Language Manual

Danish

Poul

Language Manual
Danish
Poul
Edition 41
1 January 2005

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 Labege 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

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 Density voice Poul.

1.1 Notational conventions

The following notational conventions are used in this manual:

- For linguistic entities in general, **boldface** is used.
- Input text is written in a `non proportional font`.
- Output text is written in *italics*.
- Keyboard entities are written within angle brackets < >.
- Phonetic transcriptions are written within slashes (/ /) or hash marks (# #) depending on the phonetic alphabet used.

The following abbreviations are used in this manual:

LM	Letter mode
SM	Sentence mode

See the User's Guide for a description of the two different reading modes. Note that Sentence mode sometimes is referred to as Normal mode.

2 Letters in orthographic text and control characters

Characters from **A-Å** and **a-a** may constitute a word. Swedish **Ä**, **ä**, ¨, and **ö,** may be used instead of **Æ, æ, Ø ,** and **ø**. Certain other characters are also considered as letters, notably those used as letters in other European languages, "**n**, **ü**, **c**" etc, see section 2.1. The apostrophe < ' > may also occur among letters in words; the insertion of this character can affect the pronunciation of a word, see section 3.2.4.

Characters outside of these ranges, i.e. digits and non-alphanumeric characters such as punctuation characters and currency markers etc, are not considered as letters. If any of these characters are found within a word, the word is ended where the non-letter appears and the following letters considered belonging to a new word.

2.1 Characters treated as letters in other languages

Vowels with acute accent , < ´ >, grave accent, < ` >, circumflex accent, < ^ >, or trema (< ¨ >), are mostly read as the corresponding vowels without the diacritics (accent marks etc), except that the vowel with the accent mark sometimes, but not consistently, will receive the primary word stress. In LM, the diacritic used is named.

β is assumed to be the German ligature for **ss** when it occurs next to one or more letters and is then read as the letter **s**. In other cases it is read as *beta*.

Table 1 lists the characters used as letters in other European languages that are treated as letters by the system.

Character	SM	LM
í	*i*	*i med akut accent*
ì	*i*	*i med gravis accent*
î	*i*	*i med circumfleks accent*
ï	*i*	*i med trema*
ü	*y*	*u umlaut*
Ü	*Y*	*(stort) U umlaut*
ç	*s*	*c cedille*
Ç	*S*	*(stort) C cedille*
ñ	*nj*	*n med tilde*
Ñ	*Nj*	*(stort) N med tilde*

Table 1 Characters used as letters in other European languages

2.2 Control characters

In LM (only) the control characters <RETURN> and <TAB> among others are read out according to Table 2 below. In most cases, they are read as *control* + the appropriate letter, e.g. **^T** is read as *control T*. Some other control characters are read according to their function in text applications, see below for examples.

Character	LM
^H <BACKSPACE>	*slet bagud*
^I <TAB>	*tab*
^J <LINE FEED>	*ny linie*
^M <RETURN>	*retur*
^? <DELETE>	*slet*

Table 2 Control characters

3 Non-alphanumeric characters

The processing of non-alphanumeric characters varies, depending on the reading mode, context of the character, and its function within that context. These are the types of non-alphanumeric characters to be distinguished:

- Characters always processed as punctuation, and having a direct effect on the intonation and pausing in SM.
- Characters whose pronunciation varies according to context.
- Other non-alphanumeric , non-punctuation characters, with no effect on the intonation or pausing.

Below is a discussion of the characters grouped by type. For each character, the pronunciation is given in the three basic reading modes.

3.1 Punctuation characters

Table 3 below lists punctuation characters permitted in the normal text input string and their readings in LM. In SM most of them are silent but they affect both rhythm and intonation as described in the sections below.

Character	LM	SM
.	*punktum*	(silence, see 3.1.1)
,	*komma*	(silence, see 3.1.1)
!	*udråbstegn*	(silence)
?	*spørgsmålstegn*	(silence)
:	*kolon*	(silence)
;	*semikolon*	(silence)
(	*venstreparentes*	(silence)
)	*højreparentes*	(silence)
'	*apostrof*	(silence, see 3.1.3)
"	*gåseøjne*	(silence, see 3.1.3)
»	*venstre anførselstegn*	(silence)
«	*højre anførselstegn*	(silence)

Table 3 Punctuation characters

3.1.1 Comma, colon and semicolon

Comma < **,** >, colon < **:** > and semicolon < **;** > cause a brief pause where they occur in a sentence and a rising intonation to precede the pause. Comma is used as decimal marker in numbers, see section 4.4.

