Language Manual

French

Pierre

Language Manual
French
Pierre
Edition 21
1 January 2005

This document was produced by Acapela Group. We welcome and consider all comments and suggestions. Please send them to:

Acapela Group
Box 1328
SE-171 26 Solna
Sweden

Phone +46 (0) 8 799 86 00
Fax + 46 (0) 8 799 86 01

Acapela Group
33, Boulevard Dolez
7000 Mons
Belgium

Tel: +32 (0)65 37 42 75
Fax: +32 (0)65 37 42 76

Acapela Group
3939, la Lauragaise
BP 58309
F-31683 Labege cedex
France

Tel: +33 (0)5 62 24 71 00
Fax: +33 (0)5 62 24 71 01

www.acapela-group.com

List of contents

1 General

This document discusses certain aspects of text-to-speech processing for the French text-to-speech system, in particular the different types of input characters and text that are allowed.

This version of the document corresponds to the High Density voice Pierre.

1.1 Notational conventions

The following notational conventions are used in this manual:

- For linguistic entities in general, **boldface** is used.
- Input text is written in a `non proportional font`.
- Output text is written in *italics*.
- Keyboard entities are written within angle brackets < >.
- Phonetic transcriptions are written within slashes or hash marks depending on the phonetic alphabet used.

The following abbreviations are used in this manual:

LM	Letter mode
SM	Sentence mode

See the User's Guide for a description of the two different reading modes. Note that Sentence mode is sometimes referred to as Normal mode.

2 Letters in orthographic text

Characters from **A-Z** and **a-z** may constitute a word. Characters that are used as letters in French and certain other European languages are also considered letters, see sections 2.1 and 2.2. The apostrophe character < ' > is not considered to be a letter, but may occur at the position of an elided vowel, in accordance with French orthographic conventions, e.g. **quelqu'un,** see section 3.6

Characters outside of these ranges, i.e. digits and non-alphanumeric characters such as punctuation characters and currency markers etc, are not considered as letters. If such a non-letter is included within a word, the word is ended where the non-letter appears and the following letters considered belonging to a new word.

2.1 Characters with diacritics common in French orthography

Characters written with the diacritics common in French orthography are mostly treated in a similar way. Insofar as they are used in French orthography they will be processed according to French pronunciation rules. Other letters containing the same diacritics will mostly be processed in a way that is reasonable from a French point of view. In LM the diacritic is read out as indicated in Table 1.

Character	SM	LM
Ç	*Ç*	*(maj) C cédille*
ü	*u*	*u tréma*
é	*é*	*e accent aigu*
â	*â*	*a accent circonflexe*
ä	*é/è*	*a tréma*
à	*à*	*a accent grave*
ç	*ç*	*c cédille*
ê	*ê*	*e accent circonflexe*
ë	*ë*	*e tréma*
è	*è*	*e accent grave*
ï	*ï*	*i tréma*
î	*î*	*i accent circonflexe*
ì	*i*	*i accent grave*
Ä	*É/È*	*(maj) A tréma*
É	*É*	*(maj) E accent aigu*
ô	*ô*	*o accent circonflexe*
ö	*œ*	*o tréma*
ò	*o*	*o accent grave*
û	*û*	*u accent circonflexe*
ù	*u*	*u accent grave*
ÿ	*y*	*i grec tréma*
Ö	*Œ*	*(maj) o tréma*
Ü	*u*	*(maj) U tréma*
Œ	*Œ*	*(maj) E dans l'O*
œ	*œ*	*e dans l'o*

Table 1 Non-French letters with diacritics

2.2 Characters treated as letters in other languages

The special letters of some other European languages are processed as indicated in Table 2.

Character	SM	LM
å	*a*	*a cercle*
Å	*A*	*Angstræm*
Æ	*É/È*	*(maj) E dans l'A*
æ	*é/è*	*e dans l'a*

Table 2 Non-French letters in the French system

3 Non-alphanumeric characters

The processing of non-alphanumeric characters varies, depending on the reading mode, context of the character, and its function within that context. These are the types of non-alphanumeric characters to be distinguished:

- Characters normally processed as punctuation, and having a direct effect on the intonation and pausing in SM.
- Non-punctuation characters that are always pronounced, with no effect on the intonation or pausing.
- Characters treated as special symbols
- Characters whose pronunciation varies according to context.
- Control characters

Below is a discussion of the characters grouped by type. For each character, the pronunciation is given in the three basic reading modes.

3.1 Punctuation characters

Table 3 lists punctuation characters permitted in the normal text input string and their readings. In SM they are all silent but they affect both rhythm and intonation as described in the sections below.

Character	LM	SM
.	*point*	(silence)
!	*point d'exclamation*	(silence)
?	*point d'interrogation*	(silence)
,	*virgule*	(silence)
:	*deux points*	(silence)
;	*point virgule*	(silence)
"	*guillemet*	(silence)
(	*parentèse gauche*	(silence)
)	*parentèse droite*	(silence)
«	*guillemet gauche*	(silence)
»	*guillemet droit*	(silence)

Table 3 Punctuation characters

3.1.1 Comma, colon and semicolon

Comma < **,** >, colon < **:** > and semicolon < **;** > cause a short pause to occur where the respective character is in the sentence. Comma is used as decimal mark in numbers, see section 4.3.

