



Language Manual

HQ Catalan

Language Manual: HQ Catalan

Published 22 March 2011

Copyright © 2010-2011 Acapela Group

All rights reserved

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

1. General	1
2. Letters in orthographic text	2
3. Punctuation characters	3
3.1. Comma, colon and semicolon	3
3.2. Quotation marks	3
3.3. Full stop	3
3.4. Question mark	3
3.5. Exclamation mark	3
3.6. Parentheses, brackets and braces	3
4. Other non alphanumeric characters	4
4.1. Non-punctuation characters	4
4.2. The ² and ³ signs	4
4.3. Symbols whose pronunciation varies depending on the context	5
5. Number Processing	6
5.1. Full number pronunciation	6
5.2. Leading zero	7
5.3. Decimal numbers	7
5.4. Currency amounts	7
5.5. Ordinal numbers	8
5.6. Roman numbers	8
5.7. Arithmetic operators	9
5.8. Mixed digits and letters	9
5.9. Time of day	9
5.10. Dates	10
5.11. Telephone numbers	11
6. How to change the pronunciation	13
7. Catalan Phonetic Text	14
7.1. Consonants	14
7.2. Vowels	15
7.3. Foreign phonemes	15
7.4. Lexical accent	15
7.5. Pause	15
8. Abbreviations	16
9. Web-addresses and e-mail	18

List of Tables

4.1. Non-punctuation characters	4
7.1. Symbols for the Catalan consonants	14
7.2. Symbols for the Catalan vowels	15
7.3. Foreign phonemes	15
8.1. Abbreviations and Symbols	16

Chapter 1. General

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

This version of the document relates to the High Quality (HQ) voice Laia.

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

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 from A-Z, a-z, as well as à, é, è, í, ó, ò, ú, ü may constitute a word.

Characters used in other languages, e.g. å, ð, are mapped into readable characters, for instance å is read as a.

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: , : ; “ ” . ¿ ? ¡ ! () { } [] ' .

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 '""' and ""'' 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. Following the Catalan usage, we will only use the closing question mark "?". The opening question mark causes a brief pause before and after the quoted text.

3.5. Exclamation mark

The closing exclamation mark '!' is treated in a similar manner to the full stop, causing a falling intonation pattern followed by a pause. Following the Catalan usage, we will only use the closing exclamation mark. The opening exclamation mark causes a brief pause before and after the quoted text.

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 obliqua
	barra vertical
+	més
\$	dòlar
£	lliura
€	euro
¥	ien
<	menor que
>	major que
%	per cent
^	accent circumflex
~	titlla
@	arrova
=	igual a
²	See below
³	See below
-	See below
*	See below

4.2. The ² and ³ signs

The reading of expressions with ² and ³ is:

Expression	Reading
mm ²	mil·límetres quadrats
cm ²	centímetres quadrats
m ²	metres quadrats
km ²	quilòmetres quadrats
mm ³	mil·límetres cúbics
cm ³	centímetres cúbics
m ³	metres cúbics
km ³	quilòmetres cúbics

4.3. Symbols whose pronunciation varies depending on the context

4.3.1. Hyphen

A hyphen '-' is pronounced *menys* if it is part of a mathematical expression or as *guió mig* in some other cases. [not SP] In certain date formats, in between days or years, the hyphen is pronounced *a*. In other cases the hyphen is never pronounced.

Expression	Reading	
44-3	44 guió mig 3	
44-3=41	44 menys 3 igual a 41	
gener 12-14	gener 12 a 14	[not SP]
feb 6-10	febrer 6 a 10	[not SP]
1998-2004	mil nou-cents noranta-vuit a dos mil quatre	[not SP]
02-02-2002	dos de febrer de dos mil dos	
ex-ministre	ex ministre	

4.3.2. Asterisk

Asterisk '*' is pronounced *per* if enclosed by digits and with the sign =. In other cases it is pronounced *asterisc*.

Expression	Reading
2*3	dos asterisc tres
2*3=6	2 per 3 igual a 6
*bc	asterisc 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
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 from 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 cardinal number read is 999999999999 (12 digits). Numbers higher than this are read as separate digits.

