Language Manual

French

Julie, Caroline, Claire and Bruno

Language Manual
French
Julie, Caroline, Claire and Bruno
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 systems, in particular the different types of input characters and text that are allowed.

This version of the document corresponds to the High Quality voices Julie, Claire and Bruno, and the High Density voice Caroline.

2 Letters in orthographic text

Characters from **A-Z** and **a-z**, as well as “c” and vowels with a diaeresis, grave, acute or circumflex accent, may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, i.e. “n, o, a”. These letters are not pronounced as in their native languages though, they are pronounced as regular “n, o, a” etc.

Characters outside of these ranges, i.e. numbers, punctuation characters and other non-alphanumeric characters, are not considered as letters.

3 Punctuation characters

Punctuation marks appearing in a text affect both rhythm and intonation of a sentence. The following punctuation characters are permitted in the normal input text string:
, : ; “ . ? ! () { } [ ] '

3.1 Comma, colon and semicolon

Comma < **,** >, colon < **:** > and semicolon < **;** > cause a brief pause to occur in a sentence, accompanied by a small rising intonation pattern just prior to the character.

3.2 Quotation marks

Quotes < **“** > appearing around a single word or a group of words cause a brief pause before and after the quoted text.

3.3 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. A full stop may also be used as a decimal marker in a number (see chapter 5) and in abbreviations (see chapter 8).

3.4 Question mark

A question mark <**?**> ends a sentence and causes question-intonation, first rising and then falling.

3.5 Exclamation mark

The exclamation mark <**!**> behaves in a similar manner to the full stop, causing a falling intonation pattern followed by a pause.

3.6 Parentheses, brackets and braces

Parenthesis < **( )** >, brackets < **[ ]** > and braces < **{ }** > appearing around a single word or a group of words cause a brief pause before and after the bracketed text.

4 Other non-alphanumeric characters

4.1 Non-punctuation characters

The characters listed below are processed as non-letter, non-punctuation characters. Some are pronounced at all times and others are only pronounced in certain contexts, which are described in the following sections of this chapter.

Symbol	Reading
/	slash
+	plus
$	dollar
£	livre sterling
€	euro
¥	yen
<	plus petit que
>	plus grand que
%	pour cent
^	accent circonflexe
|	barre
~	tilde
@	arobas
²	au carre
³	au cube
=	egal
-	voir ci-dessous
*	voir ci-dessous

Table 1 Non-punctuation characters

4.2 The ² and ³ signs

The reading of expressions with ² and ³ is:

Expression	Reading
mm²	millimetres carres
cm²	centimetres carres
m²	metres carres
km²	kilometres carres
mm³	millimetres cubes
cm³	centimetres cubes
m³	metres cubes
km³	kilometres cubes

4.3 Symbols whose pronunciation varies depending on the context

4.3.1 Hyphen

A hyphen < **-** > is pronounced “minus” if followed by a digit. In certain date formats, in between days, the hyphen is pronounced “au”. In between years it is pronounced “a”. In other cases the hyphen is never pronounced.

Expression	**Reading**
44-3	44 moins 3
15-20 octobre	15 au 20 octobre
6-10 nov	6 au 10 novembre
1998-2004	mil neuf cent quatre-vingt-dix-huit a deux mil quatre
02-02-2002	deux fevrier deux mil deux
arc-en-ciel	arc en ciel

4.3.2 Asterisk

Asterisk < ***** > is pronounced “fois” if enclosed by digits. In other cases it is pronounced “asterisque”.

Expression	**Reading**
2*3	deux fois trois
*bc	asterisque b c

5 Number processing

Strings of digits that are sent to the text-to-speech converter are processed in several different ways, depending on the format of the string of digits 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 recognised 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
Currency amounts
Ordinal numbers
Arithmetic operators
Mixed digits and letters
Time of day
Dates
Phone numbers
Erreur ! Source du renvoi introuvable.

5.1 Full number pronunciation

Full number pronunciation is given for the whole number part of the digit string.

