Language Manual

Danish

Mette

Language Manual
Danish
Mette
1 May 2006

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

List of Contents:

Acknowledgement

Acapela Group would like to thank the Society for Danish Language and Literature (www.dsl.dk) who kindly provided texts used during the design of the Danish voice database.

1 General

This document discusses certain aspects of text-to-speech processing for the Danish 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 Mette.

2 Letters in orthographic text

Characters from **A-Å** and **a-å** may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, e.g. **ã, á, ô**. These letters are not pronounced as in their native languages though, they are pronounced as regular “a, a, o” etc. However, when one of these letters stands on its own, they are read with an indication of the diacritic (modification of the letter), for instance (for the letters mentioned above): “a med tilde”, “a med accent aigu”, “o med cirkumfleks”.

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.

3.7 Apostrophe

Apostrophe < ' > may be used but it makes no difference to the pronunciation.

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
/	skråstreg
+	plus
$	dollar
£	pund
€	euro
¥	yen
<	mindre end
>	større end
%	procent
^	cirkumfleks
|	vertikal streg
~	tilde
@	snabel-a
²	see below
³	see below
*	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	kvadratmillimeter
cm²	kvadratcentimeter
m²	kvadratmeter
km²	kvadratkilometer
mm³	kubikmillimeter
cm³	kubikcentimeter
m³	kubikmeter
km³	kubikkilometer

4.3 Symbols whose pronunciation vary depending on the context

4.3.1 Hyphen

A hyphen < **-** > is pronounced “minus” only if the input matches the patterns X-Y=Z or -X. No white space characters are allowed in the patterns. It is pronounced “til” in date formats. If more than one hyphen appear together only one is processed.

Expression	**Reading**
44-3=41	44 minus 3 lig med 41
15.-20. oktober	femtende til tyvende oktober
6.-10. nov.	sjette til tiende november
1998-2004	nittenhundredenioghalvfems til totusindogfire
2000-07-31	to nul nul nul bindestreg nul syv bindestreg enogtredve
50-års-dag	50 års dag

4.3.2 Asterisk

Asterisk < ***** > is pronounced “gange” only if the input matches the pattern X*Y=Z. In other cases it is pronounced “asterisk”.

Expression	**Reading**
2*3=6	to gange tre lig med seks
*bc	asterisk b c

4.3.3 Equals sign

Equals sign < = > is pronounced “lig med” if preceded or followed by a digit. In all other cases it is pronounced “lighedstegn”. Examples:

Expression	**Reading**
2*3=6	to gange tre lig med seks
c d=d c	c d lighedstegn d 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
Year
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 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.
- An exception is made for year pronunciation, which occurs in four-digit strings in the range between 1100 and 2099.

Number	**Reading**
2585	totusind femhundredefemogfirs
2.585	---"---
2 585	---"---
25700	femogtyvetusind syvhundrede
25.700	---"---
25 700	---"---
2090350	to millioner halvfemstusind trehundredehalvtreds
2090350	---"---
2090350	---"---
1000000001	en milliard og en
3456789012342	tre fire fem seks syv otte ni nul en to tre fire to

34 567 890 123	fireogtredvemilliarder femhundredesyvogtresmillioner ottehundredehalvfemstusind ethundredetreogtyve

5.2 Leading zero

Numbers that begin with “0” (zero) followed by a maximum of three digits (the first of which is not “0”) are read “nul” and then the number as it would be without the “0”. Numbers that begin with “00” (zero zero) followed by a maximum of two digits (the first of which is not “0”) are read “nul nul” and then the number as it would be without the “00”. Other digit strings beginning with “0” are read out digit-by-digit.

Number	**Reading**
753	syvhundredetreoghalvtreds
020	nul tyve
0053	nul nul treoghalvtreds
00753	nul nul syv fem tre
07253	null syv to fem tre

5.3 Decimal numbers

Comma is used when writing decimal numbers.

The full number part of the decimal number (the part before comma) is read according to the rules in 5.1. 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 5.1.

Number	**Reading**
16,234	seksten komma tohundredefireogtredve
3,1415	tre komma en fire en fem
1251,04	ettusindtohundredenoghalvtreds komma nul fire
2,50	to komma halvtreds
2.50	to punktum halvtreds
3.141	tretusindethundredeenogfyrre

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 **kr, £, $, ¥** or **€** are read as currency amounts.
- Numbers with zero or two decimals followed by the words “kroner”, “dollar”, “yen” or “euro” are read as currency amounts.
- Comma is the only accepted decimal marker.
- The sequence comma followed by hyphen < **,-** > is read as “kroner”
- No spaces are allowed in the number.
- The decimal part (consisting of two digits) in currency amounts is read as “og nn øre”, “og nn pence”, and “og nn cent”.
- If the decimal part is “00” it will not be read.