3.1.2 Quotation marks

Quotation mark < **"** > and opening and closing quotation marks < > and < > may be used around a single word or a group of words in a sentence. In SM they cause a small pause to be produced where they appear in the text (unless they happen to be right at the beginning of a sentence, or followed immediately by another punctuation mark). In LM their names are read out (differently in the two modes, see Table 3).

3.1.3 Full stop

A full stop < **.** > is a sentence terminal punctuation mark which causes a falling end-of-sentence intonation pattern and is accompanied by a somewhat longer pause.

3.1.4 Question mark

A question mark <**?**> causes two different intonation patterns:

- For Yes/No type questions (closed questions, “questions fermees”), there is a rising intonation contour.
- For questions beginning with **qui**, **que**, **ou**, **quand**, **pourquoi**, **comment**, so-called WH-questions (open questions, “questions ouvertes”), there is a falling intonation pattern.

Both are accompanied by a pause.

3.1.5 Exclamation mark

The exclamation mark < **!** > has the same effect as the full stop, causing a falling intonation pattern followed by a pause. It is often helpful to use a word stress mark to signal the key word of the phrase, thus adding more expression to the sentence, see section 6.3.1.

3.1.6 Parentheses

Parentheses < **( )** > appearing around a single word or a group of words cause a brief pause before and after the bracketed text.

3.2 Non-punctuation characters

The characters listed below in Table 4 are processed as non-letter, non-punctuation characters which are pronounced at all times in all reading modes.

Character	LM/SM
@	*a commercial*
/	*barre oblique*
{	*accolade gauche*
}	*accolade droite*
\	*barre oblique inversée*
|	*barre verticale*
[	*crochet gauche*
]	*crochet droite*
$	*dollar*
&	*et*
^	*circonflexe*
`	*apostrophe inversé*
<	*inférieur à*
>	*supérieur à*
%	*pourcent*
_	*souligné*
~	*tilde*
=	*égal* (see 3.4.3)
+	*plus* (see 3.4.3)
#	*numéro*

Table 4 Non-punctuation characters

3.3 Characters treated as special symbols

The characters in Table 5 are not considered letters. Many of these are treated as special symbols and as such are processed by the system. Apart from the special cases noted below, each of the characters in Table 5 will be read as a separate word, whether or not they occur together with letters or digits.

Character	SM	LM
¢	*centime*	*centime*
£	*livre Sterling*	*livre Sterling*
¥	*yen*	*yen*
Pt	*pesetas*	*pesetas*
ª	*petit a*	*petit a*
º	*petit o* (see special cases below)	*petit o*
¿	(silence)	*point d'interrogation inversé*
½	*un demi*	*un demi*
¼	*un quart*	*un quart*
¡	(silence)	*point d'exclamation inversé*
«	(silence) */ ouvrez les guillemets* (see 3.1)	*guillemet gauche*
»	(silence) */ fermez les guillemets* (see 3.1)	*guillemet droit*
α	*alpha*	*alpha*
β	*bêta*	*bêta*
π	*pi*	*pi*
Σ	*maj sigma*	*maj sigma*
σ	*sigma*	*sigma*
Ω	*maj oméga*	*maj oméga*
δ	*delta*	*delta*
∞	*infini*	*infini*
≡	*identique à*	*identique à*
±	*plus ou moins*	*plus ou moins*
≥	*supérieur ou égal*	*supérieur ou égal*
≤	*inférieur ou égal*	*inférieur ou égal*
≈	*peu différent d'eux*	*peu différent d'eux*
°	*degré* (see special cases below)	*degré*
•	*gros point*	*gros point*
·	*point supérieur*	*point supérieur*
√	*racine carrée*	*racine carrée*
²	*au carrée* (see special cases below)	*au carrée*

Table 5 Non-letter characters recognised by the French system

The symbols **m²**, **km²**, **mm²**, **cm²** are read as *mètre(s) carré(s)*, *kilomètre(s) carré(s)*, etc. **km**, **mm** and **cm** are also read as abbreviations, see 9.1, but **m** on its own is read as the name of the letter; to have it read differently it is necessary to enter it in the user lexicon. < **²** > will be read as *au carrée* in other cases.

The character < **º** >, which in some fonts is written with a line underneath is read as *petit o* except in the collocations **Nº** and **Nºs** which are read as *numéro(s)*. The character < **°** > is read as *degré(s)*, but this is treated the same way as < **º** > in the collocations **N°** and **N°s** which are also read as *numéro(s)*.

3.4 Characters whose pronunciation varies

The pronunciation of the characters listed below varies according to their context.

Character	LM	SM
-	*trait d'union*	(see 3.4.1 and 3.4.3)
*	*astérisque*	(see 3.4.2 and 3.4.3)

Table 6 Characters with varying pronunciation

See section 3.4.3 regarding the reading of multiple occurrences of the characters < ***** **+** **-** **=** >.