Example	
2425	full number
2.425	full number
24,25	24 is a full number, 25 is the decimal part

Numbers denoting thousands, millions and billions (numbers larger than 999) may be grouped using space or full stop (not comma). In order to achieve the right pronunciation the grouping must be done correctly.

The rules for grouping of numbers are the following:

- Numbers are grouped in groups of three starting at the end.
- The first group in a number may consist of one, two, or three digits.
- If a group, other than the first, does not contain exactly three digits, the sequence of digits is not interpreted as a full number.
- The highest number read is 99999999999 (eleven digits). Numbers higher than this are read as separate digits.

Number	**Reading**
2580	deux mil cinq cent quatre-vingts
2 580	“
2.580	“
25800	vingt-cinq mil huit cents
25 800	“
25.800	“
2580350	deux millions cinq cent quatre-vingt mil trois cent cinquante
2 580 350	“
2.580.350	“
1000000000	un milliard
123456789012	un deux trois quatre cinq six sept huit neuf zero un deux
23 456 789 012	vingt-trois milliards quatre cent cinquante-six millions sept cent quatre-vingt-neuf mil douze

5.2 Leading zero

Numbers that begin with 0 (zero) are read as a zero followed by the number read as a whole.

Number	**Reading**
09253	zero neuf mil deux cent cinquante-trois
020	zero vingt

5.3 Decimal numbers

Comma or full stop may be used when writing decimal numbers.

The full number part of the decimal number (the part before comma or full stop) is read according to the rules in 5.1. If the decimals (the part after comma or full stop) are more than three, the decimal part is read as separate digits. Note: A number containing full stop followed by exactly three digits is not read as a decimal number but as a full number, following the rules in 5.1.

Number	**Reading**
16,234	seize virgule deux cent trente-quatre
3,1415	trois virgule un quatre un cinq
1251,04	mil deux cent cinquante et un virgule zero quatre
1.251,04	mil deux cent cinquante et un virgule zero quatre
2,50	deux virgule cinquante
2.50	deux point cinquante
3.141	trois mil cent quarante et un

5.4 Currency amounts

The following principles are followed for currency amounts:

- Numbers with zero or two decimals preceded or followed by the currency markers **£, $, ¥** or **€** are read as monetary amounts.
- Numbers with zero or two decimals followed by the words “livre”, “dollar”, “yen” or “euro” (singular or plural) are read as monetary amounts.
- Accepted decimal markers are comma and full stop**.**
- No spaces are allowed in the number.
- The decimal part (consisting of two digits) in monetary amounts is read as “et nn pence” and “et nn centimes”.
- If the decimal part is “00” it will not be read.

Example	**Reading**
$15.00.	quinze dollars
15.00£.	quinze livres
15.00 euro.	quinze euros
€ 200.50	deux cents euros et cinquante centimes
1.000.000 **¥**	un million de yens

There is also the possibility of writing large amounts as follows:

$ 1 million	un million de dollars

5.5 Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number "1" is followed by a month name or one of the month name abbreviations. The number may be preceded by a day or an abbreviation for a day. Examples: 1 janvier, 1 jan, mardi 1 jan.
- The number is "1er, 1ere, 2nd, 2nde".
- The number is followed by "eme, eme, e, e". Examples: 5e, 6eme, 3eme, 7e.

Valid abbreviations for months: jan, fevr, fev, avr, juil, sep, sept, oct, nov and dec.

Valid abbreviations for days: lun, mar, mer, jeu, ven, sam and dim.

The abbreviations above are only expanded to names of months and days when appearing in correct date contexts.

5.6 Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below.

Expression	**Reading**
-12	moins douze
+24	plus vingt-quatre
2*3	deux fois trois
2/3	deux divise par trois
25%	vingt-cinq pour cent
3,4%	trois virgule quatre pour cent

5.7 Mixed digits and letters

If a letter appears within a sequence of digits, the groups of digits will be read as numbers according to the rules above. The letter marks the boundary between the numbers. The letter will also be read. Examples:

Expression	**Reading**
77B84Z3	soixante-dix-sept B quatre-vingt-quatre Z trois
0092B87-B	zero zero quatre-vingt-douze B quatre-vingt-sept B

5.8 Time of day

The colon is used to separate hours, minutes and seconds. When there are no seconds, "H or h" can be used to separate hours and minutes. Abbreviations such as "A.M." and "P.M." may follow or precede the time.