Example	**Reading**
kr 20,50	tyve kroner og halvtreds øre
kr 20,00	tyve kroner

20,50 kroner	tyve kroner og halvtreds øre
$15,00	femten dollar
15,00£	femten pund
€ 200,50	tohundrede euro og halvtreds cent
1.000.000 ¥	en million yen
$1.314,57	tusind trehundredefjorten dollar og syvoghalvtreds cent

5.5 Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is followed by a full stop and 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:

  15. januar
  15. jan.
  onsdag 15. jan.
  ons. 15. jan.

- The number consists of a day interval followed by a month name/abbreviation. Example: 15.-16. januar.
- The number is part of the date format dd/mm yyyy and occurs in the dd/mm part, dd/mm must be a possible date and yyyy a year between 1100 and 2099. Example: 3/7 2003. See also section 5.10.

Valid abbreviations for months: jan, feb, febr, mar, apr, jun, jul, aug, sep, sept, okt, nov, and dec.

Valid abbreviations for days: man, tir, tirs, ons, tor, tors, fre, lør and søn.

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 tolv
+24	plus fireogtyve
2*3=6	to gange tre lig med six
6/3=2	seks divideret med tre lig med to
6:3=2	seks divideret med tre lig med to
25%	femogtyve procent
3,4%	tre komma fire 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	syvoghalfjerds B fireogfirs Z tre
0092B87-B	nul nul tooghalvfems B syvogfirs B
208Kr.	tohundredeotte kroner

5.8 Time of day

Either colon or full stop may be used to separate hours, minutes and seconds.

Possible patterns are:

a) **kl hh.mm (or h.mm)**

b) **kl hh.mm.ss (or h.mm.ss)**

c) **kl hh-hh (range of hours)**

Colon may be used instead of full stop and "klokken" instead of "kl" in both patterns. h = hour, m = minute, s = second.

Expression	**Reading**
klokken 16.15.34	klokken seksten femten fireogtredve
kl 16.15.34	klokken seksten femten fireogtredve
16.15.34	seksten femten fireogtredve

5.9 Years

Numbers between 1100 and 2099 are always read as hundreds ("year reading") with the exception of numbers containing decimals.

Expression	**Reading**
1988	nittenhundredeotteogfirs
1939-45	nittenhundredeniogtredve til femogfyrre
1998-2010	nittenhundredeotteoghalfems til totusindogti
2000	totusind
X2000	X totusind
2004	totusindogfire
1088	ettusindogotteogfirs
1900	nittenhundrede
1988	nittenhundredeotteogfirs
1988,0	ettusindnihundredeotteogfirs komma nul
1988.32	ettusindnihundredeotteogfirs punktum toogtredve
sep 2004	september totusindogfire
13. sep 2004	trettende september totusindogfire

5.10 Dates

There are four types of valid formats for dates:

1.	dd/mm/yy	25/12/04
2.	dd/mm/yyyy	25/12/2004
3.	dd-mm yyyy	25-12 2004
4.	dd/mm yyyy	25/12 2004

All the above examples are read as "femogtyvende i tolvte totusindogfire".

"**yyyy**" is a four-digit number between 1100 and 2099, "**yy**" is a two digit number, "**mm**" is a month number between 1 and 12 and "**dd**" a day number between 1 and 31.

In type 1 and 2 full stop, and slash may be used as delimiters, in type 3, only hyphen and in type 4 only dash.

In all four formats, one or two digits may be used in the "mm" and "dd" part. Zeros may be used in front of numbers below 10.

5.11 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 (see also Leading zero, section 5.2) with a pause between the regional code and the local number, and pauses between groups of numbers.

5.11.1 Ordinary phone numbers

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

- A regular Danish phone number always has 8 digits, divided into groups of two or two groups of two and six digits. We also recognize the format: two groups of four digits.
- International numbers are preceded by "+" or "00" plus country code, and for internation numbers only, the phone number may have all 8 digits in one group

Examples
22 33 44 55
22 334455
+47 22 33 44 55
00 47 22334455
1234 5678

5.11.2 Special phone numbers

The emergency numbers: 112
Service numbers: 1800 (4-digit numbers)

There are more numbers like this and they are pronounced according to the general rules for pronouncing full numbers.

