR |
|  | -id- |  | -id- |  | pRãdR@ |
|  | -id- |  | -id- |  | pRãn |
| Lex | Spell. | idx | $W^{\prime}$-rep. | R-ctat | $\boldsymbol{P}$-rep. |
| SW' | prendre | 1 | pRãdR | -C |  |
|  | -id- | 2 | pRã(~dR@§) | § |  |
|  | -id- | 3 | pRã(~dR@) | ELSE |  |
| $\begin{array}{\|l} \hline \mathrm{S}^{\prime} \mathrm{W}^{\prime} \\ \mathrm{P} \end{array}$ | prendre | 1 | pRãdR | -C | pRãdR |
|  | -id- | 2 | pRã(~dR@§) | § | pRãdR |
|  | -id- | 2 | -id- | -id- | pRãdR6 |
|  | -id- | 2 | -id- | -id- | pRãn |
|  | -id- | 3 | pRã(~dR@) | ELSE | pRãdR6 |
|  | -id- | 3 | -id- | -id- | pRãn |

Table 3 -

### 3.1.4 Forward and backward constraints induced by a word

Each word occurring in a sentence induces phonological constraints on the subsequent (resp. antecedent) word, that is forward (resp. backward) phonological constraints FC (resp. BC). These constraints are those involved in CPG-rule as illustrated by the examples introduced above.

Table 4 gives a more exhaustive list of possible BC (FC takes only the values +CS or -CS). Depending on the desired precision in the phonological and/or phonetic modelling, different lists of possible values for BC can be adopted. The second column shows two simplified versions included in MHATLex resources.

All the lexicons within MHAT may be augmented by the attributes FC and BC. We adopt this option in MHATLex because this makes it easier to generate ASR lexical networks.

| BC | BC <br> simple | explicit backw. <br> constraint | Examples |
| :--- | :--- | :--- | :--- |
| * | $*$ | $\ldots$ | * |

Table 4.

## 4 MHAT model and ASR

### 4.1 The representations involved

The following chains can be consider when speech production is modelled as multilevel Markov source emission (HMM approach):
$\mathrm{X}=\mathrm{x}_{1} \ldots \mathrm{x}_{\mathrm{i}} \ldots \mathrm{x}_{\mathrm{L}}$ a word string, where $\mathrm{x}_{\mathrm{i}}$ refers to a words at the $S^{\prime}$ level or at the $W$ level (this is equivalent because a $S^{\prime}$-representation have one and only $W$-representation).
$\mathrm{X}^{\prime}=\mathrm{x}^{\prime}{ }_{1} \ldots \mathrm{x}_{\mathrm{i}} \ldots \mathrm{x}_{\mathrm{L}}^{\prime}$ be a phonotypical-word string, where $\mathrm{x}_{\mathrm{i}}^{\prime}$ refers to a words at the $W^{\prime}$ level which is a phonotypical variant of $\mathrm{x}_{\mathrm{i}}$ (identified, for example, by a pair ( $\mathrm{x}_{\mathrm{i}}{ }_{\mathrm{i}}, \mathrm{Idx}$ ) as shown in §3.1.2)
$\mathrm{V}=\mathrm{v}_{1} \ldots \mathrm{v}_{\mathrm{i}} \ldots \mathrm{v}_{\mathrm{L}}$, a phonetic representation at $P$ level, where $v_{i}$ refers to a pronunciation variant of $x_{i}$. The symbol $v_{i}$ represents (is the name of) a string of phonetic units but is considered as one symbol.
$\mathrm{A}=\mathrm{a}_{1} \ldots \mathrm{a}_{\mathrm{k}} \ldots \mathrm{a}_{\mathrm{T}}$ the acoustical representation of the input signal, where each is an acoustic vector.

### 4.2 Pronunciation variants and decision making

### 4.2.1 (W,P) models

To take into account pronunciation variants the recognition can be based on

$$
\begin{equation*}
\left(X^{*}, V^{*}\right)=\operatorname{argmax}_{X, V}\{P(A / V) \cdot P(V / X) \cdot P(X)\} \tag{13}
\end{equation*}
$$

The decision is then in favour of the best pair $\left(\mathrm{X}^{*}, \mathrm{~V}^{*}\right)$ given A .

In (13), $\mathrm{P}(\mathrm{A} / \mathrm{V})$ is the contribution of the acoustic modelling through $\operatorname{Dic}(\mathrm{P}, \mathrm{A})$-not discussed here-; $\mathrm{P}(\mathrm{X})$ is the contribution of the language model (LM); $\mathrm{P}(\mathrm{V} / \mathrm{X})$ is the contribution of the pronunciation model. It can be evaluated as follows:

$$
\begin{equation*}
\mathrm{P}(\mathrm{~V} / \mathrm{X})=\prod_{\mathrm{i}} \mathrm{p}\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}\right) \tag{14}
\end{equation*}
$$

where $\mathrm{p}\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}\right)$ is the probability of the variant $\mathrm{v}_{\mathrm{i}}$ given the word $\mathrm{x}_{\mathrm{i}}$.