Number	Reading
2580	dos mil cinc-cents vuitanta
2 580	"
2.580	"
25800	vint-i-cinc mil vuit-cents
25 800	"
25.800	"

Number	Reading
2580350	dos milions cinc-cents vuitanta mil tres-cents cinquanta
2 580 350	"
2.580.350	"
1000000000	mil milions
123456789012	cent vint-i-tres mil quatre-cents cinquanta-sis milions set-cents vuitanta-nou mil dotze
2123456789012	dos u dos tres quatre cinc sis set vuit nou zero u dos

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 nou mil dos-cents cinquanta-tres
020	zero vint

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 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 the section *Full number pronunciation*.

Number	Reading
16,234	setze coma dos-cents trenta-quatre
3,1415	tres coma u quatre u cinc
1251,04	mil dos-cents cinquanta-u coma zero quatre
1.251,04	mil dos-cents cinquanta-u coma zero quatre
2,50	dos coma cinquanta
2.50	dos punt cinquanta
3.141	tres mil cent quaranta-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 *pesseta*, *pta*, *lliura*, *dòlar*, *ien* 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 *amb nn penics* and *amb nn cèntims*.
- If the decimal part is *00* it will not be read.

Example	Reading	
\$15.00.	quinze dòlars	
15.00£.	quinze lliures	
15.00 euros.	quinze euros	[not SP]
€ 200.50	dos-cents euros amb cinquanta cèntims	
1.000.000 ¥	un milió de iens	

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

\$ 1 milió un milió de dòlars

5.5. Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is *1r, 2n, 3r, 4t*.
- The number after '4' is followed by *è, ena*.

Example	Reading
1r	primer
2n	segon
3r	tercer
4t	quart
5è	cinquè
5ena	cinquena
6è	sisè
6ena	sisena

5.6. Roman numbers

Numbers are read as roman numbers in the following cases:

- Following the Catalan usage the numbers 1 to 10 will be read as roman numbers.

Expression	Reading
Carles II	carles segon
Lluís III	lluís tercer
Carles IV	carles quart
Felip VII	felip setè
Guifré VIII	guifré vuitè
Ferran IX	ferran novè

Note that the system will not recognise *I, V, VI, X* as roman numbers because they can also be the capital letters of: *i, v, vi, x*.

- From 11 and up numbers will be read as cardinals, following the catalan usage.

5.7. Arithmetic operators

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

Expression	Reading
-12	menys dotze
+24	més vint-i-quatre
2*3=6	dos per tres igual a sis
2/3=0.67	dos dividit per tres igual a zero punt seixanta-set
2/3	dos terços
25%	vint-i-cinc per cent
3,4%	tres coma quatre per cent

5.8. 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
77B84Z3	setenta-set B vuitanta-quatre Z tres
0092B87-B	zero zero noranta-dos B vuitanta-set B
12.345.678-V	dotze milions, tres-cents curanta cinc mil sis-centes setanta vuit V
X1234567F	X un milió dos-cents trenta-quatre mil cinc-centes seixanta-set F
6328 F X Y	sis mil tres-cents vint-i-vuit F X Y

5.9. Time of day

The colon is used to separate hours, minutes and seconds. Possible time formats are:

- a. *hh:mm* or *h:mm*
- b. *hh:mm:ss* or *h:mm:ss*
- c. *hh* or *hh-hh*

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

In pattern a:

If the *mm*-part is equal to *00*, this part will not be read. [not SP] This pattern can be preceded or followed by time indications such as *A.M.*, *AM*, *a.m.*, *am*, *P.M.*, *PM*, *p.m.*, or *pm*. The abbreviations *h* and *h.* can follow the pattern. *i* will be inserted before the *mm*-part.

In pattern b:

After the *hh*-part *hores* will be added. *i* will be inserted before the *mm*-part, and *minuts* will be added after it. After the *ss*-part, *segons* will be added. If the *ss*-part is equal to *00*, this part will not be read. [not SP] This pattern can be preceded or followed by time indications such as *A.M.*, *AM*, *a.m.*, *am*, *P.M.*, *PM*, *p.m.*, or *pm*.

In pattern c:

[not SP] The hours can appear alone, but must be followed by time indications, such as: *A.M.*, *AM*, *a.m.*, *am*, *P.M.*, *PM*, *p.m.* or *pm*. [not SP] They can also appear in a time range and must then be followed by a time indication.

[not SP] *A.M.*, *AM*, *a.m.*, or *am* is read as *de la matinada*, *del matí* or *del migdia*.

[not SP] *P.M.*, *PM*, *p.m.*, or *pm* is read as *de la tarda*, *de la nit*.

