Language Manual

British English

Peter, Graham, and Lucy

Language Manual
British English
Peter, Graham, and Lucy
19 February 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
Immeuble Les Erables
Rue du Colombier
ZAC de l'Hers
F-31670 Labege
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 British English 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 voices Peter, Graham, and Lucy.

2 Letters in orthographic text

Characters from **A-Z** and **a-z** may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, i.e. “ñ, ç, é”. These letters are not pronounced as in their native languages though, they are pronounced as regular “n, c, e” 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

Parentheses < **( )** > 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
£	pound
€	euro
¥	yen
<	less than
>	greater than
%	percent
^	circumflex
|	pipe
~	tilde
@	at
=	equals
²	see below
³	see below
-	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	millimeters squared
cm²	centimeters squared
m²	meters squared
km²	kilometers squared
mm³	millimeters cubed
cm³	centimeters cubed
m³	meters cubed
km³	kilometers cubed

4.3 Symbols whose pronunciation varies depending on the context

4.3.1 Hyphen

A hyphen < **-** > is pronounced “minus” in two cases:

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

b) if followed by a digit and an equals sign. If there is no equals sign, it is pronounced “dash”.

In certain date formats, in between days or years, the hyphen is pronounced “to”.
In compounds or between words, the hyphen is not pronounced. Examples: low-income, mother-in-law, pre- and post-war Europe.

Expression	**Reading**
-3	minus three
44-3	forty-four dash three
44-3=41	forty-four minus three equals forty-one
15-20 October	the fifteenth to twentieth of October
6-10 Nov	the sixth to tenth of November
1998-2004	nineteen ninety-eight to two thousand and four
02-02-2002	the second of February two thousand and two

4.3.2 Asterisk

Asterisk < ***** > is only pronounced as “multiplied by” if enclosed by digits and followed by equals sign. In other cases it is pronounced as “asterisk”.

Expression	**Reading**
2*3	two asterisk three
2*3=6	two multiplied by three equals six
*bc	asterisk 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 familiarize the user with the various types of formatted and non-formatted strings of digits that are recognized by the system, a brief description of the basic number processing is provided below, 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
Year
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
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 comma (not full stop). 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	two thousand five hundred and eighty
2 580	“
2,580	“
25800	twenty-five thousand eight hundred
25 800	“
25,800	“
2580350	two million five hundred eighty thousand three hundred and fifty
2 580 350	“
2,580,350	“
1000000000	one billion

123456789012	one two three four five six seven eight nine zero one two
23 456 789 012	twenty-three billion four hundred fifty-six million seven hundred eighty-nine thousand and twelve

5.2 Leading zero

Numbers that begin with 0 (zero) are read as a whole number, with a zero preceding it.

Number	**Reading**
09253	zero nine thousand two hundred and fifty-three
020	zero twenty

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. The decimals (the part after comma or full stop) are read as separate digits. Note: A number containing a comma 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	sixteen point two three four
3.1415	three point one four one five
1251.04	one thousand two hundred and fifty-one point zero four
1,251.04	one thousand two hundred and fifty-one point zero four
2.50	two point zero
2,50	two comma zero
3,141	three thousand one hundred and forty-one

5.4 Monetary amounts

The following principles are followed for monetary amounts:

- Numbers with zero or two decimal places preceded or followed by the currency markers **£, $, ¥** or **€** are read as monetary amounts.
- Numbers with zero or two decimal places preceded or followed by the words “pounds”, “dollars”, “yen” or “euros” (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 “and nn pence” or “and nn cents”.
- If the decimal part is “00” it will not be read.

Example	**Reading**
$15.00.	fifteen dollars
15.00£.	fifteen pounds
15.00 euros.	fifteen euros
€ 200.50	two hundred euros and fifty cents
1,000,000 **¥**	one million yen

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

$ 1 million	one million dollars

5.5 Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is followed by a month name or one of the month name abbreviations and the number is smaller or equal to 31. The number may be preceded by a day or an abbreviation for a day. Examples: 3 January, 3 Jan, Tuesday 3 Jan.
- The number consists of a day interval followed by a month name/abbreviation. Example 15-16 January
- The number is followed by "st, nd, rd, th, d". Examples: 1st, 1st, 2nd, 2nd, 3rd, 3rd, 4th, 4th, 23d, 23d.

Valid abbreviations for months: Jan, Feb, Mar, Apr, Jun, Jul, Aug, Sept, Oct, Nov and Dec.

Valid abbreviations for days: Mon, Tue, Wed, Thu, Thurs, Fri, Sat and Sun.