6 How to change pronunciation errors

6.1 User lexicon

Words that are not pronounced correctly by the text-to-speech converter can be entered in the user lexicon (see User's guide). When writing translations for entries in the user lexicon to change the way a word is pronounced, one method is to modify the spelling of the word (see section 6.2) and another is to write a phonetic transcription of the word (see chapter 7). Phonetic translations can also be entered directly in the text, using the PRN-tag (see User's guide)

6.2 Spelling incorrectly

Sometimes the quickest way of changing the pronunciation of the word is to change the spelling of the word directly in the text. Changing a single letter, or adding a hyphen, can often make it sound better.

7 Danish Phonetic Text

The Danish text-to-speech system uses the Danish subset of the SAMPA phonetic alphabet (Speech Assessment Methods Phonetic Alphabet) with a few exceptions. The symbols are written with a space between each phoneme. Earlier text-to-speech converters developed by Babel-Infovox use a different phonetic alphabet, called RULSYS. As some users may be accustomed to this alphabet we also provide here the corresponding RULSYS-symbols for every SAMPA-symbol. 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 Stress

Stress is marked with the numerals < 1 > and < 2 > placed immediately after the stressed vowel. Primary stress is marked with < 1 >, secondary stress is marked with < 2 >. Secondary stress almost never appears outside compounds and other poly-morphematic words (words of more than one morpheme).
In Danish, stress may have a lexical as well as a grammatical/pragmatic function. Stress has a lexical function on the word level, i.e. the stress distribution distinguishes minimal pairs such as *billigst-bilist*.

Orthographic text		**Phonetic text**
billigst	(cheapest)	[b i1 l i s d]
bilist	(car driver)	[b i l i1 s d]

Stress has a grammatical/pragmatic function on the phrasal level. In the example below, the stress pattern distinguishes the literal versus the idiomatic use of the phrase *gå i vandet*.

Orthographic text		**Phonetic text**
gå i vandet	(to walk in the water)	[g O:q1 i v a1 nq D=]
gå i vandet	(to go into the water)	[g O i v a1 nq D=]

On the sentence level, stress serves to accentuate semantically "heavy" words from semantically "light" words such as function words etc. In the example below, only *fandt* and *gaden* are stressed and the function words *jeg, den* and *på* are unstressed.

Orthographic text	**Phonetic text**
jeg fandt den på gaden (I found it in the streets)	[j a f a1 nq d d n= p O g {:1 D= n]

7.2 Vowel Length

Long vowels are marked with colon < : > placed immediately after the vowel. In Danish, length has a lexical function, and distinguishes minimal pairs such as *læse-læsse* (see below). In Danish, all the distinctive vowel qualities are found both long and short, and apart from certain phonotactic contexts, vowel length is not predictable in any way. For instance, long vowels are not found preceding consonant clusters in mono-morphematic words (words made of just one morpheme), except when preceding [s d] and [s g] and a few other special exceptions such as *æble*.

Orthographic text		**Phonetic text**
læsse	(to load)	[l E1 s @]
læse	(to read)	[l E:1 s @]

7.3 Stød

Stød is marked with a < q > placed immediately after a long vowel or a short vowel plus [D w j l r m n N] in stressed syllables. *Stød* has a lexical function, and distinguishes many minimal pairs such as *mord-mor* (see below).

Orthographic text		**Phonetic text**
mord	(a murder)	[m o1 rq]
mor	(a mother)	[m o1 r]

7.4 Symbols for the Danish consonants

Symbol	Word	Phonetic text	Comment	RULSYS
b	bat	b a1 d		B
d	dat	d a1 d		D
f	fat	f a1 d		F
g	gat	g a1 d		G
h	hat	h a1 d		H
j	ja	j a1		J
k	kat	k a1 d		K
l	ladt	l a1 d		L
m	mat	m a1 d		M
n	nat	n a1 d		N
N	længe	l E1 N @		N1
p	pat	p a1 d		P
r	bær	b {1 r		R1
s	sat	s a1 d		S
S	sjat	S a1 d		S1
t	tam	t A1 mq		T
v	vat	v a1 d		V
l=	cykel	s y1 g l=	Syllabic l	
n=	katten	k a1 d n=	Syllabic n	
x	Bach	b A1 x	German x	
w	rav	r A1 w		W
T	Beth	b E1 T	English T	
D	bad	b a1 D		D1
Dq	ræd	R a1 Dq	*stødt* D	
D=	kommet	k V1 m D=	Syllabic D	
j=	nage	n {:1 j=	Syllabic j	
jq	Thai	t A1 jq	*stødt* j	
lq	ral	r A1 lq	*stødt* l	
r’	country	k V1 n t r’ i	English r	
R	ret	r a1 d		R
rq	ser	s e1 rq	*stødt* r	
wq	tav	t A1 wq	*stødt* w	
w=	kræve	k R {:1 w=	Syllabic w	
mq	ram	r A1 mq	*stødt* m	
m=	gennem	g E1 nq m=	Syllabic m	
N=	lakken	l A1 g N=	Syllabic N	
Nq	rang	R A1 Nq	*stødt* N	
nq	rand	R A1 nq	*stødt* n	