3.4.1 Hyphen

The reading of hyphen <-> follows the following principles:

- If surrounded by digits, it is pronounced *tiret*.
- When used to link words together, it is not pronounced.
- It is discarded at the end of a line, and causes the two parts of the hyphenated word to be joined into a single word.
- If a word normally contains a hyphen, and the hyphen terminates the line, the hyphen will be discarded and the two parts of the word will be merged together. To avoid the words being merged, keep the hyphenated word together on the same line.
- If surrounded by spaces, it is pronounced as a short pause.

Expression	**Reading**
`555-4758`	*555 tiret 4758*
`Donnez-le-moi`	*donnez-le-moi*

`A la fin d'une ligne,`
`on est obligé par-<CR>`
`fois de couper un mot en deux.`
	à la fin d'une ligne, on est obligé parfois de couper un mot en deux

`Comment allez-<CR>` `vous, Madame?`	*Comment allezvous, Madame?*
`Comment allez-vous,<CR>` `Madame?`	*Comment allez vous, Madame?*
`A propos - comment le saviez-vous?`	*à propos* (pause) *comment le saviez-vous?*

3.4.2 Asterisk

Asterisk < ***** > is pronounced *fois* in SM if a digit immediately precedes and follows it; it is pronounced *astérisque* in all other cases.

Expression	Reading
2*3	*deux fois trois*
*bc	*astérisque B C*

3.4.3 Multiple occurrences of the same character

In SM, if more than three of the same character occur in sequence without a space separating the characters, only the first three occurrences will be pronounced. This is only valid for the following characters: < *** + - =** >.

Expression	Reading
*****	*astérisque astérisque astérisque*
+++++	*plus plus plus*
-----	*tiret tiret tiret*
=====	*égal égal égal*

3.5 Control characters

In LM (only) most control characters are read out. In most cases they are read as *contrôle* + the appropriate letter, e.g. **^T** is read as *contrôle T*. However, some of the control characters are read in accordance with the function that they are most commonly given in text applications, as follows:

Character	LM
^H <BACKSPACE>	*charactère précédant*
^I <TAB>	*tabulation*
^J <LINE FEED>	*nouvelle ligne*
^K <VERTICAL TAB>	*tabulation verticale*
^M <RETURN>	*retour*
^? <DELETE>	*effacement*

Table 7 Control characters with special names read in letter mode

3.6 Apostrophe

An apostrophe, < **‘** >, is allowed in the text to mark elision of one vowel in front of another, for example, **j'aime**.

3.7 Characters ignored by the system

All characters that are not described in section 2 and 3 and that are not phonetic symbols or digits, are ignored by the system. Normally, these characters are omitted but some of them may cause the sentence they appear in to be silent.

4 Number processing

Strings of digits that are sent to the text-to-speech converter are processed in several different ways, depending on the reading mode, format of the digit string, and the immediately surrounding punctuation or non-numeric characters. To familiarise the user with the various types of formatted and non-formatted strings of digits that are recognized by the system, we provide below a brief description of the basic number processing along with examples.

Number processing is subdivided into the following categories:

- Full number pronunciation
- Leading zero
- Decimal numbers
- Monetary amounts
- Ordinal numbers
- Time of day
- Arithmetic operators and other symbols
- Mixed digits and letters

Note that the system does not handle date formats.

4.1 Full number pronunciation

When LM is enabled, digit strings are read as single digits, and all punctuation marks are read.

In SM, full number pronunciation is given for the whole number part of the digit string.
Optional periods < **.** > may be used in the whole number portion to separate three-digit groups into hundreds, thousands, millions, and billions. Comma < **,** > may be used only to indicate a decimal amount, see section 4.3.

The highest number read is 999999999999 (twelve digits). Numbers higher than this are read as separate digits, with pauses between three-digit groups.

Number	**Reading**
2425	*deux mille quatre cent vingt-cinq*
2.425	*deux mille quatre cent vingt-cinq*
1.000.000.000	*un milliard*
1234567890123	*un deux trois* (pause) *quatre cinq six* (pause) *sept huit neuf* (pause) *zéro un* (pause) *deux tois*

4.2 Leading zero

Digit strings that begin with **0** (zero) are read digit by digit, with pauses between groups of digits if there are four or more digits.

Number	**Reading**
09253	*zéro neuf deux* (pause) *cinq trois*
0210	*zéro deux* (pause) *un zéro*

4.3 Decimal numbers

In decimal amounts, the digits to the left of the decimal mark < **,** > observe the rules for full number pronunciation and leading zero, see section 4.1 and 4.2 respectively.

Digits occurring to the right of the decimal mark are read as full numbers or as single digits depending on the number of digits. If there are 2 or 3 digits following the decimal mark, and the first digit is not a zero, the decimal portion is read as a full number. In all other cases the decimal portion is read as single digits, with pauses occurring between groups of digits if there are four or more digits.

