Language Manual

Finnish

Matti

Language Manual
Finnish
Matti
Edition 21
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 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

1 General

This document discusses certain aspects of text-to-speech processing for the Finnish 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 Matti.

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

Characters from **A-Z** and **a-z** and the special Finnish characters **Ä**, **ä**, **Ö** and **ö** may constitute a word. The special letters of other Scandinavian languages and vowels with accent marks common to other European languages are also considered as letters, see section 2.1 and 2.2 respectively.
The apostrophe, < **'** >, the acute accent mark,< **`** >, and the at-sign, < **@**>, may also occur among letters in words, see section 3.3.4 and 5.5.1.

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 such a non-letter is included within a word, the word is ended where the non-letter appears and the following letters considered belonging to a new word.

2.1 Other Scandinavian letters

The special letters of other Scandinavian languages are read as indicated in Table 1.

Character	SM	LM
æ	*ä*	*ä*
å	*o*	*ruotsalainen o*
Å	*O*	*(iso) ruotsalainen O*
Æ	*Ä*	*(iso) Ä*
ø	*senttiä*	*sentti*
Ø	*jeniä*	*jeni*

Table 1 Other Scandinavian letters in the Finnish system

2.2 Other European letters

Vowels with the accent marks < ´ ` ^ > or a trema (umlaut/diaeresis: < ¨ >) are mostly read as the corresponding vowels without the accent marks. If desired, words containing accent marks and other diacritic symbols can be entered in the user lexicon (see User's Guide). In LM the diacritic used is named.

C and c (C cedilla) are read as *s* in SM when next to other letters, otherwise as *C sedilji*.
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 SM). In other cases it is read as *beeta*. Table 2 lists the special characters of other European, non-Scandinavian languages.

Character	SM	LM
Ç	*C sedilji/S*	*(iso) C sedilji*
ü	*y*	*saksalainen y*
â	*a*	*a sirkumfleksi*
à	*a*	*a gravis*
ç	*c sedilji/s*	*c sedilji*
ê	*e*	*e sirkumfleksi*
ë	*e*	*e treema*
è	*e*	*e gravis*
ï	*i*	*i treema*
î	*i*	*i sirkumfleksi*
ì	*i*	*i gravis*
É	*e*	*(iso) e akuutti*
ô	*o*	*o sirkumfleksi*
ò	*o*	*o gravis*
û	*u*	*u sirkumfleksi*
ù	*u*	*u gravis*
ÿ	*y*	*y treema*
Ü	*y*	*(iso) saksalainen y*
á	*a*	*a akuutti*
í	*i*	*i akuutti*
ó	*o*	*o akuutti*
ú	*u*	*u akuutti*

Table 2 Other European letters with diacritics

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. There are three 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.
- Other non-alphanumeric, non-punctuation characters that are always pronounced, with no effect on the intonation or pausing.
- Characters whose pronunciation varies according to context.

Below is a discussion of the characters grouped by type. For each character, the pronunciation is given in the three basic reading modes. At the end of the section, there is a section on the reading of control characters and characters that are ignored by the system.

3.1 Punctuation characters

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

Character	LM	SM
.	*piste*	(silence, see 3.1.3)
,	*pilkku*	(silence, see 3.1.1)
!	*huutomerkki*	(silence)
?	*kysymysmerkki*	(silence)
:	*kaksoispiste*	(silence)
;	*puolipiste*	(silence)
(	*vasensulku*	(silence)
)	*oikeasulku*	(silence)
'	*heittomerkki*	(silence, see 3.1.2)
"	*lainausmerkki*	(silence, see 3.1.2)

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 < **,** > has a special function in digit strings, see sections 4.1 and 4.3.

3.1.2 Apostrophe

Apostrophe < ' > is read as *heittomerkki* in LM, and also in SM if it is separated from the nearest letter character by a character that is not a letter (digits, space, carriage return etc.). In other cases it is not pronounced. An exception is when an apostrophe is placed in front of a vowel that is not the first vowel of the word, this will normally cause an extra word stress to be placed on that vowel. See also section 3.3.4 and 5.5.1.

3.1.3 Quotation marks

Quotation marks (quotes) < " > are read as *lainausmerkki* in LM. In SM, quotes 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. Full stop has a special function in digit strings, see section 4.1.