3.1.2 Quotation marks

Quotes **< "** > or < « » > appearing around a single word or a group of words cause a brief pause before and after the quoted text in SM. < > are read as *gåseøjne* in LM, and < « » > are read as *højre anførselstegn* and *venstre anførselstegn* respectively in LM.

3.1.3 Apostrophe

Apostrophe < **'** > is read as *apostrof* in LM, and in SM if it is separated from the nearest letter character by a character that is not a letter (digits, space etc.); in other cases it is not pronounced. Quotation marks; in SM, they cause a short pause to be inserted where they appear in the text.

3.1.4 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 numbers, see section 4.4.

3.1.5 Question mark and exclamation mark

A sentence ending with question mark <**?**> or an exclamation mark <**!**> has the same intonation as a sentence ending with a full stop.

3.1.6 Parentheses

Parentheses < **( )** > around a single word or a group of words cause a brief pause before and after the bracketed text.

3.2 Characters whose pronunciation varies

The pronunciation of the characters listed below varies according to their context.

Character	LM	SM
-	*bindestreg*	(see 3.2.1 and 3.2.5)
=	*lighedstegn*	(see 3.2.3 and 3.2.5)
+	*plus*	*plus* (see 3.2.5)
*	*stjerne*	(see 3.2.2 and 3.2.5)
#	*nummertegn*	*nummertegn* (see 3.2.5)
@	*snabel-A*	*snabel-A*
'	*apostrof*	(see 3.2.4, 6.4.2 and 3.2.5)
`	*omvendt apostrof*	*omvendt apostrof* (in SM: silence)

Table 4 Characters with varying pronunciation

All examples below show the reading in SM.

3.2.1 Hyphen

The reading of hyphen <-> follows the following principles:

- If a digit immediately follows, it is pronounced *minus*.
- If preceded by a digit and a non-digit follows, it is ignored.
- When occurring next to another hyphen, it is pronounced *streg*
- If surrounded by spaces and/or other non-letter, non-digit characters, it is pronounced as a short pause.
- When used to mark compound words, it is not pronounced.
- When preceded by a letter and followed by a digit, punctuation mark, hyphen or other arithmetic operator, it is not pronounced.
- When followed by a space or a non-digit character, it is read *bindestreg*.
- Hyphen is discarded at the end of a line, and causes the two parts of the hyphenated word to be joined into a single word.

Expression	Reading
`44-3`	*44 minus 3*
`44 -3`	*44 minus 3*
`44- 3`	*44 3*
`23—bc`	*treogtyve streg streg B C*
`Det mener du ikke - er det sandt?`	SM*: Det mener du ikke* (pause) *er det sandt?*
`data-maskine`	*data bindestreg maskine*
`data-.`	*data*
`data- .`	*data bindestreg*
`En sætning kan være usæd-<CR>vanliglang.`	*En sætning kan være usædvanlig lang*

3.2.2 Equals sign

Equals sign < **=** > is always pronounced *er lig* if preceded and followed by digits. In all other cases, it is pronounced *lighedstegn*.

Expression	**Reading**
`2*3=6`	*to gange tre er lig seks*
`cb=bc`	*C B lighedstegn B C*

3.2.3 Asterisk

Asterisk < ***** > is pronounced *gange* if a digit immediately precedes and follows it; it is pronounced *stjerne* in all other cases.

Expression	**Reading**
`2*3`	*to gange tre*
`*bc`	*stjerne B C*

3.2.4 Apostrophe

< ' > is read as *apostrof* in LM, and also in SM if surrounded by non-letter characters; otherwise, i.e. next to one or more letter characters in SM, it is not read, but can affect the way the word is pronounced. For example, the word **ved** will most often (but not always) be pronounced without a word stress, as is normally appropriate when it functions as the preposition **ved**, as in **ved siden af**. This is normally not appropriate, however, if it is the verb form **ved**, as in **Jeg ved hvad du gjorde**; typing **v'ed** or **ve'd** results in **ved** being pronounced with a stress.

3.2.5 Multiple occurrences of the same character

In SM, if more than three of the same character occur in sequence without a space separating the characters, only the first three occurrences will be pronounced. This is only valid for the following characters: < *** + = - #**>.

