Language Manual

HQ and HD Dutch

Language Manual: HQ and HD Dutch

Published 16 March 2011

This document was produced by Acapela Group. We welcome any comments or 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 manual describes various aspects relevant to the processing of the written Dutch language that can help users to achieve the desired pronunciation by the Dutch Text-to-Speech system of Acapela. In particular, the manual sums up the different types of characters: such as punctuation, signs, numbers and textual formats, which are allowed as input, and read out in specific ways.

Current version of the document corresponds to the High Quality (HQ) voices Femke, Max, Jasmijn and Daan and the High Density (HD) voices Hilde and Jan.

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 in principle applies to all Acapela Group HQ and HD Dutch voices with some exceptions (see note below). 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 from *A-Z*, *a-z*, as well as *ÂâÄäÇçÈèÉéÊêËëÏïÖöÜü* may constitute a word. Certain other characters are also considered as letters, notably those used in other European languages, i.e. “š, œ, ž, ß, à, á, ã, å, æ, ì, í, î, ð, ñ, ò, ó, ô, õ, ø, ù, ú, û, ý, þ, ÿ”. However, the latter charactes are not pronounced as in their native languages. Instead, when occurring in a word, they are pronounced as regular Dutch “s, o, z, s, a, a, a, a, a, i, i, i, d, n, o, o, o, o, o, u, u, u, i, t, i”.

Any words formed from the characters are read out according to the standard rules of Dutch pronunciation. The words written with typo's will be read out literally.

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 that 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
/	Slash
+	Plus
\	Backslash
$	Dollar
£	Pond
€	Euro
¢	Cent
¥	Yen
<	Kleiner dan
>	Groter dan
%	Procent
‰	Promille
^	Accent circumflex
|	Rechte streep
~	Tilde
_	Underscore
¨	Trema
&	"En"-teken
§	Paragraafteken
@	At
÷	Gedeeld door
×	Maal
±	Plus minus
=	"Is"-teken
®	Geregistreerd merk
©	Copyright
™	Trademark-teken
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
mm²	vierkante millimeter
cm²	vierkante centimeter
m²	vierkante meter
km²	vierkante kilometer
mm³	kubieke millimiter
cm³	kubieke centimeter
m³	kubieke meter
km³	kubieke kilometer

4.3. Symbols whose pronunciation varies depending on the context

4.3.1. Hyphen

A hyphen '-' is pronounced *min* 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 *koppelteken*

In certain date formats, in between days or years, the hyphen is pronounced as *tot*. In other cases the hyphen is not pronounced.

Expression	Reading	
-3	min drie	
44-3	vierenveertig koppelteken drie	
44-3=41	vierenveertig min drie is gelijk aan éénenveertig	
44 - 3 = 41	vierenveertig min drie is gelijk aan éénenveertig	
15-20 Oktober	vijftien tot twintig oktober	[not SP]
6-10 Nov	zes tot tien november	[not SP]
1998-2004	negentienhonderd achtennegentig tot tweeduizendvier	[not SP]
02-02-2002	twee februari tweeduizendtwee	
zuidoost-azië	zuidoost azië	

4.3.2. Asterisk

Asterisk '*' is pronounced *maal* if enclosed by digits and followed by equals sign. In other cases it is pronounced *asterisk*.

Expression	Reading
2*3	twee asterisk drie
2*3=6	twee maal drie is gelijk aan zes
*bc	asterisk be se

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

Number	**Reading**
2580	twee duizend vijfhonderd tachtig
2 580	twee duizend vijfhonderd tachtig
2.580	twee duizend vijfhonderd tachtig
25800	vijfentwintig duizend achthonderd
25 800	vijfentwintig duizend achthonderd
25.800	vijfentwintig duizend achthonderd

Number	Reading
2580350	twee miljoen vijfhonderd tachtig duizend driehonderd vijftig
2 580 350	twee miljoen vijfhonderd tachtig duizend driehonderd vijftig
2.580.350	twee miljoen vijfhonderd tachtig duizend driehonderd vijftig
1000000000	één miljard
123456789012	honderddrieëntwintig miljard vierhonderd zesenvijftig miljoen zevenhonderd negenentachtig duizend twaalf
1234567890123	één twee drie vier vijf zes zeven acht negen nul één twee drie

