Language Manual

Brazilian Portuguese

Carlos
Paola

Text-to-Speech Converter
Language Manual
Brazilian Portuguese
September 2005, modified 11 April 2007

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 Labège cedex
France

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

www.acapela-group.com

1 General

This document discusses certain aspects of text-to-speech processing for the Brazilian 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 voice nom_de_la_voix.

2 Letters in orthographic text

Characters from **A-Z** and **a-z**, as well as “ç, ñ, õ, ã, à, á, â, é, ê, í, ó, ô, ú, ü” may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, i.e. “å, î”. These letters are not pronounced as in their native languages though, they are pronounced as regular “a, i” 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
/	barra
+	mais
$	dólar
£	libra
€	euro
¥	yeni
<	menor que
>	maior que
%	porcento
^	acento circuflecho
|	barra vertical
~	til
@	arroba
²	quadrado
³	cúbico
=	igual
-	See below
*	See below

Table 1 Non-punctuation characters

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	quilometros quadrados
mm³	milímetros cúbicos
cm³	centímetros cúbicos
m³	metros cúbicos
km³	quilometros cúbicos

4.3 Symbols whose pronunciation varies depending on the context

4.3.1 Hyphen

A hyphen < **-** > is pronounced "menos" if it is in the beginning of a line and followed by a digit, or if it is in a mathematical equation with an equals sign. 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=41	44 menos 3 igual 41
15-20 outubro	15 a 20 de outubro
6-10 nov	6 a 10 de novembro
1998-2004	mil novecentos e noventa e oito a dois mil e quatro
02-02-2002	dois de fevereiro de dois mil e dois
ultra-sensível	ultra sensível

4.3.2 Asterisk

Asterisk < ***** > is pronounced "veces" if it is in a mathematical equation with an equals sign. In other cases it is pronounced "asterisco".

Expression	**Reading**
2*3=6	dois veces três igual seis
*bc	asterisco 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

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 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	“
2580350	dois milhões quinhentos e oitenta mil trezentos e cinqüenta
2 580 350	“
2.580.350	“
1000000000	um bilhão
1234567890123	um dois três quatro cinco seis sete oito nove zero um dois três
23 456 789 012	vinte e três bilhões quatrocentos e cinqüenta e seis milhões setecentos e oitenta e nove mil doze

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 cinqüenta 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 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	dezesseis virgula duzentos trinta e quatro
3,1415	três virgula um quatro um cinco
1251,04	mil duzentos e cinqüenta e um virgula zero quatro
1.251,04	mil duzentos e cinqüenta e um virgula zero quatro
2,50	dois virgula cinqüenta
2.50	dois ponto cinqüenta
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 monetary amounts.
- Numbers with zero or two decimals followed by the words "libra", "dólar", "yeni" or "euro" (singular or plural) are read as monetary amounts.
- Accepted decimal markers are comma and full stop**.**
- The decimal part (consisting of two digits) in monetary amounts is read as "e nn peniques" and "e nn centavos".
- If the decimal part is "00" it will not be read.

Example	**Reading**
$15.00.	quinze dólares
15.00£.	quinze libras
15.00 euro.	quinze euro
€ 200.50	duzentos euros e cinqüenta centavos
1.000.000 **¥**	um milhão de yenes

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 "1" is followed by "de" and 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 de janeiro, 1 de jan, terça-feira 1 de jan.
- The number is followed by "o(s), a(s), º, ª". Examples: 5o, 6a, 3ª, 7º.

Valid abbreviations for months: jan, fev, abr, jun, jul, set, out, nov and dez.

Valid abbreviations for days: segunda, terça, quarta, quinta, sexta.

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 and an equals sign are read according to the examples below.

Expression	**Reading**
-12	menos doze
+19	mais dezenove
2*3=6	dois veces três igual seis
6÷3=2	seis a dividir por três igual dois
25%	vinte cinco porcento
3,4%	três virgula quatro porcento

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**
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. 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 "e" will be inserted before the "ss"-part, and "segundos" 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
3.yyyy.mm.dd, yyyy-mm-dd and 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.