Table 2 Danish consonants

7.5 Symbols for the Danish Vowels

Symbol	Example	Phonetic text	Comment	RULSYS
A	vams	v A1 mq s		A3
A:	arne	A:1 n @	long A	A3:
e	vind	v e1 nq		E
e:	mele	m e:1 l @	long e	E:
@	ligge	l e1 g @		E0
i	pinje	p i1 n j @		I
i:	mile	m i:1 l @	long i	I:
u	hund	h u1 nq		U
u:	mule	m u:1 l @	long u	U:
y	tyst	t y1 s d		Y
y:	syne	s y:1 n @	long y	Y:
O	ond	O1 nq		Å
O:	måle	m O:1 l @	long O	Å:
{	hver	v {1 rq		A
{:	male	m {:1 l @	long {	A:
2	pynt	p 21 nq d		Ø
2:	søle	s 2:1 l @	long 2	Ø:
a	malle	m a1 l @		A1
&	første	f &1 r s d @		Ø2
{:q	gren	g R {:q1 n	*stødt* {:	
2:q	møn	m 2:q1 n	*stødt* 2:	
9:q	frøs	f R 9:q1 s	*stødt* 9:	
A:q	barn	b A:q1 n	*stødt* A:	
E:q	pæn	p E:q1 n	*stødt* E:	
e:q	fe	f e:q1	*stødt* e:	
i:q	vin	v i:q1 n	*stødt* i:	
O:q	hån	h O:q1 n	*stødt* O:	
o:q	to	t o:q1	*stødt* o:	
Q:q	tårn	t Q:q1 n	*stødt* Q:	
u:q	tun	t u:q1 n	*stødt* u:	
y:q	fyn	f y:q1 n	*stødt* y:	
9	skønt	s g 91 nq d		Ø1
9:	høne	h 9:1 n @	long 9	Ø1:
6	ligger	l e1 g 6		
E	ven	v E1 n		Æ
E:	pæne	p E:1 n @	long E	Æ:
o	ord	o1 rq		O
o:	mole	m o:1 l @	long o	O:
Q	bordeaux	b Q d o1		Å1
Q:	årle	Q:1 l @	long Q	Å1:
V	ånd	V1 nq		Å2

Table 3 Danish vowels

7.6 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.

7.7 Pause

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

8 Abbreviations

In the current version of the Danish text-to-speech system, the abbreviations in table 4 below are recognised in all contexts. These abbreviations are case-insensitive and require no full stop in order to be recognised as an abbreviation.

Some abbreviations representing units of measurement and measures of capacity are only expanded after digits. Abbreviations connected to telephony are only expanded in front of digits.

Abbreviation	**Reading**
20 ml	tyve milliliter
25 cl	femogtyve centiliter
30 dl	tredve deciliter
40 mm	fyrre millimeter (“mm” is normally read as “med mere”)
50 cm	halvtreds centimeter
60 dm	tres decimeter
70 kg	halvfjerds kilogram
tlf 32 12 34 56	telefon toogtredve (pause) tolv (pause) fireogtredve (pause) seksoghalvtreds
mob 24 535043	mobil fireogtyve (pause) treoghalvtreds (pause) halvtreds (pause) treogfyrre

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

Abbreviation	Reading
evt.	eventuelt
tlf.	telefon
bl.a. or bla.	blandt andet (requires full stops)
ca.	cirka
d.v.s.	det vil sige
etc.	etcetera
jr.	junior
osv.	og så videre
mm.	med mere
mfl.	med flere
co.	kompagni
d.s.	det samme
dr	doktor
e.kr.	efter Kristus
f.kr.	før Kristus
f.eks.	for eksempel
f.ex.	for eksempel

Table 4 Abbreviations

9 Web-addresses and email

Web-addresses and email-addresses are read as follows:

- “www” is read as three normal “w”’s spelled letter by letter.
- Full stops are read as “punktum”, hyphens as “bindestreg”, underscore (“_”) as “understregning”, slash (“/”) as “skråstreg”.
- “dk, uk, us” and all the other abbreviations for countries are spelled out letter by letter.
- The “@” is read “ snabel-a ” .
- 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 punktum acapela bindestreg group punktum com
support@acapela -group.com	support snabel-a acapela bindestreg group punktum com