3.1.5 Question mark and exclamation mark

In the current Finnish version of the text-to-speech system, a sentence ending with a question mark < **?** > or an exclamation mark < **!** > is pronounced in the same way as one 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 Other non-alphanumeric characters

The characters listed below are (with a few exceptions) pronounced at all times in all reading modes

Character	SM	LM
%	*prosenttia*	*prosenttia*
&	*ja*	*ja*
/	*kauttaviiva*	*kauttaviiva*
<	*vasen kulmasulku*	*vasen kulmasulku*
>	*oikea kulmasulku*	*oikea kulmasulku*
$	*dollaria*	*dollaria*
£	*punta*	*punta*
Pt	*peseta*	*peseta*
¿	(silence)	*nurinkäännetty kysymysmerkki*
½	*puoli*	*puoli*
¼	*neljännes*	*neljännes*
¡	(silence)	*nurinkäännetty huutomerkki*
α	*alfa*	*alfa*
β	*beeta* (see 2.2)	*beeta*
Γ	*gamma*	*gamma*
π	*pii*	*pii*
Σ	*iso sigma*	*iso sigma*
∞	*ääretön*	*ääretön*
≡	*identtisyysmerkki*	*identtisyysmerkki*
±	*plus miinus*	*plus miinus*
≥	*suurempi tai yhtä suuri kuin*	*suurempi tai yhtä suuri kuin*
≤	*pienempi tai yhtä suuri kuin*	*pienempi tai yhtä suuri kuin*
≈	*likimäärin*	*likimäärin*
°	*astetta*	*astetta*
•	*iso pallo*	*iso pallo*
·	*pieni pallo*	*pieni pallo*
√	*neliöjuuri*	*neliöjuuri*
²	*toiseen*	*toiseen*
	tyhjälyönti	*tyhjälyönti*

Table 4 Other non-alphanumeric characters

Special case

dm², **cm²**, **km²**, **mm²**, and **m²** are read as *neliödesimetriä, neliösenttimetriä* etc. Regarding the abbreviations **dm**, **cm**, **km**, **mm** on their own, and the letter **m** on its own, see sections 8.2 and 8.3. For the possibility of having a different pronunciation of these, see section 8.3.

3.3 Characters whose pronunciation varies

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

Character	SM	LM
-	(see 3.3.1)	*väliviiva*
*	(see 3.3.2)	*tähti*
#	(see 3.3.3)	*risti*
@	(see 3.3.4 and 6.5)	*ät-merkki*
'	(see 3.3.4 and 6.5)	*heittomerkki*
`	(see 3.3.4 and 6.5)	*käänteinen heittomerkki*

Table 5 Characters with varying pronunciation

See section 3.3.5 regarding the reading of multiple occurrences of the characters < **- + = * #** >.

All examples below show the reading in SM.

3.3.1 Hyphen

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

- If a digit immediately follows, it is pronounced *miinus*.
- If preceded by a digit and followed by a non-digit character it is treated as a hyphen, i.e. the number and the following word or character are read as a compound word. In other cases it is pronounced *viiva* if it follows directly after a digit.
- The hyphen is pronounced *viiva* if it occurs next to another < **-** >.
- The hyphen is pronounced as a short pause if it is surrounded by spaces and/or other non-letter characters that are not either digits or arithmetic operators.
- The hyphen is used to mark compound words and is in this case not pronounced. If it is followed by a space or a non-digit character, then it is read as *viiva*. When it is followed by a digit, it is read as *miinus*.
- 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. In this case the hyphen is not read.

Expression	**Reading**
44-3	*44 miinus 3*
44 -3	*44 miinus 3*
44- 3	*44 viiva 3*
23—bc	*kaksikymmentäkolme viiva viiva B C*
Ja pisteenä i:n päällä - kalliit hinnat.	*Ja pisteenä iin päällä* (pause) *kalliit hinnat*
kuorma-auto	*kuormaauto*
Sanan voi jakaa ta-<CR> vuihin.	*Sanan voi jakaa tavuihin*

3.3.2 Asterisk

Asterisk < ***** > is pronounced *kertaa* in SM if surrounded by digits; it is pronounced *tähti* in all other cases.

Expression	Reading
`2*3`	*kaksi kertaa kolme*
`*bc`	*tähti B C*

3.3.3 Hash mark

The hash mark < **#** > is normally pronounced *risti* in all reading modes.