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	dez de fevereiro de dois mil e três
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	dez de fevereiro de dois mil e três
10.02.03	or	10.2.03	“
10/02/03	or	10/2/03	“
Type 3:	yyyy.mm.dd, yyyy-mm-dd and yyyy/mm/dd		
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	“

Ranges of days and years are also supported.

Examples:

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
2002/3	dois mil e dois a três
14-15 janeiro	quatorze a quinze de janeiro

Other possible formats include :

- segunda-feira, 15 de janeiro
- terça, 30 de abril de 1999
- 3 de maio de 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 spelled out, with pauses between groups of numbers.

5.10.1 Ordinary phone numbers

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

The following formats can have a potential space after the closing bracket:

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

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

The following formats are also recognized:

- xx xxx xxxx
- xx xx xxxx-xxxx

Any sequence composed of three groups of digits separated by hyphens is treated as a phone number format (with the exception of the above date formats)..

Ex 568-123548-12

5.10.2 International phone numbers

International phone numbers follow the patterns below:
International Prefix + Country code + Local number (as seen above)

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

The international prefix, country code, and local number may be separated by a space or a hyphen.

Examples:

00 32 (12)123-4567
00 33 (25 35) 5834-2850
00-33-25 648 3695

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 a PRN-tag (see User's guide).

7 Brazilian Portuguese Phonetic Text

The Brazilian Portuguese text-to-speech system uses an alphabet inspired from the SAMPA phonetic alphabet (Speech Assessment Methods Phonetic Alphabet).
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 Brazilian Portuguese consonants

Symbol	Word	Phonetic text	Comment
w	**aboliu**	\a b o l i1 w\	
y	**ciência**	\s i em1 s y a\	
p	**pai**	\p a1 y\	
t	**tenho**	\t e1 nh u\	
k	**com**	\k om\	
b	**barco**	\b a1 rr k u\	
d	**doce**	\d o1 s i\	
g	**grande**	\g r am1 d i\	
f	**falo**	\f a1 l u\	
v	**verde**	\v e1 rr d i\	
s	**céu**	\s ee1 u\	
z	**casa**	\k a1 z a\	
x	**chapéu**	\x a p ee1 u\	
j	**jóia**	\j oo1 y a\	
ss	**abonos**	\a b o1 n u ss\	
l	**labor**	\l a b o1 rr\	
lh	**trabalho**	\t r a b a1 lh u\	
r	**caro**	\k a1 r u\	
rr	**rua**	\rr u1 a\	
m	**mar**	\m a1 rr\	
n	**nada**	\n a d a\	
nh	**vinho**	\v i1 nh u\	

Table 2 Brazilian Portuguese consonants

7.2 Vowels

7.2.1 Symbols for the Brazilian Portuguese vowels

Symbol	Word	Phonetic text	Comment
a	**falo**	\f a1 l u\	
@	**ladrão**	\l a d r @1 w\	
e	**fazer**	\f a z e1 rr\	
ee	**belo**	\b ee1 l u\	
i	**lápis**	\l a1 p i ss\	
o	**lobo**	\l o1 b u\	
oo	**doc**	\d oo1 k\	
u	**jus**	\j u1 ss\	
am	**largam**	\l a1 rr g am \	
im	**fim**	\f im1\	
em	**emprego**	\em p r e1 g u\	
om	**bom**	\b om1\	
um	**um**	\um\	

Table 3 Brazilian Portuguese vowels

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 Brazilian 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.

8 Abbreviations

In the current version of the Brazilian Portuguese text-to-speech system, the abbreviations in table 4 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.

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

Abbreviation	Reading
n°	numero
n°s	numeros
a.C. *	antes de kristo
d.C. *	depois de kristo

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 "ponto", hyphens as "traço", underscore ("_") as "soblinhado", slash ("/") as "barra".
- "br, 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.com	w w w ponto acapela ponto com
http://www.acapela.com	h t t p dois pontos barra barra w w w ponto acapela ponto com
rios@yahoo.br	rios arroba yahoo ponto b r
diego_rios@yahoo.br	diego soblinhado rios arroba yahoo ponto b r