Expression	**Reading**
`*****`	*stjerne stjerne stjerne*
`+++++`	*plus plus plus*
`-----`	*bindestreg bindestreg bindestreg*
`=====`	*lighedstegn lighedstegn lighedstegn*
`#####`	*nummertegn nummertegn nummertegn*

3.3 Other non-alphanumeric characters

Each of the characters in Table 5 are read in LM, most of them also in SM.

Character	SM	LM
£	*pund*	*pund*
€	*euro*	*euro*
ª	*højt A*	*højt A*
º	*højt O*	*højt O*
¿	(silence)	*omvendt spørgsmålstegn*
½	*en halv*	*en halv*
¼	*en kvart*	*en kvart*
¡	(silence)	*omvendt udråbstegn*
α	*alfa*	*alfa*
β	*beta* (see 2.1)	*beta*
γ	*gamma*	*gamma*
π	*pi*	*pi*
∞	*uendeligt*	*uendeligt*
≡	*identisk med*	*identisk med*
±	*plus minus*	*plus minus*
≥	*større end eller lig med*	*større end eller lig med*
≤	*mindre end eller lig med*	*mindre end eller lig med*
÷	*minus*	*minus*
≈	*circa lig med*	*circa lig med*
°	*grader*	*grader*
•	*fedt punkt*	*fedt punkt*
·	*højt punkt*	*højt punkt*
²	*i anden*	*i anden*
	blanktegn	*blanktegn*

Table 5 Other non-alphanumeric characters

Special case:
km², mm², cm² and **m²** are read as *kvadratkilometer, kvadratmillimeter* etc. (**km**, **cm,** and **mm** are read as abbreviations, see chapter 8, but **m** on its own is read as the name of the letter; to have it read as e.g. *Meter* it is necessary to enter it in the user lexicon).

3.4 Characters ignored by the system

All characters that are not described in chapter 2 and 3 and that are not phonetic symbols or digits, are ignored by the system. Normally, these characters are omitted but some of them may cause the sentence they appear in to be silent.

4 Number processing

Strings of digits that are sent to the text-to-speech converter are processed in several different ways, depending on the reading mode, format of the digit string, 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:

- 4.1 Full number pronunciation
- 4.2 Leading zero
- 4.3 Year reading
- 4.4 Decimal numbers
- 4.5 Monetary amounts, time of day, and date formats
- 4.6 Arithmetic operators
- 4.7 Mixed digits and letters

The examples in this chapter show the reading in SM. In LM, all digits and punctuation marks are read.

Note that there is no provision for the reading of ordinal numbers in the current Danish system.

4.1 Full number pronunciation

Full number pronunciation is given for the whole number part of the digit string, i.e. the part to the left of the decimal marker. Numbers denoting thousands, millions and billions (numbers larger than 999) may be grouped in groups of three starting at the end using full stop as a separator. If a number string containing full stops does not consist of groups of exactly 3 digits (1, 2, or 3 digits in the case of the group preceding the first full stop) the string will be broken up into separate groups of numbers, each read according to the ordinary number processing rules. The full stops will be read as *punktum.*

The highest number read is 999999999999 (twelve digits). Numbers greater than that are read as separate digits, with pauses between groups.

Number	**Reading**
2425	*totusindfemogtyve*
1000000000	*en milliard*
123456789012	*ethundredetreogtyve milliarder firehundrede seksoghalvtreds millionersyvhundredeniogfirstusind tolv*
2.425	*totusindfirehundredefemogtyve*
22 000	*toogtyve nul nul nul*
1.0880	*en punktum nul otte otte nul*
1.2988	*en punktum totusennihundredeotteogfirs*
1.1988	*en punktum nittenhundredeotteogfirs*
1.198.238000	*en punktum ethundredeotteoghalvfems punktum tohundredeotteogtredivetusen*
1198.238.000	*ellevehundredeotteoghalvfems punktum tohundredeotteogtredive punktum nul nul nul*

4.2 Leading zero

Numbers that begin with 0 (zero) are read as single digits.

Number	**Reading**
09253	*nul ni to fem tre*
0210	*nul to et nul*

4.3 Year reading

A four digit number between 1100 and 1999 is read as hundreds (year reading).