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	minus twelve
+24	plus twenty-four
2*3	two asterisk three
2*3=6	two multiplied by three equals six
2/3	two thirds
25%	twenty-five percent
3.4%	three point four percent

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	seventy-seven B eighty-four Z three
0092B87-B	zero zero ninety-two B eighty-seven B

5.8 Time of day

The colon is used to separate hours, minutes and seconds. Abbreviations such as "A.M." and "P.M." (possible variants: a.m., am, AM, p.m., pm, PM) may follow or precede the time, with a space inserted between the time and the abbreviation.

Examples: 9 A.M.

4 pm

Possible patterns are:

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

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

c) **hh:mm'ss" (or h:mm'ss")** eg. 12:30'45"

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

In pattern a): If the "mm"-part is equal to "00", this part will not be read. Instead, "o'clock" will be added if the hours are less than 13, or "hundred hours" will be added if the hours are greater than or equal to 13.

Example:	9:00	nine o'clock
	13:00	thirteen hundred hours

In pattern b): An "and" will be inserted before the "ss"-part, and "seconds" 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 (b).

5.9 Year

Numbers between 1100 and 1999 are always read as hundreds (year reading) with the exception of numbers containing decimals. Years (2 or 4 digits) can also be followed by "s" or " 's" to indicate decades.

Expression	**Reading**
1988	nineteen eighty-eight
1939-45	nineteen thirty-nine to forty-five
1088	one thousand eighty-eight
1988.0	one thousand nine hundred and eighty-eight point zero
1988.32	one thousand nine hundred and eighty-eight point thirty-two
September 1999	September nineteen ninety-nine
1980s	nineteen eighties
70's	seventies
1980's	nineteen eighties

5.10 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	the tenth of February two thousand and three
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	the tenth of February two thousand and three
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	nineteen ninety-eight to nineteen ninety-nine
1939-45	nineteen thirty-nine to forty-five
2002/3	two thousand and two to three
14-15 January	the fourteenth to fifteenth of January
October 19-20	October the nineteenth to twentieth

Other possible formats include:

Monday, 15 January	(with or without the comma)
Mon, January 15	(with or without the comma)
30 April 1999	
April 30 1999	
May 1953	
3 May	

5.11 Phone numbers

In this section the patterns of digits that are recognized as phone numbers are described. In the pronunciation of phone numbers, all numbers are read out digit by digit with a pause between the groups.

5.11.1 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 or a hyphen:

- xxxxx xxxxxx
- xxxxx xxx xxx
- xxxxx xxxxx
- xxxx xxxxxxx
- xxxx xxx xxxx
- xxxx xxxxxx
- xxx xxxx
- xx xx xx
- xxx xxxx xxxx

- (area) xxxx xxxx
- (area) xxxxxxx
- (area) xxxxxx
- (area) xxxxx
- (area) xxx xxxx
- (area)-xxx-xxxx

(The area code is equal to 0 followed by 2 to 7 digits)

The following sequences can only appear in these formats:

- xxxx/xxx-xxxx
- xxx/xxx-xxx
- xxx-xxx-xxx
- (x)-xxx-xxx
- (xx)-xxx-xxx
- (xxx)-xxx-xxx
- (x).xxxx.xxx.xxx
- (x)-xxxx-xxx-xxx
- (xx).xxxx.xxx.xxx
- (xx)-xxxx-xxx-xxx
- (xxx).xxxx.xxx.xxx
- (xxx)-xxxx-xxx-xxx

The sequence xxx-xxx is recognized as a phone format only if preceded by "tel, mob, tel:, mob:".

5.11.2 International phone numbers

All preceding formats can be recognised if preceded by international prefix:

00x	+xx	00(xxx)
00xx	+xxx	+(x)
00xxx	00(x)	+(xx)
+x	00(xx)	+(xxx)

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

7 British English Phonetic Text

The British English uses the British English subset of 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 British English consonants

Symbol	Word	Phonetic text	Comment
b	bad	b {1 d	
t	stop, tomorrow	s t Q1 p, t @ m Q1 r @U	
t_h	top	t_h Q1 p	
p	sport, potato	s p O:1 t, p @ t_h eI1 t @U	
p_h	pad	p_h {1 d	
d	date	d eI1 t	
k	scone, campaign	s k @U1 n, k @ m p_h eI1 n	
k_h	cone	k_h @U1 n	
g	gag	g {1 g	
m	man	m {1 n	
n	nose	n @U1 z	
r	rose	r @U1 z	
l	let	l e1 t	
L	adult	{1 d V L t	
N	ring	r I1 N	
f	fat	f {1 t	
v	vote	v @U1 t	
s	sat	s {1 t	
z	zoo	z u:1	
S	shin	S I1 n	
tS	chin	tS I1 n	
Z	measure	m e1 Z @	
dZ	gin	dZ I1 n	
D	this	D I1 s	
T	thin	T I1 n	
w	wait	w eI1 t	
j	yacht	j Q1 t	
h	hit	h I1 t	
hj	exhume	e k s hj u:1 m	
hl	Llandaff	hl {1 n d @ f	Welsh allophone
x	loch	l Q1 x	Scottish allophone