[not SP] The day time ranges are defined in the following way:

Part of day	Time range
<i>matinada</i> (down)	1:00-5:59
<i>matí</i> (morning)	6:00-11:59
<i>migdia</i> (midday)	12:00-12:59
<i>tarda</i> (afternoon)	13:00-19:59
<i>nit</i> (night)	20:00-00:59

Some examples:

Expression	Reading	
3:45 AM	tres curanta-cinc de la matinada	[not SP]
6 a.m.	sis del matí	[not SP]
12.30 am	dotze trenta del migdia	[not SP]
2:30 p.m.	dos trenta de la tarda	[not SP]
10:15 PM	deu quinze de la nit	[not SP]

5.10. Dates

The valid date formats are:

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

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. [not SP] Dates expressed as *dd/mm* are only recognized if preceded by the word "dia".

Examples of valid formats and their readings:

Type 1:

10-02-2003 or 10-2-2003	deu de febrer de dos mil tres
10.02.2003 or 10.2.2003	“
10/02/2003 or 10/2/2003	“

Type 2:

10-02-03 or 10-2-03	deu de febrer de dos mil tres
10.02.03 or 10.2.03	“
10/02/03 or 10/2/03	“

Type 3:	[not SP]
dia 25/12	dia vint-i-cinc de desembre

[not SP] Ranges of days and years are also supported.

Expression	Reading	
1998-1999	mil nou-cents noranta-vuit a mil nou-cents noranta-nou	[not SP]
1939-45	mil nou-cents trenta-nou a mil nou-cents quaranta-cinc	[not SP]
2002/3	dos mil dos a dos mil tres	[not SP]
14-15 gener	catorze a quinze de gener	[not SP]
abril 2-3	abril dos a tres	[not SP]
dia 10/04	dia deu d'octubre	[not SP]

Other possible date formats include:

Expression	Reading	
Dilluns, 15 de gener	dilluns quinze de gener	
Dijous, 30 d'abril de 1999	dijous trenta d'abril de mil nou-cents noranta-nou	
3 de maig de 1953	tres de maig de mil nou-cents cinquanta-tres	
15 de set. de 2009	quinze de setembre de dos mil nou	[not SP]
2 d'ag. de 1998	dos d'agost de mil nou-cents noranta-vuit	[not SP]

[not SP] Abbreviations of months and days in date formats:

Months:

gen, gen., febr, febr., abr, abr., jul, jul. ag, ag., set, set., oct, oct., nov, nov., des, des.

Days:

dt., dc., dj., dv., ds.

5.11. Telephone 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 section *Leading zero*) with pauses between groups of numbers. Groups that contain more than three digits are read out digit by digit.

5.11.1. Local phone numbers

Expression	Reading
93 321 24 25	noranta-tres tres-cents vint-i-u vint-i-quatre vint-i-cinc
972 456 789	nou-cents setanta-dos milions quatre-cents cinquanta-sis mil set-cents vuitanta-nou
93 321 2425	noranta-tres tres-cents vint-i-u dos quatre dos cinc
972 45 67 89	nou-cents setanta-dos cuarenta-cinc seixanta-set vuitanta-nou

5.11.2. International phone numbers

International phone numbers follow the pattern below:

International prefix + Country code + space or parenthesis + Local number.

International prefix: 00 or +
Country code: 1-3 digits
Local number: 6-9 digits

Expression	Reading	
0034 (971) 123-4567	zero zero tres quatra noucents setanta-u cent vint-i-tres quatre cinc sis tres	[not SP]
0034 971 123456	“	[not SP]
001 21- 123-45-56	zero zero u cent vint-i-tres curanta-cinc cinquanta-sis	[not SP]
+34 971 12 34 56	més trenta-quatre nou-cents setanta-u dotze trenta-quatre cinquanta-sis	
+34 971 123 456	més trenta-quatre nou-cents setanta-u cent vint-i-tres quatre-cents cinquanta-sis	[not SP]
(+34) 971 12 34 56	més trenta-quatre nou-cents setanta-u dotze trenta-quatre cinquanta-sis	
+34 93 123 4567	mes trenta-quatre noranta-tres cent vint-i-tres quatre cinc sis set	[not SP]
+34 93 123 45 67	mes trenta-quatre noranta-tres cent vint-i-tres curanta-cinc seixanta-set	
0034 93 123 45 67	zero zero trenta-quatre noranta-tres cent vint-i-tres curanta-cinc seixanta-set	
(+34) 93 123 45 67	mes trenta-quatre noranta-tres cent vint-i-tres curanta-cinc seixanta-set	