Number	**Reading**
1088	*ettusindotteogfirs*
1900	*nittenhundrede*
1988	*nittenhundredeotteogfirs*
1.988	*ettusindnihundredeotteogfirs*
1988,0	*nittenhundredeotteogfirs komma nul*
1988.0	*nittenhundredeotteogfirs punktum nul*

4.4 Decimal numbers

Decimal numbers should normally be written with a comma. If full stop is used, the decimal part is read according to the number processing rules described above. The decimal part is read as single digits. Note that a number containing full stop is not read as a decimal number.

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

4.5 Monetary amounts, time of day and date formats

The current version of the Danish system does not recognise specific formats for monetary amounts, time of day, and dates.

A full stop < **.** > is read as *punktum* if the special format with groups of exactly three digits is not used, and a colon < **:** > between digits is read as *kolon*. Monetary values can be written as a number followed by **kr** or with **kr** embedded between the kroner amount and the øre amount.

Expression	**Reading**
2988:45	*totusindnihundredeotteogfirs kolon femogfyrre*
2988.45	*totusindnihundredeotteogfirs punktum femogfyrre*
2988kr45	*totusindnihundredeotteogfirs kroner femogfyrre*
200:-	*tohundred*
200.-	*tohundrede*
200kr	*tohundrede kroner*
200kr50	*tohundrede kroner halvtreds*
19 30	*nitten tredive*

4.6 Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below. See also section 3.2.

Expression	**Reading**
-12	*minus tolv*
+24	*plus fireogtyve*
2*3	*to gange tre*
25%	*femogtyve procent*
3,4%	*tre komma fire procent*
,05%	*nul fem procent*

4.7 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. If there is a sequence of letters within a digit string, the sequence will be read according to the normal pronunciation rules.

Expression	**Reading**
77B84Z3	*syvoghalvfjerds B fireogfirs Z tre*
77BOB84Z3	*syvoghalvfjerds bob fireogfirs Z tre*
0092B87-B	*nul nul ni to B syvogfirs B*
208kr	*tohundredeotte kroner*

5 Danish Phonetic Text

In the current version of the text-to-speech system, SAMPA (Speech Assessment Methods Phonetic Alphabet) is used when making lexicons or using phonetic strings within texts. In earlier versions, RULSYS was used. For the voices based on RULSYS, a conversion is made automatically from SAMPA to RULSYS inside the system.

We recommend new users to use only SAMPA since this is the notation that will be used in future development. Users who are already familiar with the RULSYS alphabet still have the possibility to use it when making user lexicons for all RULSYS-based voices (among them the Danish voice Poul). There will be a description of RULSYS in the next chapter.

For the sake of clarity, SAMPA transcriptions are written within slashes (/ /) and RULSYS transcriptions within hash marks (# #). Note that neither the slashes nor the hash marks are part of the actual transcription.

The Danish system uses a phonetic alphabet similar to the Danish subset of SAMPA. The phonetic alphabet is described below.

If the pronunciation is incorrect the user may write phonetic transcriptions in the text. Then, a PRN-tag is needed to switch to phonetic mode, see User's Guide. It is also possible to make user lexicons (see User's Guide), or change the orthography of a word (see chapter 7) in order to achieve the preferred pronunciation.

5.1 Consonants

Table 6 lists the phonetic symbols used for the Danish consonants along with example words (the letters corresponding to the consonant sound are in boldface) and their transcriptions.

Consonant symbol	Example	Transcription
b	**b**us	/b u4 s/
	ku**p**	/k u4 b/
p	**p**us	/p u4 s/
v	**v**and	/v a4 n ?/
w	ha**v**	/h A4 w/
f	gi**ft**	/g i4 f d/
m	du**m**	/d O4 m ?/
d	**d**røm	/d r _| &4 m ?/
	ka**t**	/k a4 d/
D	ba**d**	/b a4 D/
	vå**d**	/v O:4 D/
t	**t**ak	/t A4 g/
s	**s**ult	/s u4 l ? d/
n	**n**ok	/n V4 g/
l	**l**im	/l i:4 ? m/
R	**r**od	/r o:4 ? D/
r	bo**rd**	/b o:4 ? r/
j	**j**ul	/j u:4 ? l/
	ka**j**	/k A4 j ?/
g	**g**al	/g {:4 ? l/
	la**k**	/l A4 g/
k	**k**op	/k V4 b/
S	**sj**æl	/S E:4 ? l/
	chokolade	/S o k o l {:4 D @/
N	ga**ng**e	/g A4 N ?/
	læ**ng**e	/l E4 N @/
h	**h**øst	/h 24 s d/

Table 6 Danish consonant symbols in SAMPA

5.2 Vowels

Table 7 lists the phonetic symbols used for the Danish vowels along with example words and their transcriptions.