The estimation of such probability is not easy. Large transcribed corpora are needed. In (Mailland and coll, 1995) an experiment of automatic training - $\mathrm{P}(\mathrm{V} / \mathrm{X})$ can be replaced by

$$
\begin{equation*}
\varepsilon(\mathrm{V} / \mathrm{X})=\prod_{\mathrm{i}} \varepsilon\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}\right) \tag{15}
\end{equation*}
$$

where $\varepsilon\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}\right)$ is a binary function taking the value 1 if $\mathrm{v}_{\mathrm{i}}$ is accepted as a pronunciation for $\mathrm{x}_{\mathrm{i}}$, the value 0 else. Then $\varepsilon(\mathrm{V} / \mathrm{X})=1$ if V is accepted as a pronunciation for X , $\varepsilon(\mathrm{V} / \mathrm{X})=0$ else.

Depending on the recogniser, several definitions of $\varepsilon$ are possible : $\varepsilon\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}\right)=1$ iff $\mathrm{v}_{\mathrm{i}}$ is a possible (resp. the most probable, one of the k-most probable...) pronunciation of $\mathrm{x}_{\mathrm{i}}$.
(W,P)-Models defined above do not control the validity of pronunciation variants with respect to their contexts in sentences.

### 4.2.2 ( $\mathbf{W}^{\prime}, \mathbf{P}$ )-models

To take into account these contexts we can replace (13) (14) (15) by (16) (17) (18)

$$
\begin{align*}
& \left(X^{*}, V^{*}\right)=\operatorname{argmax}_{X, V}\left\{P(A / V) \cdot P\left(V / X^{\prime}\right) \cdot P\left(X^{\prime}\right)\right\}  \tag{16}\\
& \mathrm{P}\left(\mathrm{~V} / \mathrm{X}^{\prime}\right)=\prod_{\mathrm{i}} \mathrm{p}\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}^{\prime}{ }_{\mathrm{i}}\right)  \tag{17}\\
& \varepsilon(\mathrm{V} / \mathrm{X})=\prod_{\mathrm{i}} \varepsilon\left(\mathrm{v}_{\mathrm{i}} / \mathrm{x}_{\mathrm{i}}^{\prime}\right) \tag{18}
\end{align*}
$$

In these models variants and LM are related to phonotypical variants.

The contribution of LM , that is $\mathrm{P}(\mathrm{X})$ in standard approaches, is now $\mathrm{P}\left(\mathrm{X}^{\prime}\right)$. As pointed out in §3.1.2 there is a one to one relationship between X and $\mathrm{X}^{\prime}$ so long as $\mathrm{X}^{\prime}$ is legal, that is if the phonotypical variants $\mathrm{x}_{\mathrm{i}}^{\prime}$ are selected in accordance to their phonological contexts in $\mathrm{X}^{\prime}$.

Let $\delta\left(\mathrm{x}_{\mathrm{i}}{ }_{\mathrm{i}}, \mathrm{x}_{\mathrm{i}+1}{ }^{\prime}\right)$ be the binary function taking the value 0 when no contextual constraints are violated in the bigram and 1 otherwise (that is the bigram is legal). $\delta$ can be extended to control the legality of $\mathrm{X}^{\prime}$ by

$$
\begin{equation*}
\delta\left(\mathrm{X}^{\prime}\right)=\prod_{\mathrm{i}} \delta\left(\mathrm{x}_{\mathrm{i}}^{\prime}, \mathrm{x}_{\mathrm{i}+1}^{\prime}\right) \tag{19}
\end{equation*}
$$

It follows that we can obtain the bigram-based LM contribution in ( $\mathrm{W}^{\prime}, \mathrm{P}$ ) model by using (20) or (21):

$$
\begin{align*}
& \mathrm{P}\left(\mathrm{X}^{\prime}\right)=\prod_{\mathrm{i}} \mathrm{P}\left(\mathrm{x}_{\mathrm{i}+1}^{\prime} / \mathrm{x}_{\mathrm{i}}^{\prime}\right)  \tag{20}\\
& \mathrm{P}\left(\mathrm{X}^{\prime}\right)=\delta\left(\mathrm{X}^{\prime}\right) \cdot \mathrm{P}(\mathrm{X})=\prod_{\mathrm{i}} \delta\left(\mathrm{x}_{\mathrm{i}}^{\prime}, \mathrm{x}^{\prime}{ }_{\mathrm{i}+1}\right) \cdot \mathrm{P}\left(\mathrm{x}_{\mathrm{i}+1} / \mathrm{x}_{\mathrm{i}}\right) \tag{21}
\end{align*}
$$

This last formula allows a second computation method of the phonotypical LM contribution. It uses the same bigrams as in the standard case (where $\mathrm{P}(\mathrm{X})=\prod_{i}$ $\left.\mathrm{P}\left(\mathrm{x}_{\mathrm{i}+1} / \mathrm{x}_{\mathrm{i}}\right)\right)$. What is specific here is the control of the legality of $\mathrm{X}^{\prime}$ by the function $\delta$. In this case training material can be the same for the two recognition models.