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

Chapter 7. Catalan Phonetic Text

The Catalan text-to-speech system uses the Catalan subset of the SAMPA phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*). 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. Consonants

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

Table 7.1. Symbols for the Catalan consonants

Symbol	Word	Phonetic text
p	pare	p a1 r @
t	terra	t E1 rr @
k	casa	k a1 z @
	Quelcom	k @ l k o1 m
b	bala	b a1 l @
d	dona	d O1 n @
g	goma	g o1 m @
	Guillem	g i L E1 m
f	foc	f O1 k
tS	andratx	@ n d r a1 tS
	Puig	p u1 tS
dZ	platja	p l a1 dZ @
s	sal	s a1 l
z	rosa	rr O1 z @
S	caixa	K a1 S @
Z	Jordi	Z O1 r d i
l	laberint	l @ B @ r i1 n
L	coll	K O1 L
	Dall	d a1 L
r	cara	k a1 r @
rr	roca	rr O1 k @
m	molí	m u l i1
n	nata	n a1 t @
J	canya	k a1 J @
	senyor	s @ J o1
N	sang	s a1 N
B	alba	a1 l B @

Symbol	Word	Phonetic text
D	seda	s E1 D @
G	agafar	@ G @ f a1
w	pau	p a1 w
j	aire	a1 j r @
jj	Moya	m o1 jj a for Spanish words
T	Mercedes	m e r T e1 D e s for Spanish words
x	ojos	o1 x o s for Spanish words

7.2. Vowels

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

Table 7.2. Symbols for the Catalan vowels

Symbol	Word	Phonetic text
a	sala	s a1 l @
e	vent	b e1 n
E	terra	t E1 rr @
i	crisi	k r i1 z i
o	senyor	s @ J o1
O	porta	p O1 r t @
u	fum	f u1 m
@	cosa, pere	k O1 z @ p e1 r @

7.3. Foreign phonemes

Table 7.3. Foreign phonemes

Symbol	Word	Phonetic text
W	winona	W i n o1 n a
h	handicap	h a1 n d i k @ p
v	Vancouver	b @ N k u1 v @ r

7.4. Lexical accent

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. For the correct pronunciation of a word, it is important to include the symbol marking the word stress.

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

7.5. Pause

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

Chapter 8. Abbreviations

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

Table 8.1. Abbreviations and Symbols

Abbreviation/Symbol	Reading
Exc.	excel·lència
SA*	Societat Anònima
sr.	senyor
srs.	senyors
sres.	senyores
sra.	senyora
srtà.	senyoreta
TVE*	Televisión Española
cl.	centilitres
cm.	centímetres
cts.	cèntims
dta.	dreta
dl.	decilitres
dm.	decímetres
dupl.	duplicat
etc.	etcètera
esq.	esquerra
kg.	kilograms
km.	kilòmetres
mg.	mil·ligrams
ml.	mil·lilitres
mins	minuts
mm.	mil·límetres
núm.	número
orig.	original
pral.	principal
prov.	província
tel.	telèfon
°C *	graus celsius
°K *	graus kelvin
°F *	graus fahrenheit
apt.	apartat
aving. (o av. avd. avda)	avinguda
dr.	doctor
dra.	doctora

Abbreviation/Symbol	Reading
cant.	cantonada
gov.	Govern
gral.	General
eng.	enginyeria
pàg.	pàgina
prof.	professor
vol.	Volum
NIF*	NIF

For month and day abbreviations, see section *Dates* .

Chapter 9. Web-addresses and e-mail

Web-addresses and e-mail-addresses are read as follows:

- *www* is read as three *w*'s spelled letter by letter.
- Full stops '.' are read as *punt*, hyphens '-' as *guió mig*, underscore '_' as *guió baix*, slash '/' as *barra obliqua*.
- *es*, *uk*, *fr* and all the other abbreviations for countries are spelled out letter by letter.
- The '@' is read *arrova*.
- 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	be doble, be doble, be doble punt acapela guió mig grup punt com
http://www.acapela-group.com	h t t p dos punts barra obliqua barra obliqua be doble, be doble, be doble punt acapela guió mig grup punt com
garcia@yahoo.es	garcia arrova yahoo punt e esa
ana_garcia@yahoo.es	ana guió baix garcia arrova yahoo punt e esa