Number	**Reading**
16,234	*seize virgule deux cent trente-quatre*
3,1415	*trois virgule un quatre* (pause) *un cinq*
1251,04	*mille deux cent cinquante-et-un virgule zéro quatre*
2,50	*deux virgule cinquante*
2.50	*deux* (pause) *cinquante*
.65	*soixante-cinq*

4.4 Monetary amounts

Dollar amounts using the < **$** > sign are the only monetary amounts currently supported in this version of the French language. French, Belgian and Swiss franc amounts are not supported.

In a dollar amount, optional commas may be used in the dollar portion of the string to delimit thousands, hundred thousands, etc.

Expression	**Reading**
$1988.45	*mille neuf cent quatre-vingt-huit dollar et quarante-cinq cents*
$1,988.45	*mille neuf cent quatre-vingt-huit dollar et quarante-cinq cents*
$200	*deux cent dollars*
$.45	*quarante-cinq cents*

4.5 Ordinal numbers

When the ordinalizer **ieme** or **ieme** (or **eme** or **eme** or even **e)** is appended to a digit string, the ordinal version of the full number is spoken in SM only. **1er** and **1ere** or **1ere** are spoken as *premier* and *premièr**e*** respectively.

Expression	**Reading**
21ième	*vingt et unième*
21ieme	*vingt et unième*
21ème	*vingt et unième*
21eme	*vingt et unième*
21e	*vingt et unième*
42ième	*quarante-deuxième*
123ieme	*cent vingt-troisième*
95eme	*quatre-vingt-quinzième*

4.6 Time of day

Time of day is read in SM if the following format is observed:

HhM

H represents the hours
M represents the minutes

The hours and minutes must be separated by the letter **h**. The minutes field is optional.

Time	**Reading**
10h30	*dix heures trente*
1h15	*une heure quinze*
08h45	*huit heures quarante-cinq*
14h	*quatorze heures*

4.7 Arithmetic operators and other symbols

Digit strings with arithmetic operators and miscellaneous symbols are processed according to the examples below.

Expression	**Reading**
25%	*vingt-cinq pourcent*
3,4%	*trois virgule quarante-cinq pourcent*
,05%	*virgule zéro cinq pourcent*
-12	*minus douze*
+24	*plus vingt-quatre*
2*3	*deux fois trois*

4.8 Mixed digits and letters

In SM, no alphabetic characters are allowed in a digit string. The occurrence of one or more letters in a digit string will cause the digit string to be terminated and output according to the number processing described above.

Expression	**Reading**
208FR	*deux cent huit F R*
77B84Z3	*soixante-dix-sept B quatre-vingt-quatre Z trois*
0092B87-B	*zéro zéro* (pause) *neuf deux B quatre-vingt-sept tiret B*

5 French Phonetic Text

In the current version of the text-to-speech system, SAMPA (Speech Assessment Methods Phonetic Alphabet) is used when making lexicons or using phonetic strings within texts. In earlier versions, RULSYS was used. For the voices based on RULSYS, a conversion is made automatically from SAMPA to RULSYS inside the system.

We recommend new users to use only SAMPA since this is the notation that will be used in future development. Users who are already familiar with the RULSYS alphabet still have the possibility to use it when making user lexicons for all RULSYS-based voices (among them the French voice Pierre). There will be a description of RULSYS in the next section.

For the sake of clarity, SAMPA transcriptions are written within slashes (/ /) and RULSYS transcriptions within hash marks (# #). Note that neither the slashes nor the hash marks are part of the actual transcription.

The French system uses a phonetic alphabet similar to the French subset of SAMPA. The phonetic alphabet is described below.

If the pronunciation is incorrect the user may write phonetic transcriptions in the text. Then, a PRN-tag is needed to switch to phonetic mode, see User's Guide. It is also possible to make user lexicons (see User's Guide), or change the orthography of a word (see section 8) in order to achieve the preferred pronunciation.

5.1 Consonants

The table below lists the phonetic symbols in SAMPA used for the French consonants along with example words (the letters corresponding to the consonant sound are in boldface) and their transcriptions.

Consonant symbol	Example	Transcription
b	**b**a o **b** a **b**	/b a O b a b/
d	**d**i n **d**e	/d e~ d/
f	**f**i **f**e	/f i f/
g	**g**a **g**	/g a g/
j	vi **ei** **ll**e	/v j e j/
k	**c**o **q**	/k O k/
l	**l**i **ll**e	/l i l/
m	**m**ê **m**e	/m E m/
n	**n**o **nn**e	/n O n/
p	**p**i **p**e	/p i p/
R	**r**a **r**e	/R a R/
s	**s**au **c**e	/s o s/
t	**t**ou **t**e	/t u t/
v	**v**i **v**e	/b i v/
w	**ou**i	/w i/
z	o **s**e r	/o z e/
S	**ch**er **ch**e	/S E R S/
Z	ju **g**e	/Z y Z/
H	**su**is	/S H i/
N	danci **ng**	/d a~ s i N/
n j	li **gn**e	/l i n j/

Table 8 French consonant symbols

5.2 Vowels

The table below lists the phonetic symbols in SAMPA used for the French vowels along with example words and their transcriptions.