Possible patterns are:

a) **hh:mm (or h:mm)**

b) **hh:mm:ss (or h:mm:ss)**

c) **hhHmm (or hHmm)** ex 12H30 (3h30)

h = hour, m = minute, s = second.

In pattern a): If the "mm"-part is equal to "00", this part will not be read.

In pattern b): An "et will be inserted before the "ss"-part, and "secondes" will be added after it. If the "ss"-part is equal to "00", this part will not be read.

Pattern (c) follows the rules for pattern (a).

5.9 Dates

The valid formats for dates are:

1.dd-mm-yyyy, dd.mm.yyyy, and dd/mm/yyyy
2.dd-mm-yy, dd.mm.yy, and dd/mm/yy

"**yyyy**" is a four-digit number, "**yy** is a two-digit number, "**mm**" is a month number between 1 and 12 and "**dd**" a day number between 1 and 31.

Hyphen, full stop, and slash may be used as delimiters.

In all formats, one or two digits may be used in the "mm" and "dd" part. Zeros may be used in front of numbers below 10.

Examples of valid formats and their readings:

Type 1:	dd-mm-yyyy, dd.mm.yyyy, and dd/mm/yyyy		
10-02-2003	or	10-2-2003	dix fevrier deux mil trois
10.02.2003	or	10.2.2003	"
10/02/2003	or	10/2/2003	"

Type 2:	dd-mm-yy, dd.mm.yy, and dd/mm/yy		
10-02-03	or	10-2-03	dix fevrier deux mil trois
10.02.03	or	10.2.03	"
10/02/03	or	10/2/03	"

Ranges of days and years are also supported.

Examples:

1998-1999	mil neuf cent quatre-vingt-dix-huit a mil neuf cent quatre-vingt-dix-neuf
1939-45	mil neuf cent trente-neuf a quarante-cinq
2002/3	deux mil deux a trois
14-15 janvier	quatorze au quinze janvier

Other possible formats include :

- Lundi, 15 janvier
- Mar, 30 avril 1999
- 3 mai 1953

5.10 Phone numbers

In this section the patterns of digits that are recognised as phone numbers are described. In the pronunciation of phone numbers each group of digits is read as a full number (see also Leading zero section 5.2) with pauses between groups of numbers. Groups that contain more than three digits are read out digit by digit.

5.10.1 Ordinary phone numbers

Sequences of digits in the following formats are treated as phone numbers.

The following sequences of digits can be separated by a space, a period, or a hyphen:

- xx (xx) xxx xx xx
- xx (x)x xx xx xx xx
- xx (x) x xx xx xx xx
- (xx) xxxx xxx xxx
- (xx) xx xx xx xx xx
- xx x xx xx xx xx
- xxx xx xx xx
- xx xx xx xx xx

The following sequences can only appear in these formats:

- xxx/xx xx xx
- xxx/xx.xx.xx
- xx xxx xx xx
- xx/xxx xx xx
- xx xxx xxx xx
- xxxx/xx xx xx

5.10.2 International phone numbers

International phone numbers follow the patterns below:
International Prefix + Country code + Regional number + Local number

International prefix: “00” or “+”
Country code: 1-3 digits
Regional number: 1-3 digits with or without parentheses (see below for exact formats)
Local number: 6-8 digits

Examples:

0032 71 12 34 56	(can also be separated by a period rather than a space)
0032 (02) 123 45 67	(can also be separated by a period or a hyphen rather than a space)
0032 (0)71 12 34 56	
0033 (0)3 123 456 78	
0033 (0)3 12 34 56 78	
0033 (0)5 12 34 56	
0032 (0) 71-12.34.56	
0032-071 12 34 56	
0033 3 123 456 78	
0033 3 123 45 67	
0033 3 12 34 56 78	

6 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). In this lexicon, the user enters a phonetic transcription of the word (see chapter 7). Phonetic translations can also be entered directly in the text, using the PRN-tag (see User's guide).