Vowel symbol	Example	Transcription
i	v**i**ld	/v i4 l ?/
y	t**y**k	/t y4 g/
u	pl**u**s	/p l u4 s/
e	st**i**k	/s d e4 g/
	f**e**d	/f e:4 ? D/
2	l**y**kke	/l 24 g @/
	h**ø**st	/h 24 s d/
9	s**ø**n	/s 94 n/
	h**ø**ne	/h 9:4 n @/
&	m**ø**rk	/m &4 r g/
_| &	gr**ø**nne	/g R _| &4 n @/
o	f**o**t**o**	/f o4 t o/
	f**o**d	/f o:4 ? D/
E	h**e**st	/h E4 s d/
	bl**æ**k	/b l E4 g/
{	v**a**ne	/v {4 n @/
a	r**e**t	/R a4 d/
	k**a**t	/k a4 d/
A	m**a**j	/m A4 j ?/
	st**e**g	/s d A4 j ?/
A	h**a**m	/h A4 m/
	m**a**rk	/m _A A:4 g/
O	f**u**nd	/f O4 n ?/
	m**å**ne	/m O:4 n @/
Q	d**or**sk	/d Q:4 s g/
	år	/Q:4 ?/
V	sl**o**t	/s l V4 d/
	b**å**nd	/b V4 n ?/
@	mal**e**	/m {:4 l @/
	komm**e**	/k O:4 m@/

Table 7 Danish vowel symbols in SAMPA

Note that /_| &/ and /_A A/ are not proper SAMPA symbols but they may still be used when making transcriptions. There are corresponding symbols in RULSYS, see next chapter.

5.3 Comments on phonetic symbols for vowels

5.3.1 Long and short vowels

Vowels can be short or long. Long vowels are marked as such by a colon < **:** > following the vowel symbol. Long vowels can contain "stød" and are in that case followed by < **?** > after the colon.

Examples

din	`/d i:4 ? n/`	(long vowel with "stød")
dine	`/d i:4 n @/`	(long vowel without "stød")
dit	`/d i4 d/`	(short vowel without "stød")

5.3.2 Unstressed /e/

An unstressed /**e**/ is transcribed as /**@**/ and is used primarily in unstressed endings (but not when the letter **r** follows or precedes).

Example **male** `/m {:4 l @/`

The unstressed ending **-ER** should normally be transcribed as /V/, and the ending **-ERE** as /V V/.

Examples	**(en) maler**	`/m {:4 l V/`	(noun, singular)
	malere	`/m {:4 l V V/`	
	lærer	`/l E:4 V/`	
	lærere	`/l E:4 V V/`	

5.4 Extra symbols for phonetic details

In the current version of the Danish synthesis certain phonetic details can be specified in phonetic text. This can be exploited in case the user wishes to achieve an unusual pronunciation, or if the transcription automatically generated by the system is inaccurate.

5.4.1 Creaky voice ("stød")

A "stød", can be realised phonetically as a kind of creaky voice. It is a modification of a long (primarily or secondarily) stressed vowel sound or a voiced consonant sound following a short (primarily or secondarily) stressed vowel. "Stød" is represented by the phonetic symbol /?/ in SAMPA,

Examples

kal	`/k O:4 ? L/`	
vand	`/v a4 n ?/`	

Compare:

mand	`/m a4 n ?/`	(with "stød")
man	`/m a4 n/`	(without "stød")
jeg læser	`/j A l E:4 ? s V/`	
en læser	`/e n l E:4 s V/`	

5.4.2 Lexical stress

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 Danish, there are three levels of stress available within a word. The strongest stress, primary stress, is represented by a < **4** > placed after the vowel to receive the stress.

Examples **time** /t i:4 m @/
hus /h u:4 ? s/

A slightly weaker level of stress, secondary stress, is represented by a < **1** >. Vowels without stress marks are unstressed and receive no stress at all.

Example **vandmanden** /v a4 n m a1 n ? @ n/

Generally there should only be one primary stress mark per word, although more than one secondary stress mark may occur in a word.