Vowel symbol	Example	Transcription
a	b**a**t	/b a/
_A1 a	b**a**s	/b _A1 a/
a~	b**an**c	/b a~/
e	b**é**e	/b e/
E	b**ai**e	/b E/
e~	b**ain**	/b e~/
@	br**e**bis	/b R @ b i/
i	b**i**s	/b i s/
O	b**o**l	/b O l/
o	b**eau**	/b o/
o~	b**on**	/b o~/
u	b**ou**t	/b u/
y	b**u**	/b y/
9	b**eu**re	/b 9 r/
2	b**oeu**fs	/b 2/
9~	**un**	/b 9~/

Table 9 French vowel symbols

Note that /_A1 a / is not a proper SAMPA symbol but it may still be used when making transcriptions. There is a corresponding symbol in RULSYS, see next section.

5.2.1 Comments on phonetic symbols for vowels

The vowel sound referred to as "e-muet" or "e-caduc" is represented by the phonetic symbol /@/. It is found in words like l**e**, pr**e**nons, and p**e**tit. The symbol /@/ should only be used when transcribing orthographic **e**:

petit	/p @ t i/
monsieur	/m @ s j 2/

The phonetic symbol /9/ should be used to represent the full vowel represented by the letters **eu** in normal orthography:

peu	/p 9/

The French rules take care of @-deletion/retention within a word, at the end of a word, and across word boundaries:

samedi	/s a m d i/
barque	/b a R k/
une amie	/y n _spc a m i/

@ is retained if more than two consonants would occur together due to @ deletion:

table de travail	/t a b l @ _spc d @ _spc t R a v a j/

If @-deletion is undesirable in a particular word, transcribing the word phonetically usually avoids the deletion. For example, **devenir** is normally pronounced *devnir* by the French system. To force the

second **e** to be pronounced, the word can be phonetically transcribed as `/d @ v @ n i R/` and all three syllables will be pronounced.

5.3 Extra symbols for phonetic details

In the current version of the French synthesis certain phonetic details can be specified in phonetic text. This can be exploited in case the user wishes to achieve an unusual pronunciation, or if the transcription automatically generated by the system is inaccurate.

5.3.1 Stress

In words with more than one syllable, one (and normally only one) of the syllables is more prominent than the others. This is referred to as word stress, or lexical stress. Words of one syllable also have word stress when spoken in isolation, although many may lose the stress in certain contexts.

In French, the word-level stress normally occurs on the last full vowel of a word in isolation, e.g. **petite.** Not all words have stress. The so-called *function* words (prepositions, pronouns, articles, etc.) are usually unstressed. When a word is combined with other words in a phrase, the stress that occurs when the word is in isolation may be lost in favour of stressing the last word of a phrase, e.g. **la petite fille**.

A word can also be stressed in a sentence to emphasise its importance, e.g. **La petite fille, je veux dire, pas la grande.** The text-to-speech system automatically determines the stress pattern of orthographically entered text. To alter the stress pattern produced by the system one can indicate phrase-level stress. In orthographic text, this is done by placing < **_X** >, where **X** represents a single digit between **0** and **9**, within a PRN-tag (see User's Guide) immediately before the word whose prominence is to be altered. The emphasis mark can also be used in transcriptions in a user lexicon.

_2	normal stress for most words
_0	makes a word non-stressed
_1	gives stress to a normally unstressed word
_3 - **_9**	gives levels of emphatic stress

Compare how the meaning is changed when the emphatic stress is varied in the sentence below:

Examples
```
C'est la fille de nos voisins.
C'est la /_4/ fille de nos voisins.
C'est la fille de /_6/ nos voisins.
```

Note that the necessary PRN-tag is not included in the examples above.

When using RULSYS notation, it is also possible to alter the word level stress, see section 6.3.1, this is not possible when using SAMPA.

5.3.2 Punctuation marks

The punctuation marks < **. ! ? ,** > used in phonetic text have the same effect on intonation as when appearing in orthographic text. In SAMPA the punctuation marks are denoted `/_/`, `/_! _/`, `/_? _/`, and `/_com _/` respectively.

5.3.3 Hyphen

In phonetic text, hyphen (in SAMPA underscore + hyphen, < **_-** >) can be used to separate parts of a compound word. If the hyphen separating two parts of a word comes at the end of a line, the word is not read until the second part on the next line is typed. For a description of the use of the hyphen character in normal orthographic text, see section 3.3.1.

6 The RULSYS phonetic alphabet

Note that we recommend new users to use only SAMPA since this is the notation that will be used in future development. Note also that it is only possible to use RULSYS when making user lexicons, not in the input text string.

The following differentiates RULSYS from SAMPA in the French system:

- no spaces are used within words in transcriptions
- it is possible to denote word stress

Note that the hash marks (# #) are used to indicate RULSYS transcriptions and to differentiate them from SAMPA transcriptions; the hash marks are not part of the actual transcriptions.