5.2. Leading zero

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

Number	Reading
09253	nul negen duizend tweehonderd driëenvijftig
020	nul twintig

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* . The decimals (the part after comma or full stop) are read as separate digits if there are more than 3 digits after the comma. Note: A number containing a point 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	zestien komma tweehonderd vierendertig
3,1415	drie komma één vier één vijf
1251,04	duizend tweehonderd éénenvijftig komma nul vier
1.251,04	duizend tweehonderd éénenvijftig komma nul vier
2.50	twee punt vijftig
2,50	twee komma vijftig
3,141	drie komma honderd eenenveertig

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 *pounds, dollars, yen* or *euros* (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 *en nn pence* and *en nn cents*.
- If the decimal part is *00* it will not be read.

Example	Reading	
$15,00	vijftien dollar	
15,00£	vijftien pond	
15,00 euros	Vijftien euro	[not SP]
€ 200,50	tweehonderd euro en vijftig cent	
1.000.000 ¥	één miljoen yen	

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

$ 1 miljoen	één miljoen dollar	
$ 1 mln.	één miljoen dollar	[not SP]
$ 1 miljard	één miljard dollar	
$ 1 mrd.	één miljard dollar	[not SP]

5.5. Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is followed by *de, ste*.

  Examples: 1ste, 2de.

5.6. Arithmetic operators

Numbers together with arithmetical operators are read as in the examples below:

Expression	Reading
-12	min twaalf
44-3	vierenveertig koppelteken drie
44-3=41	vierenveertig min drie is gelijk aan éénenveertig
+24	plus vierentwintig
2+3	twee plus drie
2+3=5	twee plus drie is gelijk aan vijf
2*3	twee asterisk drie
2*3=6	twee maal drie is gelijk aan zes
2/3	twee derde
6/2=3	zes gedeeld door twee is gelijk aan drie
25%	vijfentwintig procent
3.4%	drie punt vier procent

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
77B84Z3	zevenenzeventig b vierentachtig zed drie
0092B87-B	nul nul tweëennegentig b zevenentachtig b

5.8. Time of day

The colon is used to separate hours, minutes and seconds.

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"*

Example: 12:30'45"

d. *hh:mm u*

Example: 15:25 u

e. *hh.mm u*

Example: 15.25 u

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

In pattern a:

If the *mm*-part is equal to *00*, this part will not be read.

Expression	Reading
9:00	negen uur
13:00	dertien uur
12:00	twaalf uur
0:00	middernacht

In pattern b:

An *en* will be inserted before the *ss*-part, and *seconden* will be added after it. If the *ss*-part is equal to *00*, this part will not be read.

Expression	Reading
10:24:00	tien uur vierentwintig
10:24:20	tien uur vierentwintig minuten en twintig seconden

In pattern c:

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

5.9. Year

[not SP] Full numbers (4-digit strings lacking decimal parts) between *1100 and 1900* , are interpreted as years and read as hundreds (year reading).

Expression	Reading
1100	elfhonderd
1888	achttienhonderd achtentachtig
1839-45	achttienhonderd negenendertig tot vijfenveertig

4-digit strings between 1901 and 1999 are equally interpreted as years and split as tens.

Expression	Reading
1988	negentien achtentachtig
1939-45	negentien negenendertig tot vijfenveertig
September 1999	september negentien negenennegentig

Other 4-number digits (including all the ones with decimal parts), are read as regular thousands.