5.4.3 Emphasis and reduction

Some words are in most cases pronounced without word stress by the system. These are the so-called function-words; they are mostly prepositions, pronouns, some auxiliary verbs and certain adverbs and conjunctions, i.e. words with a particular grammatical function in the sentence, e.g. **til**, **jeg**, **vil**, **sa**, **men**. Most other words are given a word stress by the system in normal circumstances. Sometimes, however, it is desirable to assign a different stress pattern than that which the system produces automatically. This can be done by the user entering so-called emphasis markers in the text. The only markers recognised by the Danish system are the symbols **_0** and **_1**. To function as emphasis markers these have to be entered as phonetic text, i.e. within a PRN-tag. **_1** gives stress to an otherwise unstressed word, whereas **_0** removes the stress from an otherwise stressed word.

Compare how the meaning is changed when the stress pattern is varied in the sentence below.

Examples Jeg henter bilen.
/_1/ jeg henter bilen.
/_1/ jeg /_0/ henter bilen.
jeg henter /_0/ bilen.

If these emphasis markers appear inside phonetic text it is not necessary to add extra phonetic prefixes.

5.4.4 Punctuation marks

The punctuation marks < **. ! ? ,** > used in phonetic text have the same effect on intonation as when appearing in orthographic text. In SAMPA the punctuation marks are denoted /**_.**/, /**_!**/, /**_?**/, and /**_com**/ respectively.

5.4.5 Hyphen

Hyphen < **-** > in phonetic text can be used to separate parts of a compound word. In SAMPA it is denoted < **_-** >. If the hyphen separating two parts of a word comes at the end of a line, the word is not spoken until the second part on the next line is also read in (SM only). A word written in phonetic text which contains (one or more) hyphens is spoken as a complete word when the system is in SM. For a description of the use of the hyphen character in normal orthographic text, see section 3.2.1.

6 The RULSYS phonetic alphabet

Note that we recommend new users to use only SAMPA since this is the notation that will be used in future development. Note also that it is only possible to use RULSYS when making user lexicons, not in the input text string.

The following differentiates RULSYS from SAMPA in the Danish system:

- no spaces are used within words in transcriptions
- the lexical accent is placed before the vowel to be stressed, not after as in SAMPA
- it is possible to denote secondary stress
- there are symbols for blocking retroflex sounds

Note that the hash marks (# #) are used to indicate RULSYS transcriptions and to differentiate them from SAMPA transcriptions; the hash marks are not part of the actual transcriptions.

If the pronunciation is incorrect the user may write phonetic transcriptions in the text. Then, a PRN-tag is needed to switch to phonetic mode, see User's Guide. It is also possible to make user lexicons (see User's Guide), or change the orthography of a word (see chapter 7) in order to achieve the preferred pronunciation.

6.1 RULSYS Consonants

Table 8 lists the phonetic symbols in RULSYS used for the Danish consonants along with example words and their transcriptions.

Consonant symbol	Example	Transcription
B	**b**us	#B'US#
	ku**p**	#K'UB#
P	**p**us	#P'US#
V	**v**and	#V'A1NQ#
W	ha**v**	#H'A3W#
F	gi**ft**	#G'IFD#
M	du**m**	#D']MQ#
D	**d**røm	#DR'\3MQ#
	ka**t**	#K'A1D#
D1	ba**d**	#B'A1D1#
	vå**d**	#V']:QD1#
T	**t**ak	#T'A2G#
S	**s**ult	#S'ULQD#
N	**n**ok	#N']2G#
L	**l**im	#L'I:QM#
R	**r**od	#R'O:QD1#
R1	bo**rd**	#B'O:QR1#
J	**j**ul	#J'U:QL#
	ka**j**	#K'A2JQ#
G	**g**al	#G'A:QL#
	la**k**	#L'A3G#
K	**k**op	#K']2B#
S1	**sj**æl	#S1'[:QL#
	chokolade	#S1OKOL'A:D1E0#
N1	ga**ng**e	#G'A3N1Q#
	læ**ng**e	#L'[N1E0#
H	**h**øst	#H'\SD#

Table 8 RULSYS consonants

6.2 RULSYS Vowels

Table 9 lists the phonetic symbols in RULSYS used for the Danish vowels along with example words and their transcriptions.