Table 2 British English consonants

7.2 Vowels

7.2.1 Symbols for the British English vowels

Symbol	Word	Phonetic text	Comment
A:	father	f A:1 D @	
O:	four	f O:1	
I	bit	b I1 t	
i:	neat	n i:1 t	
u:	zoo	z u:1	
V	hut	h V1 t	
U	put	p U1 t	
{	pat	p {1 t	
e	net	n e1 t	
@	allow	@ l aU1	
eI	main	m eI1 n	
aI	high	h aI1	
OI	boy	b OI1	
@U	nose	n @U1 z	
aU	pout	p aU1 t	
3:	fur	f 3:1	
Q	dot	d Q1 t	
I@	near	n I@1	
e@	there	D e@1	
U@	moor	m U@1	
i	locally	l @U1 k @ l i	
u	punctual	p_h V1 N k tS u @ L	
A~	renaissance	r e n eI1 s A~ s	French vowel
E~	vin	v E~1	French vowel
O~	avignon	{1 v i n j O~	French vowel
l=	battleaxe	b {1 t l= { k s	(only before vowels)
L=	battle	b {1 t L=	(word finally or before of consonants)
m=	grantham	g r {1 n T m=	
n=	sudden	s V1 d n=	
r=	history	h I1 s t r= i	

Table 3 British English vowels

Note that "1" is a stress mark and not part of a consonant or vowel. See section 7.3.

7.3 Lexical accent

A lexical accent is used to indicate the level of prominence (or emphasis) of a syllable in a word. In British English, some words can be differentiated by the position of this lexical accent. The word "record" is an example of this since it can be both a noun (a record: /r e1 k O: d/) or a verb (to record: /r I k_h O:1 d/). Practically all words in British English 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, primary accent is indicated by the symbol "1" placed directly after (no space) the accented vowel. Secondary accent is indicated by the symbol "2". Some examples:

devastating /d e1 v @ s t eI2 t I N/

devastation /d e2 v @ s t eI1 S n=/

devote /d I v @U1 t/

devotee /d e2 v @ t_h i:1/

7.5 Glottal stops

A glottal stop, represented by the phonetic symbol /?/, is a small sound which is often used to separate two words when the second word starts with a stressed vowel. This sound can be inserted in a transcription in order to improve the pronunciation.

7.6 Pause

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

8 Abbreviations

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

As previously mentioned, there are also abbreviations for the days of the week and the months, see chapter 5.6.

Abbreviation	Reading
kg	kilograms
°C	degrees Celsius
°F	degrees Fahrenheit
°K	degrees Kelvin
asap	A S A P
b/f	before
blvd	boulevard
cm	centimeters
corp	corporation
DM*	deutschmark
eg	for example
etc	et cetera
ft	feet
gal	gallons
gov	governor
hr	hour
hrs	hours
ie	that is
jr	junior
km	kilometers
Km/h	kilometers per hour
mg	milligrams
ml	milliliters
mm	millimeters
mph	miles per hour
mr	mister
mrs	missis
ms	miss
mt	mount
prof	professor
sgt	sergeant
sr	senior
tsp	teaspoon
vs	versus
gen	general
ltd	limited
dept	department
ct	court
rd	road
av	avenue
ctrl	control
lb	pound

Table 4 Abbreviations

Some abbreviations are expanded differently depending on their position in the sentence. For example, “dr” and “st” are expanded into “drive” and “street” if they appear after a capitalized noun. They are expanded into “doctor” and “saint” when they appear before a capitalized noun.

Examples	**Readings**
Main st.	Main street
St John.	Saint John
Bayview dr.	Bayview drive
Dr. Jones.	Doctor Jones

“m, g” and “in.” are expanded only when appearing after a number.

Examples	**Readings**	
25 m	twenty-five meters	
30 in.	thirty inches	(note that the period is mandatory here)
45 g	forty-five grams	

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 "dot", hyphens as "dash", underscore ("_") as "underscore", slash ("/") as "slash".
- "us, uk, fr" and all the other abbreviations for countries are spelled out letter by letter.
- The "@" is read "at".
- 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 dot babeltech dot com
http://www.babeltech.com	h t t p colon slash slash w w w dot babeltech dot com
smith@yahoo.us	smith at yahoo dot u s
jane_smith@yahoo.us	jane underscore smith at yahoo dot u s