7 French Phonetic Text

The French text-to-speech system uses the French subset of the SAMPA phonetic alphabet (Speech Assessment Methods Phonetic Alphabet), with the exception of the symbol /J/ which was replaced by the sequence /n j/ (ex. oignon). The symbols are written with a space between each phoneme.

Only SAMPA may be used in phonetic transcriptions. Symbols not listed here are not valid in phonetic transcriptions and will be ignored if included in the user lexicon or in a PRN tag.

7.1 Consonants

7.1.1 Symbols for the French consonants

Symbol	Word	Phonetic text	Comments
j	junior	Z y n j O R	glide
w	trois	t R w a	glide
H	huit	H i t	glide
p	papa	p a p a	
t	tante	t a~ t	
k	cacao	k a k a o	
b	bord	b O R	
d	dort	d O R	
g	galette	g a l E t	
f	femme	f a m	
s	sans	s a~	
S	chat	S a	
v	vol	v O l	
z	zero	z e R o	
Z	jouet	Z u E	
l	long	l o~	
R	rat	R a	
m	mangue	m a~ g	
n	navette	n a v E t	
N	pudding	p u d i N	

Table 2 French consonants

7.2 Vowels

7.2.1 Symbols for the French vowels

Symbol	Word	Phonetic text	Comment
i	ville	v i l	
e	et	e	
E	cher	S E R	
a	chat	S a	
O	nord	n O R	
o	gauche	g o S	
u	lourd	l u R	
y	but	b y t	
2	bleu	b l 2	
9	neuf	n 9 f	
@	demain	d @ m e~	
e~	main	m e~	Nasal
a~	grand	g R a~	Nasal
o~	rond	R o~	Nasal
9~	brun	b R 9~	Nasal

Table 3 French vowels

7.3 Pause

An underscore < **_** > in a phonetic transcription generates a small pause.

8 Abbreviations

In the current version of the French text-to-speech system, the abbreviations in table 4 below are recognised in all contexts. These abbreviations are mostly case-insensitive (except for those indictated below by "*") and require no full stop in order to be recognised as an abbreviation.

As previously mentioned, there are also abbreviations for the days of the week and the months.

Abbreviation	Reading
kg	Kilo
°C	Degres Celsius
°F	Degres Fahrenheit
°K	Degres Kelvin
bd	Boulevard
bld	Boulevard
bef	Franc belge
cie	Compagnie
cm	Centimetre
dB*	Decibel
DM*	Deutschmark
dm	Decimetre
dpt	Departement
dr	Docteur
ed	Editeur
etc	Et cetera
ff	Franc francais
gr	grams
jr	Junior
km	Kilometres
kmh	Kilometres heure
mgr	Monseigneur
mlle	Mademoiselle
MM*	Messieurs
mm	Millimetre
mme	Madame
mr	Monsieur
ms	Millisecondes
n°	Numero
nb	Nota bene
no	Numero
nr	Numero
rte	Route
ste	Sainte
st	Saint
sts	saints
tel	telephone
mt	Mont
sr	Senior
BC*	Avant Jesus-Christ
AD*	Apres Jesus-Christ
ml	Millilitre
cl	Centilitre
dl	Decilitre

Table 4 Abbreviations

9 Web-addresses and email

Web-addresses and email-addresses are read as follows:

- "www" is read as three w's spelled letter by letter.
- Full stops are read as "point", hyphens as "tiret", underscore ("_") as "underscore", slash ("/") as "slash".
- "be, uk, fr" and all the other abbreviations for countries are spelled out letter by letter.
- The "@" is read "arobas" .
- Words/strings (including "org", "com" and "edu") are pronounced according to the normal rules of pronunciation in the system and in accordance with the lexicon.

String	**Reading**
www.babeltech.com	w w w point babeltech point com
http://www.babeltech.com	h t t p deux points slash slash w w w point babeltech point com
dubois@infonie.fr	dubois at infonie point f r
jane_dubois@infonie.fr	jane underscore dubois at infonie point f r