Vowel symbol	Example	Transcription		
I	v**i**ld	#V'ILQ#		
Y	t**y**k	#T'YG#		
U	pl**u**s	#PL'US#		
E	st**i**k	#SD'EG#		
	f**e**d	#F'E:QD1#		
Ø or \	l**y**kke	#L'\GE0#	or	#L'ØGE0#
	h**ø**st	#H'\SD#	or	#H'ØSD#
Ø1 or \1	s**ø**n	#S'\1N#	or	#S'Ø1N#
	h**ø**ne	#H'\1:NE0#	or	#H'Ø1:NE0#
Ø2 or \2	m**ø**rk	#M'\2R1G#	or	#M'Ø2R1G#
Ø3 or \3	gr**ø**nne	#GR'\3NE0#	or	#GR'Ø3NE0#
O	f**o**to	#F'OTO#		
	f**o**d	#F'O:QD1#		
Æ or [	h**e**st	#H'[SD#	or	#H'ÆSD#
	bl**æ**k	#BL'[G#	or	#BL'ÆG#
A	v**a**ne	#V'A:NE0#		
A1	r**e**t	#R'A1D#		
	k**a**t	#K'A1D#		
A2	m**a**j	#M'A2JQ#		
	st**e**g	#SD'A2JQ#		
A3	h**a**m	#H'A3M#		
	m**a**rk	#M'A3:G#		
Å or]	f**u**nd	#F']NQ#	or	#F'ÅNQ#
	m**å**ne	#M']:NE0#	or	#M'Å:NE0#
Å1 or]1	d**o**rsk	#D']1:SG#	or	#D'Å1:SG#
	år	#']1:Q#	or	#'Å1:Q#
Å2 or]2	sl**o**t	#SL']2D#	or	#SL'Å2D#
	b**å**nd	#B']2NQ#		
E0	mal**e**	#M'A:LE0#		
	komm**e**	#K']2ME0#		

Table 9 RULSYS vowels

Note that the following phonetic symbols are equivalent:

- #Ø# and #\#
- #Æ# and #[#
- #Å# and #]#.

6.3 Comments on phonetic symbols for vowels

6.3.1 Long and short vowels

Vowels can be short or long. Long vowels are marked as such by a colon < **:** > following the vowel symbol. Long vowels can contain "stød" and are in that case followed by < **Q** > after the colon.

Examples

din	`#D'I:QN#`	(long vowel with "stød")
dine	`#D'I:NE0#`	(long vowel without "stød")
dit	`#D'ID#`	(short vowel without "stød")

6.3.2 Unstressed E

An unstressed E is transcribed as #**E0**# (E zero), and is used primarily in unstressed endings (but not when the letter **r** follows or precedes).

Example **male** #M'A:LE0#

The unstressed ending **-ER** should normally be transcribed as #]2#, and the ending **-ERE** as #]2]2#.

Examples	**(en) maler**	`#M'A:L]2#` (noun, singular)
	malere	`#M'A:L]2]2#`
	lærer	`#L'[:]2#`
	lærere	`#L'[:]2]2#`

6.4 Extra symbols for phonetic details

In the current version of the Danish synthesis certain phonetic details can be specified in phonetic text. This can be exploited in case the user wishes to achieve an unusual pronunciation, or if the transcription automatically generated by the system is inaccurate.

6.4.1 Creaky voice ("stød")

"Stød" is described in section 5.4.1., It is represented by the phonetic symbol /Q/ in RULSYS.

Examples	**kal**	`#K']:QL#`
	vand	`#V'A1NQ#`
Compare:	**mand**	`#M'A1NQ#` (with "stød")
	man	`#M'A1N#` (without "stød")
	jeg læser	`#JA3 L'[:QS]2#`
	en læser	`#EN L'[:S]2#`

6.4.2 Lexical stress

For a description of lexical stress, see section 5.5. In RULSYS, primary stress is denoted by an apostrophe < **'** > and secondary stress by a grave accent mark < **`** >.

Examples	**time**	`#T'I:ME0#`
	hus	`#H'U:QS#`
	vandmanden	`#V'A1NM`A1NQE0N#`

6.4.3 Emphasis and reduction

In RULSYS, phrase level stress is denoted in the same way as in SAMPA, see section 5.4.3.

6.4.4 Punctuation marks

The punctuation marks < **, . ! ?** > used in phonetic text have the same effect on intonation as when appearing in orthographic text.

The characters < ' > and < ` > have a completely different function when writing in phonetic text than in orthographic text. They are reserved characters used to mark primary and secondary stress in a word, see section 6.4.2. They cannot be used to quote text or single words in phonetic text.

