Language Manual

HQ and HD Portuguese

Language Manual: HQ and HD Portuguese

Published 22 March 2011

This document was produced by Acapela Group. We welcome and consider all comments and suggestions. Please, use the *Contact Us* link at our website:
http://www.acapela-group.com

Table of Contents

List of Tables

Chapter 1. General

This document discusses certain aspects of text-to-speech processing for the Portuguese 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 Quality (HQ) voice Célia and the High Density (HD) voices Ester and Elia.

Please note that the *User's Guide*, mentioned several times in the manual, is called *Help* in some applications.

Note: This language manual is general and applies to all Acapela Group voices specified above. One or more of the voices may be included in a certain Acapela Group product.

Note: For efficiency reasons, the processing described in this document has a different behaviour in some Acapela Group products. Those products are:

- Acapela TTS for Windows Mobile
- Acapela TTS for Linux Embedded
- Acapela TTS for Symbian

For these products, the default processing of numbers, phone numbers, dates and times has been simplified for the low memory footprint (LF) voice formats. Developers have the possibility to change the default behaviour from *simplified* to *normal* preprocessing by setting corresponding parameters in the configuration file of the voice. Please see the documentation of these products for more information. In the following chapters, each simplification will be described by the indication *[not SP]* following the description of the standard behaviour. The *SP* in the indication stands for *Simplified Processing*.

Chapter 2. Letters in orthographic text

Characters in the ranges of *A-Z* and *a-z* as well as the characters *ç, õ, ã, à, á, â, é, ê, í, ó, ô, ú* may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, e.g. *å, ü*.

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

Chapter 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 *Number processing*) and in abbreviations (see chapter *Abbreviations*).

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 '*!*' is treated 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.

Chapter 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.

Table 4.1. Non-punctuation characters

Symbol	Reading
/	barra
+	mais
$	dólar
£	libra
€	euro
¥	iene
<	menor que
>	maior que
%	porcento
^	acento circunflexo
|	barra vertical
~	til
@	arroba
*	asterisco
=	igual
²	See below
³	See below
-	See below

4.2. The ² and ³ signs

The reading of expressions with ² and ³ is:

Expression	Reading
mm²	milímetros quadrados
cm²	centímetros quadrados
m²	metros quadrados
km²	quilómetros quadrados
mm³	milímetros cúbicos
cm³	centímetros cúbicos
m³	metros cúbicos
km³	quilómetros cúbicos

4.3. Symbols whose pronunciation varies depending on the context

4.3.1. Hyphen

A hyphen '-' is pronounced *menos* in two cases:

1. if followed by a digit and no other digit is found in front of the hyphen

2. if followed by a digit and an equals sign. If there is no equals sign '=', it is pronounced *traço*

[not SP] In certain date formats, when the hyphen is indicating time period [from – to] it is pronounced *a*. When it occurs between days and years it is pronounced *de*. Hyphen can also function as a delimiter in a date, see section *Dates* . In other cases the hyphen is never pronounced.

Expression	**Reading**	
-3	menos 3	
44-3	44 menos 3	
44-3=	44 menos 3 igual a	
44-3=41	44 menos 3 igual a 41	
15-20 Outubro	15 a 20 de Outubro	
6-10 Nov.	6 a 10 de Novembro	[not SP]
1998-2004	mil novecentos e noventa e oito a dois mil e quatro	[not SP]
ultra-sensível	ultra sensível	

4.3.2. Asterisk and 'x'

Asterisk '*' is pronounced *asterisco*.

'*x*' is pronounced *vezes* if it is in a mathematical equation with an equals sign. In other cases it is pronounced *xis*.

Expression	**Reading**
2x3=6	dois vezes três igual a seis
xbc	x b c
2*3	dois asterisco três
*bc	asterisco b c

Chapter 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
Telephone numbers

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
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 999999999999 (twelve digits). Numbers higher than this are read as separate digits.