### 4.3 Lexical resources needed for (W,P) and ( $\mathbf{W}^{\prime}, \mathbf{P}$ ) recognisers

Knowledge involved in the above formulas -(13) to (21) - can be found in the lexicons introduced in §3.1.

### 4.3.1 Pronunciation variants

Possible pronunciations of a word (resp. a phonotypical word) are given by $\operatorname{Lex}\left(\mathrm{S}^{\prime}, \mathrm{W}, \mathrm{P}\right)$ (resp. Lex( $\left.\mathrm{S}^{\prime}, \mathrm{W}^{\prime}, \mathrm{P}\right)$ ). This allows an initial implementation of the function $\varepsilon$. Moreover they provide the phonetic transcription of the pronunciation variants.

### 4.3.2 Language models

The (W,P) model needs only a list of words. When this list is obtained from corpora, a mere lexicon Lex( $\mathrm{S}^{\prime}$ ) --list of word spellings- may be useful.

Lex $\left(\mathrm{S}^{\prime}, \mathrm{W}^{\prime}\right)$ is necessary to implement and/or to train a (W',P)-based system.

For example the estimation of the phonotypicalbigram probabilities (or more generally k-gram probabilities) need phonotypical texts transcribed using Lex(S', W').

It allows also the evaluation of the $\delta$ bigrams thanks to the L-ctx, R-ctx, FC and BC attributes.

## 5 MHATLex resources

These resources implement the basic lexicon Lex(S',W) -including backward- and forward-constraints- and the morpho-syntactic attributes inherited from BDLex. We call MHATLex-W the basic lexicon obtained in this way.

After the first experiments of the MHATLex material, new ideas related to flexible adaptation of these resources were brought out.

### 5.1 Flexible generation of specific lexicons

### 5.1.1 Projection of the lexicon

Depending on the particular application, various forms of lexicon may be useful.

Basically the environment of MHATLex allows the projection of MATLexW

- at level $W$, thanks to a GPC table, that is by performing $\mathrm{C}\left(\mathrm{W}, \mathrm{W}^{\prime}\right)$; MHATLexW' is the resulting phonotypical lexicon which includes L- and R-ctxt attributes;
- at level $P$ thanks to a MPG table, that is $\operatorname{Dic}\left(\mathrm{W}^{\prime}, \mathrm{P}\right)$; the resulting lexicon is called MHATLexP
All these projections may be restricted to a subvocabulary and to a subset of attributes.


### 5.1.2 Adaptation of the phonological model

Generally an application does not need the whole phonological component. So, since pronunciation variants are obtained through CPG- and MPG-tables, the user can implement its own pronunciation model by simple modification of these tables and he can also specify its own list of backward- and forward-constraints.

A particular simplified context-free pronunciation model, MHATLex-standard, is included in MHATLex resources.

### 5.1.3 Adaptation of the phonetic level

MHATLex allows the adaptation of the phonetic level according to the ASR acoustic-phonetic dictionary.

New phonetic alphabets lead only to the modification of the GPM-table.

### 5.2 MHATLex figures

### 5.2.1 Figures for the basic materials

Table 5 gives figures related to the amount of data in MHATLexW and in the CPG- and MPG-tables for a given pronunciation model included in the resources. For this model simplifications have been introduced. In particular the list of backward-constrain values is restricted to four values $*,+\mathrm{N},-\mathrm{N}$ and -C (as shown in table 4, column 2).

| Type of <br> W entries | Entry <br> number | GPC <br> number | GPM <br> number |
| :---: | :---: | :---: | :---: |
| Canonical | 49962 | 252 | 75 |
| Inflected | 437998 | 478 | 84 |

Table 5

### 5.2.2 Figures for the derived lexicons

Table 6 gives the corresponding figures at level $\mathrm{W}^{\prime}$. As all the CPG has been rewritten they disappear at this level, but new MPG's are introduced.

| Type of <br> W' entries | Entry <br> number | GPM <br> number |
| :---: | ---: | :---: |
| Canonical | 81456 | 148 |
| Inflected | 854452 | 220 |

Table 6
The projection of the basic lexicon at levels $\mathrm{W}^{\prime}$ and P
extends the number of entries by an ANV factor (average number of variants).

| Level | Entry nb | ANV |
| :---: | ---: | ---: |
| W | 437998 | 1.00 |
| $\mathrm{~W}^{\prime}$ | 854452 | 1.95 |
| P | 1708913 | 3.90 |

Table 7
Table 7 gives these figures for the above model of pronunciation.