3.3.4 Apostrophe, acute accent and @

At-sign, < **@** >, is read as *ät-merkki*, apostrophe, < **'** >, is read as *heittomerkki*, and acute accent mark, < **`** >, is read as *käänteinen heittomerkki* in LM, and also in SM if surrounded by non-letter characters. Otherwise (i.e. next to one or more letter characters in SM) they are not read, but may affect the way the word is spoken.

Inserting one of the characters < **' ` @** > in a word can sometimes be used to good effect to force the system to a different pronunciation. For example, the word **mitä** will most often (but not always) be pronounced without a word stress, as is normally appropriate when it functions as a relative pronoun, as in **se, mitä sanottiin**. This is normally not appropriate, however, if it is the question word **mitä**, as in **Mitä oikein sanoit**. Typing

`m'itä` or `m`itä` or `m@itä`

results in **mitä** being pronounced with a stress. It is advisable to use this method with interrogative sentences that start with a question word; e.g. **Mitä oikein tarkoitat**

Expression	Reading
`tämä @`	*tämä ätmerkki*
`t'ämä '`	*tämä heittomerk* (with more stress on **tämä** than in the previous example)

3.3.5 Multiple occurrences of the same character

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

Expression	Reading
`*****`	*tähti tähti tähti*
`+++++`	*plus plus plus*
`-----`	*viiva viiva viiva*
`=====`	*on yhtä kuin on yhtä kuin on yhtä kuin*
`#####`	*risti risti risti*

3.4 Control characters

In LM (only), most control characters are read out. In most cases they are read as *kontrolli* + the appropriate letter, e.g. **^T** is read as *kontrolli T*. However, some of the control characters are read in accordance with the function that they are most commonly given in text applications, as follows:

Character	LM
^H <BACKSPACE>	*backspace*
^I <TAB >	*sarkain*
^J <NEW LINE>	*uusi rivi*
^K <VERTICAL TAB>	*vertikaali sarkain*
^M <RETURN>	*rivinvaihto*
^? <DELETE>	*delete*

Table 6 Control characters with special names read in letter mode

3.5 Characters ignored by the system

All characters that are not described in section 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 Year reading
- 4.3 Leading zero
- 4.4 Decimal numbers
- 4.5 Monetary amounts
- 4.6 Arithmetic operators
- 4.7 Mixed digits and letters

Note that there is no provision for the reading of ordinal numbers, the time of day, or dates.

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

4.1 Full number pronunciation

In SM, full number pronunciation is given for the whole number portion of the digit string. Optional full stops may be used in the whole number portion to separate groups of exactly three digits (counting from the end) into hundreds, thousands, millions, and billions. If a digit string containing full stops does not conform to the three-digit-group format each group will be read as the appropriate number and each full stop will be read as *piste*.

Comma < **,** > is used to indicate a decimal amount. When there are more than two decimals, the decimals are read out digit by digit. When LM is enabled, digits strings are read as single digits, and all punctuation symbols are read out. Thousands, millions, and billions cannot be grouped by using space characters. The longest number string pronounced as a full number in SM is **999999999999**. Numbers greater than that (i.e. more than 12 digits) will be read out digit by digit.

All the readings below are in SM.

Number	**Reading**
2425	*kaksituhatta neljäsataakaksikymmentäviisi*
1000000000	*yksi miljardi*
123456789012	*satakaksikymmentäkolmemiljardia neljäsataaviisikymmentäkuusimiljoonaa seitsemänsataakahdeksankymmentäyhdeksäntuhatta kaksitoista*
123.456.789.012	*satakaksikymmentäkolmemiljardia neljäsataaviisikymmentäkuusimiljoonaa seitsemänsataakahdeksankymmentäyhdeksäntuhatta kaksitoista*
1234.56.789.012	*tuhat kaksisataakolmekymmentäneljä piste viisikymmentäkuusi piste seitsemänsataakahdeksankymmentäyhdeksän piste nolla yksi kaksi*
2,25	*kaks pilkku kaksikymmentäviisi*
2,425	*kaksi pilkku neljä kaksi viisi*
22 000	*kaksikymmentäkaksi nolla nolla nolla*

4.2 Leading zero

Numbers that begin with 0 (zero) are read digit by digit.

Number	**Reading**
09253	*nolla yhdeksän kaksi viisi kolme*
0021	*nolla nolla kaksi yksi*

4.3 Decimal numbers

Decimal amounts should normally be written with a comma. The decimal part is read as a whole number when there are two decimals, but digit by digit if there are more than two or if the first of two decimals is **0**.