If the pronunciation is incorrect the user may write phonetic transcriptions in the text. Then, a PRN-tag is needed to switch to phonetic mode, see User's Guide. It is also possible to make user lexicons (see User's Guide), or change the orthography of a word (see section 8) in order to achieve the preferred pronunciation.

6.1 RULSYS Consonants

The table below lists the phonetic symbols in RULSYS used for the French consonants along with example words and their transcriptions.

Consonant symbol	Example	Transcription
B	ba o ba b	#BAOBAB#
D	d in d e	#DE9D#
F	f if e	#FIF#
G	g ag	#GAG#
J	v ie ill e	#VJEJ#
K	co q	#KOK#
L	l ill e	#LIL#
M	mê m e	#ME1M#
N	no nn e	#NON#
P	p ip e	#PIP#
R	ra r e	#RAR#
S	s auc e	#SO1S#
T	tout e	#TUT#
V	viv e	#VIV#
W	ou i	#WI#
Z	o s er	#O1ZE#
SH	ch erch e	#SHE1RSH#
ZH	ju g e	#ZHYZH#
J1	su is	#SJ1I#
NG	danci ng	#DA9SING#
NJ	li gn e	#LINJ#

Table 10 RULSYS consonants

6.2 RULSYS Vowels

The table below lists the phonetic symbols in RULSYS used for the French vowels along with example words and their transcriptions.

Vowel symbol	Example	Transcription
A	b**a**t	#BA#
A1	b**a**s	#BA1#
A9	b**an**c	#BA9#
E	b**é**e	#BE#
E1	b**ai**e	#BE1#
E9	b**ain**	#BE9#
E0	br**e**bis	#BRE0BI#
I	b**i**s	#BIS#
O	b**o**l	#BOL#
O1	b**eau**	#BO1#
O9	b**on**	#BO9#
U	b**ou**t	#BU#
Y	b**u**	#BY#
\	b**eu**re	#B\R
\1	b**oeu**fs	#B\1#
\9	**un**	#\9#

Table 11 RULSYS vowels

6.2.1 Comments on phonetic symbols for vowels

The vowel sound referred to as "e-muet" or "e-caduc" is represented by the phonetic symbol #E0#. It is found in words like **le**, **prenons**, and **petit**. The symbol #E0# should only be used when transcribing orthographic **e**:

petit	#PE0TI#
monsieur	#ME0SJ\1#

The phonetic symbol #\1# should be used to represent the full vowel represented by the letters **eu** in normal orthography:

peu	#P\1#

The French rules take care of E0-deletion/retention within a word, at the end of a word, and across word boundaries:

samedi	#SAMDI#
barque	#BARK#
une amie	#YN AMI#

#E0# is retained if more than two consonants would occur together due to E0-deletion:

table de travail	#TABLE0 DE0 TRAVAJ#

If E0-deletion is undesirable in a particular word, transcribing the word phonetically usually avoids the deletion. For example, **devenir** is normally pronounced *devnir* by the French system. To force the second **e** to be pronounced, the word can be phonetically transcribed as #DE0VE0NIR#, and all three syllables will be pronounced.

6.3 Extra symbols for phonetic details

In the current version of the French synthesis certain phonetic details can be specified in phonetic text. This can be exploited in case the user wishes to achieve an unusual pronunciation, or if the transcription automatically generated by the system is inaccurate.

6.3.1 Stress

For a description of stress, see section 5.3.1. In the current version of the French system, it is possible in RULSYS, but not in SAMPA, to denote word stress. This is done by placing an apostrophe < ' > before the vowel that is to receive the stress. However, this is not compulsory, as the system will stress this vowel automatically.

Example	**ami**	`#AM'I#`
	impossible	`#E9POS'IBLE0#`

Remember that only vowels are stressed, i.e., a stress mark must be followed by a vowel written in phonetic characters. Be sure, for example, that you do not leave a real apostrophe in phonetic text.

It is also possible to denote phrase level stress, see section 5.3.1.

6.3.2 Punctuation marks

When using RULSYS, punctuation marks are also permitted in phonetic text, and have the same effect as in normal text, affecting both the rhythm and intonation of the sentence. These punctuation characters are permitted in phonetic text:

, . ? ! -

The character < ' > has a completely different function when writing in phonetic text than in ordinary text. It is a reserved character that may be used to mark stress in a word,see section 6.3.1. It cannot be used to quote text or mark the place of an elided vowel in phonetic text.

6.3.3 Hyphen

Hyphen, < **-** >, in phonetic text can be used to separate parts of a compound word.

Example	**rez-de-chaussee**	`#RE-DE0-SHO1S'E#`

If the hyphen separating two parts of a word comes at the end of a line, the word is not spoken until the second part on the next line is also read in. A word written in phonetic text that contains one or more hyphens is spoken as a complete word in SM. For a description of the use of the hyphen character in normal orthographic text, see section 3.4.1.

7 Liaisons

In spoken French, a "liaison" links the final consonant of a word not normally pronounced to the following word if this begins with a vowel. For example, in the phrase **le petit chat** the **t** of **petit** is not pronounced since **chat** begins with a consonant, while in **son petit ami** a liaison between **petit** and **ami** is made and the **t** is pronounced.