6.4.5 Hyphen

Hyphen < **-** > in phonetic text can be used to separate parts of a compound word. If the hyphen separating two parts of a word comes at the end of a line, the word is not spoken until the second part on the next line is also read in (SM only). A word written in phonetic text which contains (one or more) hyphens is spoken as a complete word when the system is in SM. For a description of the use of the hyphen character in normal orthographic text, see section 3.2.1.

7 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). There are two ways to do this: either, the user enters a phonetic transcription of the word (see chapter 6), or, the user rewrites the word orthographically. Phonetic transcriptions can also be entered directly in the text, using a PRN-tag (see User's guide).

7.1 Change the orthography

7.1.1 Spelling incorrectly

It is possible to intentionally misspell a word by trying to spell a word in a more phonetic manner, i.e., choosing non-ambiguous letter combinations to represent difficult sounds.

Examples **Kierkegaard** can be misspelled **Kirkegard**

7.1.2 Use of hyphen

A hyphen character can be used within a word to separate two letters that might otherwise be incorrectly pronounced together.

Example **ispind** can be written **is-pind**

7.1.3 Expanding acronyms

Not very many acronyms are handled by the current Norwegian system (see chapter 8). Therefore, it may be very useful to expand them in the user dictionary. Since acronyms should be expanded to more than one word it may be difficult to enter a proper transcription. It is much easier to enter the words in question orthographically.

Example **DGI** Danske Gymnastik- og Idrætsforeninger

7.2 Using phonetic text

When you are unable to correct a pronunciation error by misspelling the word, phonetic text should be used to produce the desired pronunciation. When phonetic text is used, the system bypasses the normal spelling pronunciation rules, and pronounces each phonetic symbol "literally", according to the examples listed in Tables 6 and 7.

7.2.1 Choosing the right phonetic symbols

A helpful way to transcribe in phonetic text is to work with a dictionary. Normally, dictionaries give the pronunciation for each word. They also provide a pronunciation key to show how to pronounce the special symbols used in the pronunciation guide. Similarly, Tables 6 and 7 give the pronunciation key for the special phonetic symbols used in Danish for the text-to-speech system.

Using a dictionary, look up the word you want to transcribe. Next to the word you should find the pronunciation. Working with the dictionary's pronunciation key and Tables 6 and 7, convert the dictionary pronunciation symbols to the appropriate Danish symbols for the text-to-speech converter. Symbols that are used in the dictionary to mark syllable or word boundaries should be ignored. Be sure to include the stress assignment information since lexical stress is an important part of a word's pronunciation.

8 Abbreviations

The following abbreviations are case-insensitive, and do not require a full stop in order to be processed as an abbreviation. In SM, if a full stop accompanies the abbreviation, the sentence is terminated at the abbreviation and spoken.

8.1 Abbreviations that function in SM

In the current version of the Danish text-to-speech system, the abbreviations in Table 10 are recognised in all contexts in SM.

Abbreviation	LM	SM
ca	*C A*	*cirka*
dr	*D R*	*doktor*
etc	*E T C*	*etcetera*
evt	*E V T*	*eventuelt*
fx	*F X*	*for eksempel*
hr	*H R*	*herre*
kg	*K G*	*kilo*
kl	*K L*	*klokken*
kr	*K R*	*kroner*
mv	*M V*	*med videre*
nr	*N R*	*nummer*
osv	*O S V*	*og så videre*
pct	*P C T*	*procent*
pga	*P G A*	*på grund af*
pr	*P R*	*per*
tlf	*T L F*	*telefon*
cm	*C M*	*centimeter*
km	*K M*	*kilometer*
mm	*M M*	*millimeter*

Table 10 Abbreviations recognised in SM

The user lexicon may be used to redefine any of these abbreviations, or to create new ones.

8.2 Abbreviations recognised only in SM

The abbreviations in Table 11 are read as abbreviations in SM only.

Abbreviation	LM	SM
bl a	*B L A*	*blandt annet*
f eks	*F eks*	*for eksempel*
p g a	*P G A*	*på grund af*
cm²	*C M i anden*	*kvadratcentimeter*
km²	*K M i anden*	*kvadratkilometer*
mm²	*M M i anden*	*kvadratmillimeter*
m²	*M i anden*	*kvadratmeter*

Table 11 Abbreviations recognised in SM only