Number	**Reading**
16,234	*kuusitoista pilkku kaksi kolme neljä*
1251,04	*tuhat kaksisataaviisikymmetäyksi pilkku nolla neljä*
2,50	*kaksi pilkku viisikymmentä*

4.4 Monetary amounts

The current version of the Finnish language only recognises one specific format for monetary amounts. No specific format exists for reading the time of day or dates. If the user wants certain digit strings to be read in a particular way, to represent monetary amounts for instance, it will be necessary to format the input text accordingly. Monetary values can be written as a number followed by **mk** or or **fim**, or with a colon followed by a hyphen < **:-** >.

Expression	**Reading**
200:-	*kaksisataa markkaa*
200.-	*kaksisataa*
200mk	*kaksisataa markkaa*

4.5 Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below. See also section 3.3.

Expression	**Reading**
-12	*miinus kaksitoista*
+24	*plus kaksikymmentäneljä*
2*3	*kaksi kertaa kolme*
25%	*kaksikymmentäviisi prosenttia*
3,4%	*kolme pilkku neljä prosenttia*
,05%	*pilkku nolla viisi prosenttia*

4.6 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`	*seitsemänkymmentäseitsemän B kahdeksankymmentäneljä Z kolme*
`092B8-B`	*nolla yhdeksän kaksi B kahdeksan viiva B*
`208mk`	*kaksisataakahdeksan markkaa*

4.7 Inflected numerals

The Finnish system includes a facility for inflecting numbers in accordance with an ending which can be added to the number. The ending must be separated from the number with a colon. The system inflects the whole sequence of digits according to the case indicated by the ending. No provision for vowel harmony is necessary; this allows some room for incorrect spellings, i.e. the system inflects the numerals correctly as long as the case is correct.

Numbers are only inflected in SM. The following cases are supported:

Case	Ending
Inessive	**:ssa** or **:ssä**
Allative	**:lle**
Adessive	**:lla** or **:llä**
Ablative	**:lta** or **:ltä**
Translative	**:ksi**
Essive	**:na** or **:nä**
Elative	**:sta** or **:stä**
Partitive	**:ta**, **:tä**, **:aa** or **:ää**
Genitive	**:n**
Illative	**:aan**, **:een** or **:ään**

Table 7 Case endings supported in numerals

Expression	Reading in SM
`1234:lle`	*tuhannelle kahdellesadalle kolmellekymmenelle neljälle*
`1234:ää`	*tuhatta kahtasataa kolmeakymmentä neljää*
`1234:sta`	*tuhannesta kahdestasadasta kolmestakymmenestä neljästä*
`1234:n`	*tuhannen kahdensadan kolmenkymmenen neljän*

5 Finnish 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 Finnish voice Matti). There will be a description of RULSYS in the next section.

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 Finnish system uses a phonetic alphabet based on standard SAMPA.

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 section 7) in order to achieve the preferred pronunciation.

5.1 Consonants

Table 8 lists the phonetic symbols used for the Finnish consonants along with example words (the letters corresponding to the consonant sound are in boldface) and their transcriptions.

Consonant symbol	Example	Transcription
b	**b**ussi	/b u4 ss i/
p	**p**ussi	/p u4 ss i/
d	**d**ata	/d A4 t A/
t	**t**akka	/t A4 kk A/
g	**g**aala	/g a4 l A/
k	**k**oppi	/k O4 pp i/
m	**m**atto	/m A4 tt O/
n	**n**okka	/n O4 kk A/
N	tu**ng**e	/t u4 N e/
f	**f**iini	/f i:4 n i/
s	**s**ula	/s u4 l A/
S	**sh**ekki	/S e4 kk i/
h	**h**attu	/h A4 tt u/
v	**v**ain	/v A4 i n/
j	**j**uttu	/j u4 tt u/
l	**l**oma	/l O4 m A/
r	**r**akas	/r A4 k A s/

Table 8 Finnish consonant symbols in SAMPA

5.2 Comments on phonetic symbols for consonants

5.2.1 Long and short consonants

In addition to the consonants in Table 8, there are also long variants for all of them (see section 5.4.1 for the similar distinction in vowels). The long consonant sounds are represent by two identical letters in orthographic text, and, similarly, two identical symbols in phonetic text.