In the French text-to-speech system, liaisons are generated automatically. Sometimes the system makes an undesirable liaison between two words. If this happens, the liaison can be "blocked" by inserting a special phonetic symbol between the two words where the liaison occurs. To block a liaison, insert the symbol `/_|/` within an PRN-tag (see User's Guide) between the two words.

It is also possible to create a liaison when the system does not automatically generate one. To force a liaison between two words, use an ordinary hyphen < **-** >.

Input	**Reading**
`Quand /_|/ est-il venu?`	*Quan est-il venu?*
`tout-à l'heure`	*toute à l'heure*

Note that the PRN-tag is replaced by slashes < **/ /** >in the example above.

7.1.1 Liaisons in phonetic text

When writing in phonetic text, a liaison can be blocked by simply omitting the linking character from the phonetic text string.

Example **Quand est-il venu?** can be phonetically transcribed in RULSYS as `#KA9 ETIL VE0NY?#`

A liaison can be created by including the linking consonant in the phonetic text string.

Example **Pied a terre** can be transcribed as `#PJET A TE1R#`

8 How to change pronunciation errors

Words that are not pronounced correctly by the text-to-speech converter can be entered in the user lexicon (see User's guide). There are two ways to do this: either, the user enters a phonetic transcription of the word (see section 5), or, the user rewrites the word orthographically. Phonetic transcriptions can also be entered directly in the text, using a PRN-tag (see User's guide).

8.1 Change the orthography

8.1.1 Spelling incorrectly

It is possible to intentionally misspell a word by trying to spell a word in a more phonetic manner, i.e., choosing non-ambiguous letter combinations to represent difficult sounds. For example, the letters **ch** in **archa que** might better be represented by the letter **k** , which is closer to the actual sound in the word.

Examples **computer** can be misspelled **compiouteur**
Andrew can be misspelled **Androu**

Adding an **e** to the end of a word usually causes a word-final consonant to be pronounced. If the final consonant is **m** or **n**, the addition of a final **e** will change the preceding vowel from a nasal vowel to an oral vowel.

Example The noun **est** denoting the direction (east) can be misspelled **este**
to avoid confusion with the verb **etre**
(the word est will sometimes be correctly pronounced as the direction word, for instance in the collocation **a l'est de**).

PROM can be misspelled **promme** to rhyme with **pomme**.

8.1.2 Use of hyphen

A hyphen character can be used within a word to separate two letters that might otherwise be incorrectly pronounced together.

Example **deshonneur** can be written **des-honneur**

8.1.3 Expanding acronyms

Very few acronyms are handled by the current French system (see section 9). Therefore, it may be very useful to expand them in the user dictionary. Since acronyms should be expanded to more than one word it may be difficult to enter a proper transcription. It is much easier to enter the words in question orthographically. The examples below show some acronyms and their expanded readings.

Examples **UE** Union europeenne

8.2 Using phonetic text

When you are unable to correct a pronunciation error by misspelling the word, phonetic text should be used to produce the desired pronunciation. When phonetic text is used, the system bypasses the normal pronunciation rules, and pronounces each phonetic symbol "literally", according to the examples listed in Table 8 and Table 9.

8.2.1 Writing with phonetic text

A helpful way to transcribe in phonetic text is to work with a dictionary. Often, dictionaries give a pronunciation guide for each word. They also provide a pronunciation key to show how to pronounce the special symbols used in the pronunciation guide. Similarly, Table 8 and Table 9 in this document give the pronunciation key for the special phonetic symbols used in French for the text-to-speech converter.

To transcribe a word phonetically, "sound out" the word slowly. Working with Table 8 and Table 9, find the symbols that most closely correspond to each of the sounds in the word you want to transcribe. To mark stress, decide which vowel is the most prominent, and place a primary stress mark < **1**> after the vowel.

Example

Suppose we want to transcribe the word **schooner**. We consult a dictionary and find an entry that looks like this:

> **schooner** [skuno:r] *n.m.* Petit navire a deux mats.

Among the symbols listed in the dictionary's pronunciation key are:

> **s** (s**s**alle), **k** (c**c**ou), **u** (t**ou**t), **n** (**n**ous), **o:** (s**oeu**r), **r** (**r**ue)

Using Tables 8 and 9, we find the following corresponding symbols:

> **s** (**s**auce), **k** (**c**o**q**), **u** (b**ou**t), **n** (**n**o**nn**e), **9** (b**eu**rre), **R** (**r**a**r**e)

Now we can transcribe **schooner** using the French phonetic symbols for the text-to-speech converter.

```
/s k u n 9 R/
```

9 Abbreviations

In the current version of the French text-to-speech system, the abbreviations in Table 10 are recognised in all contexts in SM only. These abbreviations are case-insensitive, and do not require a period in order to be processed as an abbreviation. In SM, if a period accompanies the abbreviation, the sentence is terminated at the abbreviation and output.

The user lexicon may be used to redefine any of the abbreviations below, or to create your own.