Number	**Reading**
2580	dois mil quinhentos e oitenta
2 580	"
2.580	"
25800	vinte cinco mil e oitocentos
25 800	"
25.800	"

Number	Reading
2580350	dois milhões quinhentos e oitenta mil trezentos e cinquenta
2 580 350	"
2.580.350	"
1000000000	mil milhões
23 456 789 012	vinte e três mil milhões quatrocentos e cinquenta e seis milhões setecentos e oitenta e nove mil e doze
1234567890123	um dois três quatro cinco seis sete oito nove zero um dois três

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 nove mil duzentos e cinquenta e três
020	zero vinte

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 the section *Full number pronunciation* . If the decimals (the part after comma) 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 the section *Full number pronunciation* .

Number	Reading
16,234	dezasseis vírgula duzentos e trinta e quatro
3,1415	três vírgula um quatro um cinco
1251,04	mil duzentos e cinquenta e um vírgula zero quatro
1.251,04	mil duzentos e cinquenta e um vírgula zero quatro
2,50	dois vírgula cinquenta
2.50	dois ponto cinquenta
3.141	três mil cento e quarenta e um

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 currency amounts.
- Numbers with zero or two decimals followed by the words *libra, dólar, iene* or *euro* (singular or plural) are read as currency amounts.
- Accepted decimal markers are comma and full stop.
- The decimal part (consisting of two digits) in currency amounts is read as *e nn pence*, and *e nn cêntimos*.

- If the decimal part is *00* it will not be read.

Expression	Reading	
$15.00	quinze dólares	
15.00£	quinze libras	
15.00 euros	quinze euros	[not SP]
€ 200.50	duzentos euros e cinquenta cêntimos	
1.000.000 ¥	um milhão de ienes	

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

$ 1 milhão	um milhão de dólares

5.5. Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is followed by *o(s), a(s), º, ª*.

Expression	Reading
5o	quinto
6a-feira	sexta-feira
3ª	terceira
7º	sétimo
5os	quintos
6as	sextas

5.6. Arithmetic operators

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

Expression	Reading
-12	menos doze
+19	mais dezanove
2x3=6	dois vezes três igual a seis
6/3=2	seis a dividir por três igual a dois
25%	vinte cinco porcento
3,4%	três vírgula quatro porcento

Observe that *x* is pronounced *vezes* if enclosed by digits and followed by equals sign '='. In other cases it is pronounced *xis*.

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.

Expression	Reading
77B84	setenta e sete B oitenta e quatro
0092B87-B	zero zero noventa e dois B oitenta e sete 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. The time words *pm* and *am* are recognized when occurring after an hour alone. [not SP] When there are minutes and seconds, time words are recognized in front of the format, as well as after.

Possible patterns are:

a. *hh:mm* or *h:mm*

b. *hh:mm:ss* or *h:mm:ss*

c. [not SP] *hhHmm* or *hhhmm* (*H* may be in upper or lower case)

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

In pattern a:

The word *hora(s)* will be added after the *hh*-part. If the *mm*-part is equal to a number, an *e* will be insterted before *mm*, and *minutos* will be added after it. In other cases an *e* will be added before the *mm*-part and the word *minuto(s)* after.

In pattern b:

The word *hora(s)* will be added after the *hh*-part. The word *minuto(s)* will be added after the *mm*-part. If the *mm*-part is equal to *00*, this part will not be read. An *e* will be inserted before the *ss*-part, and *minutos* will be added after *mm*-part and *segundos* will be added after *ss*-part. If the *ss*-part is equal to *00*, the expression will be read as specified for pattern (a).

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

Expression	**Reading**	
9 A.M.	nove da manhã	[not SP]
9 P.M.	nove da noite	[not SP]
13:20	treze horas e vinte minutos	
3:20	três horas e vinte minutos	
12:00	meio-dia	
00:00	meia-noite	
13:20:20	treze horas vinte minutos e vinte segundos	
3:20:20	três horas vinte minutos e vinte segundos	
12H30	doze horas e trinta minutos	[not SP]
3h30	três horas e trinta minutos	[not SP]

5.9. Dates

5.9.1. Valid formats

The valid formats for dates are:

1	dd.mm.yyyy	dd-mm-yyyy	dd/mm/yyyy
2	dd.mm.yy	dd-mm-yy	dd/mm/yy
3	yyyy.mm.dd	yyyy-mm-dd	yyyy/mm/dd

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.