Example: **takka** is transcribed `/t A4 kk A/`

5.3 Vowels

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

Vowel symbol	Example	Transcription
A	k**a**la	/k A4 l A/
A:	t**aa**kka	/t a4 kk A/
e	k**e**ttu	/k e4 tt u/
e:	**ee**va	/e:4 v A/
i	k**i**ta	/k i4 t A/
i:	v**ii**ri	/v i:4 r i/
O	k**o**ko	/k O4 k O/
O:	t**oo**ni	/t O:4 n i/
u	m**u**ki	/m u4 k i/
u:	k**uu**la	/k u:4 l A/
y	**y**li	/y4 l i/
y:	t**yy**li	/t y:4 l i/
{	t**ä**mä	/t {4 m {/
{:	t**ää**llä	/t {: ll {:/
2	k**ö**hä	/k 24 h {/
2:	t**öö**lö	/t 2:4 l 2/

Table 9 Finnish vowel symbols in SAMPA

5.4 Comments on phonetic symbols for vowels

5.4.1 Long and short vowels

In Finnish there are two distinct vowel lengths for all the vowels. The long variants of the vowels are roughly twice the length of the short ones. The long vowel sounds are in SAMPA represented by the same vowel symbol as their short counterparts immediately followed by a colon, i.e **/A:/**, **/e:/**, **/i:/**, **/ O:/** etc.

5.5 Extra symbols for phonetic details

In the current version of the Finnish 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.

5.5.1 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.

Each purely Finnish word should have a primary stress mark after the first vowel. Many longer words have a secondary stress on one or more syllables later in the word. In most cases the system does not assign secondary stress, but the user can insert a secondary stress mark.

Primary stress is in SAMPA denoted by a < **4** > placed immediately after the stressed vowel. Secondary stress, common in many longer words, is denoted by a < **1** >.

Generally there should only be one stress mark per word. If no stress marks appear in a sentence at all, the system will produce a monotone reading of the sentence.

Remember that only vowels are stressed, i.e., a stress mark must be preceded by a vowel written in phonetic characters.

It is also possible to mark the stress in orthographic text. If the primary stress is on another vowel than the first (which is common in foreign words), it can be marked by an apostrophe. Secondary stress in orthographic text is denoted by an at-sign < **@** > or the acute accent mark < `>. See also section 3.3.4.

Example The word **sanomalehti** is transcribed /s A4 n O m A l e4 h t i/.
The same pronunciation is achieved if the word is written
sanomal@ehti or **sanomal`ehti**.

5.5.2 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.

5.5.3 Hyphen

In phonetic text, hyphen 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 read until the second part on the next line is typed.

For a description of the use of the hyphen character in normal orthographic text, see section 3.3.1.

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 Finnish 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

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 section 7) in order to achieve the preferred pronunciation.

6.1 RULSYS Consonants

Table 10 lists the phonetic symbols in RULSYS used for the Finnish consonants along with example words and their transcriptions.

Consonant symbol	Example	Transcription
B	**b**ussi	#B'USSI#
P	**p**ussi	#P'USSI#
D	**d**ata	#D'ATA#
T	**t**akka	#T'AKKA#
G	**g**aala	#G'AALA#
K	**k**oppi	#K'OPPI#
M	**m**atto	#M'ATTO#
N	**n**okka	#N'OKKA#
NG	tu**ng**e	#T'UNGE#
F	**f**iini	#F'IINI#
S	**s**ula	#S'ULA#
2S	**sh**ekki	#2S'EKKI#
H	**h**attu	#H'ATTU#
V	**v**ain	#V'AIN#
J	**j**uttu	#J'UTTU#
L	**l**oma	#L'OMA#
R	**r**akas	#R'AKAS#

Table 10 RULSYS consonants

6.2 Comments on phonetic symbols for consonants

6.2.1 Long and short consonants

In addition to the consonants in Table 10, there are also long variants for all of them (see section 6.4.1 for the similar distinction in vowels). The long consonant sounds are represented by two similar letters in orthographic text. For example, the '**kk**' sound in **takka** is represented phonetically by uppercase #KK#.

6.3 RULSYS Vowels

Table 11 lists the phonetic symbols in RULSYS used for the Finnish vowels along with example words and their transcriptions.