Expression	Reading
1088	duizend achtentachtig
2088	tweeduizend achtentachtig
1988,0	duizend negenhonderd achtentachtig komma nul
1988,32	duizend negenhonderd achtentachtig komma tweëendertig
1988,0	duizend negenhonderd achtentachtig komma nul

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:	Reading
10-02-2003 or 10-2-2003	tien februari tweeduizend drie
10.02.2003 or 10.2.2003	tien februari tweeduizend drie
10/02/2003 or 10/2/2003	tien februari tweeduizend drie

Type 2:	Reading
10-02-03 or 10-2-03	tien februari tweeduizend drie
10.02.03 or 10.2.03	tien februari tweeduizend drie
10/02/03 or 10/2/03	tien februari tweeduizend drie

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

Expression	Reading
1998-1999	negentien achtennegentig tot negentien negenennegentig
1939-45	negentien negenendertig tot vijfenveertig
2002/3	tweeduizend twee tot drie
14-15 januari	veertien tot vijftien januari
oktober 19-20	oktober negentien tot twintig

[not SP] Other possible formats include:

Expression	Reading	Comment
Maandag, 15 januari	Maandag vijftien januari	with or without the comma
30 april 1999	dertig april negentien negenennegentig	
april 30 1999	april dertig nengentien negenennegentig	
januari 1953	januari negentien drieenvijftig	
3 januari	drie januari	

[not SP] For months and days, the following abbreviations can be used in the preceding formats:

Valid abbreviations for months are: *: jan, feb(r), mrt, apr, jun, jul, aug, sep(t), okt, nov* and *dec*.

Valid abbreviations for days are: *: ma, di, wo, do, vr, za* and *zo*.

5.11. Phone numbers

[not SP] Some digit patterns are recognised as phone numbers, and pronounced accordingly. If a pattern is recognised as a phone number, the strings longer than three digits are read out as groups of two or three full cardinal numbers with pauses inbetween. Leading zero's are pronounced separately digit by digit.

For example, "(02)12345678" is split as "0 2 12 34 56 78".

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, a period, or a hyphen:

Format	Example
xxx-xxxxxx	071-2586336
xxx.xxxxxx	071.2586336
xxx xxxxxx	071 2586336
xxxx-xxxxxx	0219-271447
xxxx.xxxxxx	0219.271447
xxxx xxxxxx	0219 271447

Format	Example
xx-xxxxx-xxx	02-19271-447
xx.xxxxx.xxx	02.19271.447
xx xxxxx xxx	02 19271 447

5.11.2. International phone numbers

International phone numbers follow the pattern below:

International prefix + *Country code* + *space or hyphen* + *Local number*.

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

Format	Example
+xx-x-xxxxx-xxx	+31.6.12353.323
+xx.x.xxxxx.xxx	
+xx x xxxxx xxx	
00-xx-x-xxxxx-xxx	00.31.6.12353.323
00.xx.x.xxxxx.xxx	
00 xx x xxxxx xxx	

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

Chapter 7. Dutch Phonetic Text

The Dutch text-to-speech system primarily uses the Dutch subset of the SAMPA phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*). The symbols are written with a space after 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. The symbol "1" immediately following vowels in the tables below indicates lexical stress present on prominent syllable (for further details c.f. Section "Lexical Stress").

7.1. Consonants

Table 7.1. Symbols for the Dutch consonants

Symbol	Word	Phonetic text
p	pad	p A1 t
t	tak	t A1 k
tj	potje	p O1 tj @
k	kat	k A1 t
b	bad	b A1 t
d	dak	d A1 k
g	zakdoek	z A1 g d u k
f	fiets	f i1 t s
s	sap	s A1 p
S	sjaal	S a1 l
x	lach	l A1 x
v	vat	v A1 t
z	zat	z A1 t
Z	plantage	p l A n t a1 Z @
G	regen	r e1 G @
h	huis	h 9y1 s
w	sneeuwen	s n e1 w @
j	aaien	a1 j @
l	alle	A1 l @
r	haar	h a1 r
m	mat	m A1 t
n	nat	n A1 t
N	lang	l A1 N
nj	anjer	A1 nj @ r

7.2. Vowels

Table 7.2. Symbols for the Dutch vowels

Symbol	Word	Phonetic text
I	bid	b I1 t

Symbol	Word	Phonetic text
E	bed	b E1 t
Y	buts	b Y1 t s
O	bos	b O1 s
@	rede	r e1 d @
i	bied	b i1 t
e	beet	b e1 t
y	buut	b y1 t
2	beuk	b 21 k
a	baat	b a1 t
u	boek	b u1 k
o	boot	b o1 t
Ei	bijt	b Ei1 t
9y	buit	b 9y1 t
Au	bout	b Au1 t
E~	timbre	t E~1 b r @
A~	chanson	S A~ s O~1
O~	bonbon	b O~ b O~1
A	bak	b A1 k
Oe	oeuvre	Oe1 v r @
E:	meubilair	m 2 b i l E:1 r

7.3. Lexical stress

In words with more than one syllable, one of the syllables can be percieved as more prominent than the others. This is referred to as *word stress*, or *lexical stress*. Words consisting of one syllable also bear primary word stress when spoken in isolation, although many may lose stress in certain contexts. For the correct pronunciation of a word, it is important to include the symbol marking word stress.