Type 1:	**Reading**
10-02-2003 or 10-2-2003	dez de Fevereiro de dois mil e três
10.02.2003 or 10.2.2003	"
10/02/2003 or 10/2/2003	"
Type 2:	**Reading**
10-02-03 or 10-2-03	dez de Fevereiro de dois mil e três
10.02.03 or 10.2.03	"
10/02/03 or 10/2/03	"
Type 3:	**Reading**
2003-02-10 or 2003-2-10	dez de Fevereiro de dois mil e três
2003.02.10 or 2003.2.10	"
2003/02/10 or 2003/2/10	"

Other possible formats include:

Segunda-feira, 15 de Janeiro
Terça, 30 de Abril de 1999
Ter, 30 de Abr de 1999 [not SP]
3 de Maio de 1953

5.9.2. Ranges of days and years [not SP]

Ranges of days and years are also supported.

Expression	**Reading**
1998-1999	mil novecentos e noventa e oito a mil novecentos e noventa e nove
1939-45	mil novecentos e trinta e nove a quarenta e cinco
2002/5	dois mil e dois a cinco
14-15 Janeiro	catorze a quinze de Janeiro

5.9.3. Months and days abbreviations [not SP]

Valid abbreviations for months: *Jan, Fev, Abr, Jun, Jul, Set, Out, Nov*, and *Dez*.

Valid abbreviations for days: *Seg, Ter, Qua, Qui, Sex, Sáb*, and *Dom*.

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

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 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:

- (x) xxx.xxxx
- (xx) xxx.xxxx
- (xxx) xxx.xxxx
- 2xx xxx xxx
- xx – xx xx xx xx
- 0x – xxxxxx
- 093x xxx.xxxx
- 9xx xxx xxx
- 0800 xxx xxx
- 800-xxx-xxxx
- 0800 xxx xxxx
- 1800 xxx xxx
- 00 800 xxxx xxxx

The following sequences can be preceded by the special numbers *646*, *800*, *808*, *707*, and *760*:

- special_number xxx xxx

Any sequence composed of three groups of digits separated by hyphens is treated as a phone number format.

Example:
568-123548-12

5.10.2. International phone numbers

International phone numbers follow the patterns below:

International Prefix + *Country code* + *space* or *hyphen* + *Local number* (as seen above)

International prefix:	*00* or +
Country code:	1-3 digits

Examples
0032 625497
0033 (325) 651.2462
00351 265 254 367

Other valid international formats are:

International prefix + *country code* + *space or hyphen* + *any of the patterns below*

- 2xxxxxxxx
- 9xxxxxxxx
- 6xxxxx
- 6xxxxx000
- special_numberxxxxxx

Chapter 6. How to change the pronunciation

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 *Portuguese Phonetic Text*). Phonetic transcriptions can also be entered directly in the text, using the *PRN* tag (see *User's guide*).

Chapter 7. Portuguese Phonetic Text

The Portuguese text-to-speech system uses the Portuguese subset of the SAMPA phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*), with the exception of the symbols */a~/, /uj/, /6j/, /6~w~/* that have been added. The symbols are written with a space between each phoneme.

Only the symbols listed here 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. Symbols for the Portuguese consonants

Table 7.1. Symbols for the Portuguese consonants

Symbol	Word	Phonetic text	Comment
w	água	a1 G w 6	
j	ciência	s j e~1 s j 6	
p	pai	p aj1	
t	tenho	t 6j1 J u	
k	com	k o~	
b	barco	b a1 r k u	
d	doce	d o1 s @	
g	grande	g r 6~1 d @	
f	falo	f a1 l u	
v	verde	v e1 r d @	
s	céu	s Ew1	
z	casa	k a1 z 6	
S	chapéu	S 6 p Ew1	
Z	jóia	Z Oj1 6	
l	lanche	l 6~1 S @	
L	trabalho	t r 6 B a1 L u	
r	caro	k a1 r u	
R	rua	R u1 6	
m	mar	m a1 r	
n	nada	n a1 D 6	
J	vinho	v i1 J u	
W	mil	m i1 W	
B	sabe	s a1 B @	
D	saída	s 6 i1 D 6	
G	seguir	s @ G i1 r	