Vowel symbol	Example	Transcription		
A	k**a**la	#K'ALA#		
AA	t**aa**kka	#T'AAKKA#		
E	k**e**ttu	#K'ETTU#		
EE	**ee**va	#'EEVA#		
I	k**i**ta	#K'ITA#		
II	v**ii**ri	#V'IIRI#		
O	k**o**ko	#K'OKO#		
OO	t**oo**ni	#T'OONI#		
U	m**u**ki	#M'UKI#		
UU	k**uu**la	#K'UULA#		
Y	**y**li	#'YLI#		
YY	t**yy**li	#T'YYLI#		
Ä or [	t**ä**mä	#T'ÄMÄ#	or	#T'[M[#
ÄÄ or [[	t**ää**llä	#T'ÄÄLLÄ#	or	#T'[[LL[#
Ö or \	k**ö**hä	#K'ÖHÄ#	or	#K'\H[#
ÖÖ or \\	t**öö**lö	#T'ÖÖLÖ#	or	#T'\\L\#

Table 11 RULSYS vowels

Note that #Ä# and #[# are equivalent as well as #Ö# and #\#.

6.4 Comments on phonetic symbols for vowels

6.4.1 Long and short vowels

In Finnish there are two distinct vowel lengths for all the vowels. The long variants of the vowels are roughly twice the length of the short ones Long vowel sounds are in RULSYS represented by two identical letters in the orthography and, similarly, two identical symbols in phonetic text.

Example: **kuula** is transcribed #K'UULA#

6.5 Extra symbols for phonetic details

In the current version of the Finnish 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.

6.5.1 Lexical stress

For a description of lexical stress, see section 5.5.1. In phonetic text, primary stress is in RULSYS denoted by an apostrophe < ' > placed immediately before the stressed vowel. Secondary stress is denoted by an acute accent mark (reversed apostrophe) < ` >.

It is also possible to mark the stress in orthographic text. If the primary stress is on another vowel than the first (which is common in foreign words), it can be marked by an apostrophe.

Example: **Karl'eelliasviten.**

Secondary stress in orthographic text is denoted by an at-sign < **@** > or the acute accent mark < `>.

Example The word **sanomalehti** is transcribed `/s A4 n O m A l e4 h t i/.`
The same pronunciation is achieved if the word is written
sanomal@ehti or **sanomal`ehti**.

It is important to have stress marks in a sentence written in phonetic text. Generally there should only be one stress mark per word. If no stress marks appear in a sentence at all, the system will produce a monotone reading of the sentence.

Remember that only vowels are stressed, i.e., a stress mark must be followed by a vowel written in phonetic characters. Be sure, for example, that you do not leave a real apostrophe or real quotes in phonetic text.

6.5.2 Punctuation marks

Punctuation marks are also permitted in phonetic text, and have the same effect as in normal text, affecting both the rhythm and intonation of the sentence. These punctuation characters are permitted in phonetic text:

. , ? !

The character < ' > has a completely different function in phonetic text than in orthographic text. It is a reserved character used to mark stress in a word, see section 6.5.1. It cannot be used to quote text or single words in phonetic text.

6.5.3 Hyphen

The 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 3.3.1.

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 section 5), 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

As Finnish orthography follows Finnish phonology very closely, the transformation into phonetic symbols will be correct in most cases. Unlike some other language versions where a deliberate misspelling can be used to produce a correct pronunciation when the system's automatic conversion produces the wrong result, the pronunciation of the Finnish system will rarely benefit from misspellings. Exceptions are cases where the pronunciation of a word deviates markedly from the spelling, as may be the case with non-Finnish names and abbreviations, for instance.

The system does not correctly predict secondary stresses in most cases; therefore it may be possible to improve on the pronunciation of, for instance, compound words and other long words by inserting either a hyphen < **-** > or a reverse apostrophe < ` > (or < **@** >), as explained in Section 5.5.1

7.1.2 Expanding acronyms

Not very many acronyms are handled by the current Finnish system (see section 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.

Examples	**IS**	Ilta Sanomat
	VPK	Vapaa Palokunta

7.2 Using phonetic text

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 8 and 9.

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 8 and 9 give the pronunciation key for the special phonetic symbols used in Finnish 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 8 and 9, convert the dictionary pronunciation symbols to the appropriate Finnish 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

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.