Abbreviation	LM	SM
dr	*D R*	*docteur*
mlle	*M L L E*	*mademoiselle*
mme	*M M E*	*madame*
st	*S T*	*saint*
ste	*S T E*	*sainte*
mm	*M M*	*millimètre(s)*
cm	*C M*	*centimètre(s)*
km	*K M*	*kilomètre(s)*
ml	*M L*	*millilitre(s)*
cl	*C L*	*centilitre(s)*

Table 12 Abbreviations in the French system

As mentioned in section 4.6, the abbreviation **h** represents **heures** when placed immediately after a digit:

Input	Reading (SM)
20h30 .	*Vingt heures trente*
20 h .	*Vingt H*

9.1 Entering abbreviations in the user lexicon

The single letters **m**, **l**, **s**, etc. are read as the names of the letters unless they are given a special transcription in the user lexicon. In LM they will always be read out as the names of the letters, even if they have been given a different transcription in the user lexicon. Thus, if one wishes the letter **M** on its own to be read as **Monsieur** or as **metre/metres** it is possible to enter suitable transcriptions in the user lexicon. Likewise one can give any other single letter, or combination of letters, a user-defined pronunciation by entering them in the user lexicon.

In order to avoid the transcription given to a single consonant letter in the user lexicon being applied also in front of an apostrophe, it is necessary to use a phonetic transcription and to start this with an equals sign < = >. When an apostrophe follows a word which has been given a transcription that begins with < = > followed by a consonant, this consonant will be read and the rest of the transcription discarded. Transcriptions are given below in both RULSYS and SAMPA notation.

User lexicon entry	Input	Reading (SM)
M #ME0SJ\1# /m @ s j 2/	Je m'appelle Pierre.	*Je monsieur appelle Pierre*
M #=ME0SJ\1# /_= m @ s j 2/	Je m'appelle Pierre.	*Je m'appelle Pierre .*
S #SE9# /s e~/	Il s'appelle Pierre.	*Il saint appelle Pierre .*
S #=SE9# /_= s e~/	Il s'appelle Pierre.	*Il s'appelle Pierre*

Since the letter **M** is often used as an abbreviation for both **Monsieur** and **metre(s)**, there are special rules for these readings. The system contains rules for dealing with the following entries in a user lexicon as readings for the single letter **m**:

User lexicon entry	Reading
`M #===ME1TR+E0#` `/_= _= _= m E t R _+ @/`	*mètre(s)* only when a number precedes *Monsieur* otherwise *m* before apostrophe
`M #==ME1TR+E0#` `/_= _= m E t R _+ @/`	*mètre(s)* only when a number precedes *M* otherwise but *m* before apostrophe
`M #=ME1TR+E0#` `/_= m E t R _+ @/`	*mètre(s)* in all cases *m* before apostrophe
`M #ME1TR+E0#` `/_= m E t R _+ @/`	**m** on its own will be read as *mètre* in all cases, even when followed by apostrophe
`M #=ME0SJ\1#` `/_= m @ s j 2/`	*Monsieur* in all cases except before apostrophe

The two-letter word **MM** will by default be read as *millimètre(s)* by the system. To have this read as *Messieurs* it is necessary to enter it in the user lexicon. As for the single-letter abbreviation **M** there are special formats allowing different pronunciations of this abbreviation.

User lexicon entry	Reading
`MM #===MILIME1TR+E0#` `/_= _= _= m i l i m E t R _+ @/`	*millimètre(s)* only when a number precedes *Messieurs* otherwise but *m* before apostrophe
`MM #==MILIME1TR+E0#` `/_= _= m i l i m E t R _+ @/`	*millimètre(s)* only when a number precedes *M M* otherwise *m* before apostrophe
`MM #=MILIME1TR+E0#` `/_= m i l i m E t R _+ @/`	*millimètre(s)* in all cases but *m* before apostrophe

For **L** the following user lexicon entries can be used:

User lexicon entry	Reading
`L #==LITR+E0#` `/_= _= l i t R _+ 2/`	*litre(s)* in all cases other than before an apostrophe
`L #LITR+E0#` `/l i t R _+ 2/`	*litre(s)* in all cases including before an apostrophe

The text-to-speech system does not normally distinguish between small (lower-case) letters and capital (upper-case) letters. However, a user lexicon can be made case-sensitive (in some systems, the user lexicon is always case-sensitive; see the User's Guide for further information). This can be exploited in order to distinguish between for instance **M** as *Monsieur* and **m** as *mètre(s)*. A useful procedure would be to have a file containing the following items (see next page):

```
M           #===ME1TR+E0#
            /_= _= _= m E t R _+ @/

MM          #===MILIME1TR+E0#
            /_= _= _= m i l i m E t R _+


Mm          #===MILIME1TR+E0#
            /_= _= _= m i l i m E t R _+

m           #==ME1TR+E0#
            /_= _= _= m E t R _+ @/
```

Then the desired items can be copied from this file into the user lexicon. Note: use either the RULSYS transcription (indicated by # #) or the SAMPA transcription (indicated by / /).