7.2. Vowels

Table 7.2. Symbols for the Portuguese vowels

Symbol	Word	Phonetic text	Comment
a	falo	f a1 l u	
@	felizes	f @ l i1 z @ S	
6	cama	k 61 m 6	
e	fazer	f 6 z e1 r	
E	belo	b E1 l u	
i	lápis	l a1 p i S	
o	lobo	l o1 B u	
O	porta	p O1 r t 6	
u	justo	Z u1 S t u	
a~	casa antiga	k a1 Z a~ t i1 G 6	
i~	fim	f i~1	
e~	emprego	e~ p r e1 G u	
o~	bom	b o~1	
u~	um	u~	
6~	andar	6~ D a1 r	
aw	aula	aw1 l 6	
iw	cumpriu	k u~ p r iw1	
ew	debateu	d @ B 6 t ew1	
Ew	véu	v Ew1	
aj	mais	m aj1 S	
6j	seis	s 6j1 S	
Ej	réis	R Ej1 S	
Oj	jóias	Z Oj1 6 S	
oj	noivo	n oj1 v u	
uj	azuis	6 z uj1 S	
6~j~	cabem	k a1 B 6~j~	
o~j~	noções	n u s o~j~1 S	
u~j~	muita	m u~j~1 t 6	
6~w~	obrigam	O B r i1 G 6~w~	

7.3. Lexical accent

A lexical accent is used to indicate the level of prominence (or emphasis) of a syllable in a word. Practically all words in Portuguese have a lexical accent even if it does not always serve to differentiate between two different words. It is therefore important to include stress marks when writing phonetic transcriptions.

In the phonetic transcriptions, the lexical accent is indicated by the symbol *1* placed directly after (no space) the accented vowel.

7.4. Pause

An underscore /_/ in a phonetic transcription generates a small pause.

Chapter 8. Abbreviations

In the current version of the Portuguese text-to-speech system, the abbreviations in the table below are recognised in all contexts. These abbreviations are case-insensitive 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.

Table 8.1. Abbreviations

Abbreviation	Reading	Comment
esq.	esquerdo	
dto	direito	
fte	frente	
av	avenida	
exmo	excelentíssimo	
exma	excelentíssima	
exmos	excelentíssimos	
exmas	excelentíssimas	
dr	doutor	
v.exa	vossa excelência	
ext	extensão	
n°	número	
n°s	números	
sr	senhor	
sra	senhora	
srta	senhorita	
etc	et cetera	
eng.o	engenheiro	
eng.a	engenheira	
s. ex.a	sua excelência	
vol.	volume	
dig.mo	digníssimo	
mons.	monsenhor	
p.f.	por favor	
s.f.f.	se faz favor	
obg.	obrigado	
prof.	professor	
prof.a	professora	
sto.	santo	
sta.	santa	
NB	note bem	
cf.	confira	
cit.	citado	
pág.	página	

Abbreviation	Reading	Comment
tel.	telefone	
°C	graus centígrados	when preceded by digit
°F	graus Fahrenheit	when preceded by digit
°R	graus Réamur	when preceded by digit

Chapter 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 *ponto*, hyphens '-' as *traço*, underscore '_' as *underscore*, slash '/' as *barra*.
- *pt, uk, fr* and all the other abbreviations for countries are spelled out letter by letter.
- The @ is read *arroba*.
- 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.acapela-group.com	w w w ponto acapela traço ponto grup com
http://www.acapela-group.com	h t t p dois pontos barra barra w w w ponto acapela traço grup ponto com
pereira@netcabo.pt	pereira arroba netcabo ponto p t
diogo_silva@netcabo.pt	diogo underscore silva arroba netcabo ponto p t