In phonetic transcriptions, word stress is indicated by the symbol */1/* immediately following the stressed vowel (with no space between the vowel and the stress symbol).

Secondary stress can also occur in Dutch (see examples below). Secondary stress refers to full (i.e. very clear) pronunciation of a vowel, but there are no main intonational changes on such vowels (unlike with primary stress). The secondary stress is indicated by the symbol */2/* immediately following the stressed vowel in the same way as for primary stress.

Many vowels such as schwa (/@ /) mostly remain unstressed.

Examples	Transcription
posttrein	p O1 s t r Ei2 n
beslist	b @ s l I1 s t
ervaringswerelden	E r v a1 r I N s w e2 r @ l d @

7.4. Glottal stop

A glottal stop is a small glottal sound, which is often produced in speech to separate two words (or compound parts of the word) when the second word/part starts with a vowel. In

phonetic transcriptions, it is represented by the phonetic symbol /?/. This sound is often produced in clear speech. Glottal stop can optionally be inserted in transcriptions in order to improve pronunciation, as in the examples below.

Examples of glottal stop	**Transcription (without vs. with glottal stop)**
gezinsauto	/x @ z I1 n s Au2 t o/ vs. /x @ z I1 n s ? Au2 t o/
van achter	/v A n A1 x t @ r/ vs. /v A n ? A1 x t @ r/
beantwoorden	/b @ A1 n t w o r d @/ vs. /b @ ? A1 n t w o r d @/

7.5. Pause

An underscore symbol /_/ in phonetic transcriptions generates a small pause.

Chapter 8. Abbreviations

In current version of the Dutch text-to-speech system, abbreviations are recognized in all contexts. Examples of such abbreviations are given in the table below. Some of these abbreviations are case-sensitive: i.e. particularly those containing upper-case characters in the below table, while others are case-insensitive. Also note that some of the abbreviations require full stops in order to be recognized as an abbreviation.

As mentioned above, there are also abbreviations for the days of the week and the months (see chapter *Dates*), as well as in numbers (see chapter *Number Processing*) and signs (see chapter *Signs*).

Table 8.1. Abbreviations

Abbreviation	Reading
Hz	hertz
MHz	megahertz
dwz	dit wil zeggen
m.b.t.	met betrekking tot
t.e.m.	tot en met
bit/s	bits per seconde
dB	decibel
kV	kilovolt
afk.	afkorting
mevr. or mvr.	mevrouw
dhr.	deheer
juf.	juffrouw
prof.	professor
bv.	bijvoorbeeld
ca.	circa
enz.	enzovoort
i.o.	in opdracht
o.a.	onder andere
o.m.	ondermeer
blz.	bladzijde
tel.	telefoon
t.a.v.	ter attentie van
km	kilometer
km/h	kilometer per uur
µg	microgram
a.u.b. or aub	alstublieft
d.m.v.	door middel van
e.a.	en andere
etc	et cetera
i.f.v.	in funktie van
i.p.v.	in plaats van

Abbreviation	Reading
i.v.m.	in verband met
m.a.w.	met andere woorden
m.n.	met name
mvg	met vriendelijke groeten
nr.	nummer
P.S.	post sciptum

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 *punt*, hyphens '-' as *koppelteken*, 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 as *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.acapela-group.com	w w w punt acapela koppelteken group punt com
http://www.acapela-group.com	h t t p dubbele punt slash slash w w w punt acapela koppelteken group punt com
smith@yahoo.us	smith at jahoe punt u s
jane_smith@yahoo.nl	jane underscore smith at jahoe punt n l