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

8.1 Abbreviations recognised in SM

Abbreviation	SM
ab	*aabee*
cm	*senttimetriä**
db	*desibeliä*
dl	*'desilitraa*
dm	*desimetriä**
em	*edellämainittu*
emt	*edellämainittu teos*
ens	*ensimmäinen*
ent	*entinen*
erik	*erikoisesti*
esim	*esimerkiksi*
etc	*etsetera*
fim	*suomen markkaa*
fk	*filosofian kandidaatti*
hg	*hehtogrammaa*
hl	*hehtolitraa*
hm	*hehtometriä*
hra	*herra*
hz	*hertsiä*
jatk	*jatkuu*
jksk	*joksikin*
jllak	*jollakin*
jllek	*jollekin*
jltk	*joltakin*
jm	*juoksumetriä*
jms	*ja muuta sellaista*
jnak	*jonakin*
jne	*ja niin edelleen*
jnk	*jonkin*
jtk	*jotakin*
kal	*kaloria*
kg	*kiloa*
klo	*kello*
ko	*kyseessä oleva*
kpl	*kappale*
kr	*kruunua*
lk	*luokka*
mg	*milligrammaa*
milj	*miljoonaa*
min	*minuuttia*
mk	*markkaa*
ml	*millilitraa*
mm	*muun muuassa*/***
mrd	*miljardia*
nk	*niin kutsuttu*
ns	*niin sanottu*
nti	*neiti*
ok	*okei*
oy	*ooyy*
pnä	*päivänä*
puh	*puhelin*
pv	*päivää*
pvm	*päivämäärä*
rva	*rouva*
ry	*ärryy*
tl	*teelusikallista*
tn	*tonnia*
tri	*tohtori*
ts	*toisin sanoen*
vk	*viikkoa*
vrk	*vuorokautta*
vrt	*vertaa*
yht	*yhteensä*
ym	*ynnä muuta*
yms	*ynnä muuta sellaista*

Table 12 Abbreviations recognised in SM

* Regarding the reading of **cm²**, **dm²**, **km²**, and **mm²** see 8.3

** For the reading of **mm** as **millimetriä** see 8.2 and 8.3.

8.2 Abbreviations only recognised in SM in connection with digits

The abbreviations in Table 13 are read as abbreviations after digits only, and (with the exception of **mm** and **ha**) only in SM.

Abbreviation	LM	SM
g	*G*	*grammaa*
h	*H*	*tuntia*
ha	*HA*	*hehtaaria*
j	*J*	*joulea*
l	*L*	*litraa*
m	*M*	*metriä**
mm	*M M*	*millimetriä***
s	*S*	*sekuntia*
t	*T*	*tuntia*
v	*V*	*vuotta*

Table 13 Abbreviations only recognised in SM in connection with digits

* **m** will be read as the name of the letter unless a different reading (for instance as *metriä*) is entered into the user lexicon. **m²** is, however, read as *neliömetri* by default. Again, if the user wants a different reading it is necessary to enter this in the user lexicon. See section 8.3 for instructions on what to enter into the user lexicon.

** **mm** is read as *millimetriä* if a digit immediately precedes it, and in the combination **mm²**. In other cases it is read as *muun muuassa*. See section 8.3 for instructions on changing the reading.
The letter **n** on its own is treated as an abbreviation for **noin**. In other cases the letter **n** is read as the name of the letter.

8.3 Abbreviations of units of area and distance

The abbreviations **cm²**, **dm²**, **km²**, **mm²**, and m² are read as units of area in SM only. In other cases < **²** > is read as *toiseen*, Table 14.

Abbreviation	LM	SM
cm²	*C M toiseen*	*neliösenttimetriä*
dm²	*D M toiseen*	*neliödesimetriä*
km²	*K M toiseen*	*neliökilometriä*
mm²	*M M toiseen*	*neliömillimetriä*
m²	*M toiseen*	*neliömetriä*

Table 14 Abbreviations of area that are recognised in SM only

It is possible, using the user lexicon, to achieve special readings of these instead of the system's default readings, see Table 15.

User lexicon entry		Effect on reading, all reading modes
M	#'EM# or /e4 m/	**m** always read as *m*
M	#@METR# or /_@ m e t r/	**m** always read as *metriä*
MM	#@MM# or /_@ m m/	**mm** always read as *millimetriä*
MM	#@MUMU# or /_@ m u m u/	**mm** always read as *muun muuassa*

Table 15 Entries in user lexicon that result in special readings of the area abbreviations