It can be observed that even a simplified model increases the size of a phonetic lexicon by a factor of about 4. All these phonetic variants are not always necessary.

This has been demonstrated by several ASR experiments - see for example (Strik \& Cucchiarini, 1998) various approaches are discussed-. But, in (Pérennou \& Pousse, 1998; Pousse \& Pérennou, 1999) the evaluations comparing ( $\mathrm{W}, \mathrm{P}$ ) and ( W ', P) approaches, where MHATlex resources are used, has shown a clear advantage of the $\left(\mathrm{W}^{\prime}, \mathrm{P}\right)$ model.

A distinction must be done between recognition and training. Indeed, in a first stage, particularly when automatic phonetic alignments are involved, a careful preparation of training data may require a lexicon with a wide coverage of the pronunciation variants. In a second stage, the pronunciation model can be reduced in the recogniser by taking into account the statistics extracted from the specific training data (see §4.2).

## 6 Conclusion

The MHATlex resources can be used to generate specific lexicons for applications in various domains going from corpus lemmatisation to ASR. Basically the environment of MHATLex allows the generation of various types of lexicons according to particular pronunciation models.

More information on linguistic resources BDLex and MHATLex can be found in the web site http://www.irit.fr/ACTIVITES/EQ_IHMPT /ress_ling/.

Future developments involve ASR domain, annotation of text and speech corpora and extension of the model to other languages.

## 7 References

Aubert, X. Dugast, Ch. (1995). Improved AcousticModelling in Dictation System by Handling Liaisons and Multiple Pronunciations. Proc. Eurospeech'95 (pp. 767-70). Madrid, Spain.
Chomsky, N. Halle, M. (1968). The Sound and Pattern of English. Harper \& Row Pub.
De Calmès, M. Pérennou, G. (1998). BDLEX : a Lexicon for Spoken and Written French. In A.Rubio, N. Gallardo, R. Castro, A. Tejada, (Eds), 1st International Conference on Langage Resources \& Evaluation (pp. 1129-36), Grenade. ELRA, Paris.
Dell, F. (1985). Les règles et les sons. Herman.
Gauvain J.L. Lamel, L.F. Adda, G. Adda-Decker, M. (1993). Speaker-Independent Continuous Speech Dictation. Proc. EuroSpeech93

Mailland, A. de Calmès, M., Pérennou, G. (1995). Learning of a Phonological Component from BREF Corpus. Proc. ICPhS 95 (pp.13-9), Stockholm
Pérennou, G. (1995). Phonological Component of an Automatic Speech Recognition, The Case of Liaison Processing. In C. Sorin, J. Mariani, H. Meloni, J. Schoentgen (Eds.), Levels in Speech Communication, Relations and Communications (pp.211-24). Elsevier.
Pérennou, G. (1996). Les règles et les niveaux en phonologie: du générativisme aux modèles markoviens. In H. Méloni (Ed.), Fondements et perspectives en traitement automatique de la parole (pp.185-204). AUPELF-UREF, HACHETTE or ELLIPSES.
Pérennou, G. de Calmès, M. (1987). BDLEX Lexical Data and Knowledge Base of Spoken and Written French. European Conference on Speech Technology (pp. 393-6), Edinburgh, Scotland.
Pérennou, G. Pousse L. (1998). Phonological Component in Automatic Speech Recognition . In H. Strik, J.M. Kessens, M. Webster (Eds.), Proceedings of ESCA Tutorial and Research Workshop on

Modelling Pronunciation Variation for Automatic Speech Recognition (pp. 91-6), Rolduc, Kerkrade, Netherlands. A2RT, Dept. of Language \& Speech, University of Nimegen.
Pousse, L. Pérennou, G. (1999). Language Model Level vs. Lexical Level for Modelling Pronunciation Variation in a French CSR. In G. Olaszy, G. Németh, K. Erdohegyi (Eds.) EUROSPEECH'99 (pp. 1771-4, vol 4, 5), Budapest, Hungary.
Selkirk, E. (1974). French Liaison and X"-notation. Linguistic Inquiry 5, pp. 573-90.
Strike, H; Cucchiarini, C. (1998). Modeling Pronunciation Variation for ASR: Overview and Comparison of Methods. In H. Strik, J.M. Kessens, M. Webster (Eds.), Proceedings of ESCA Tutorial and Research Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition (pp. 137-44), Rolduc, Kerkrade, Netherlands. A2RT, Dept. of Language \& Speech, University of Nimegen.

