Language Manual

Icelandic

Snorri

Language Manual
Icelandic
Snorri
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 Icelandic 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 Snorri.

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 Icelandic characters are considered letters and may thus constitute a word. The apostrophe character < **'** > and the at-sign < **@** > can also occur among letters within words.

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 Icelandic letters

Most of the special Icelandic characters are converted internally in the system to a two-character code. Table 1 lists the Icelandic characters that are converted and their codes. The internal codes may be used instead of the characters in orthographic text. If appearing alone, the characters and their codes have identical readings in SM, i.e. both ð and `'d` are read as *eðð*.

Letter	Example	Internal code	Example
Þ	Þorgeir	'T	'Torgeir
þ	þota	't	'tota
Đ	BIÐJA	'D	BI'DJA
ð	baða	'd	ba'da
Á	Árni	'A	'Arni
á	láta	'a	l'ata
É	Ég	'E	'Eg
é	mér	'e	m'er
Í	Ísland	'I	'Island
í	blístra	'i	bl'istra
Ó	Ómar	'O	'Omar
ó	tómur	'o	t'omur
Ú	Úlfur	'U	'Ulfur
ú	bútur	'u	b'utur
Ý	Ýrr	'Y	'Yrr
ý	flýta	'y	fl'yta

Table 1 Icelandic characters and their internal codes

The characters **Æ**, and **æ**, **o** are not converted and are read as *Æ, Ö* and *æ, ö* in all reading modes.

2.2 Other Scandinavian letters

The special letters of other Scandinavian languages are read as indicated in Table 2 below.

Character	SM	LM
ä	*e*	*a með tvípunkti*
å	*a*	*bollu a*
Å	*A*	*(stórt) bollu A*
Ä	*E*	*(stórt) A með tvípunkti*
ø	*ö*	*gegnumstrikað o*
Ø	*Ö*	*(stórt) gegnumstrikað O*

Table 2 Scandinavian letters in the Icelandic system

2.3 Other European letters

ß 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 *beta*. **C** and **c** (C cedilla) are read as *s* in SM when next to other letters.

Vowels with non-Icelandic accent marks < ` ^ > or a trema (diaeresis:< ¨ >) are mostly read as the corresponding vowels without the marks. In LM the diacritic used is named. If desired, words containing accent marks and other diacritic symbols can be entered in the user lexicon. Note that accented letters in the look-up word are converted to the internal one- or two-character code).

Character	SM	LM
ü	*u*	*u með tvípunkti*
â	*a*	*a með sirkumflex*
à	*a*	*a með bakfallsbroddi*
ê	*e*	*e með sirkumflex*
ë	*e*	*e með tvípunkti*
è	*e*	*e með bakfallsbroddi*
ô	*o*	*o með sirkumflex*
û	*u*	*u með sirkumflex*
Ü	*u*	*u með tvípunkti*
ç	*s*	*lykkju c*
Ç	*S*	*(stórt) lykkju C*

Table 3 Reading of non-Scandinavian 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. 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 that are always pronounced, with no effect on the intonation or pausing.
- Control characters

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 4 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
,	*komma*
:	*tvípunktur*
;	*semíkomma*
"	*gæsalöpp*
.	*punktur*
?	*spurningarmerki*
!	*upphrópunarmerki*
(	*svigi opnast*
)	*svigi lokast*

Table 4 Punctuation characters

3.1.1 Comma, colon and semicolon

Comma < **,** >, colon < **:** > and semicolon < **;** > cause a short pause to occur along with a rising intonation. Comma is used as decimal marker in numbers, see section 4.5.

3.1.2 Quotation marks

Quotes < **“** > appearing around a single word or a group of words cause a brief pause before and after the quoted text.

3.1.3 Full stop

A full stop < **.** > terminates a sentence and causes a falling, sentence-final intonation pattern to occur along with a somewhat longer pause. Full stop has a special function in numbers, see section 4.1.

3.1.4 Question mark

A question mark < **?** > terminates a sentence, but may cause either a rise or a fall in the intonation pattern depending on the type of sentence.

- For Yes/No type questions (i.e. a question requiring a Yes or No as an answer), there is a rising intonation contour.

- For questions beginning with HV (i.e. a question that begins with question words like **hver**, **hvað**, **hvar**, **hvenær**, **hvers vegna**, **hvernig**, etc.), there is a falling intonation pattern.

3.1.5 Exclamation mark

An exclamation mark < **!** > functions in a way similar to the full stop, but gives a somewhat sharper fall in the intonation pattern. It is often helpful to use the emphatic stress mark (see section 5.6.2) to signal the key word of a phrase, thus adding more expression to the sentence.

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
-	*strik*	(see 3.2.1 and 3.2.5)
=	*samasem*	(see 3.2.3 and 5.2.5)
+	*plús*	*plús* (see 5.2.5)
*	*stjarna*	(see 5.2.2 and 5.2.5)
#	*kross*	*kross* (see 5.2.5)

Table 5 Characters with varying pronunciation

See section 3.2.4 for the reading of multiple occurrences of the characters < **- = + * #** >.

All examples below show the reading in SM.

3.2.1 Hyphen

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

- If surrounded by spaces or punctuation marks, it is pronounced as a short pause.
- When used to mark compound words, it is not pronounced.
- 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; the word is not spoken until the second part in the next line has also been processed.
- In most other cases, hyphen is pronounced *strik*.

Expression	**Reading**
`550522-4758`	*550522 strik 4758*
`þú segir ekki - er þetta satt?`	*þú segir ekki* `(silence)` *er þetta satt?*
`inn-póstur`	*innpóstur*
`Orðum er gjarnan skipt með band-<CR>` `striki.`	*Orðum er gjarnan skipt með bandstriki*
`kerfi-<CR>` `nu`	*kerfinu*
`--bc.`	*strik strik B C*

3.3 Equals sign

Equals sign < = > is pronounced *eru* if a number follows it, in other cases it is pronounced *samasem*.

Expression	**Reading**
`2*3=6`	*tveir sinnum þrír eru sex*
`cb==bc`	*C B samasem B C*

3.3.1 Asterisk

Asterisk < ***** > is pronounced *sinnum* in SM if a digit immediately precedes and follows it. It is pronounced *stjarna* in all other cases.

Expression	**Reading**
`2*3`	*tveir sinnum þrír*
`*bc`	*stjarna B C*

3.3.2 Multiple occurrences of the same character

In SM, if more than three of the same character 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**
`*****`	*stjarna stjarna stjarna*
`+++++`	*plús plús plús*
`-----`	*strik strik strik*
`=====`	*samasem samasem samasem*
`#####`	*kross kross kross*

3.4 Other non-alphanumeric characters

Apart from the special case noted below, each of the characters in Table 6 are pronounced in all reading modes.

Character	SM	LM
#	*kross*	*kross*
%	*prósent*	*prósent*
&	*og*	*og*
/	*skástrik*	*skástrik*
'	*úrfellingarmerki*	*úrfellingarmerki*
@	*at merki*	*at merki*
<	*minna en*	*minna en*
>	*stærra en*	*stærra en*
£	*pund*	*pund*
Pt	*peseti*	*peseti*
¿	(silence)	*öfugt spurningarmerki*
½	*en halv*	*hálfur*
¼	*en kvart*	*einn fjórði*
¡	(silence)	*öfugt upphrópunarmerki*
α	*alfa*	*alfa*
β	*beta* (see 2.3)	*beta*
Γ	*stórt gamma*	*stórt gamma*
π	*pí*	*pí*
Σ	*summa*	*summa*
∞	*óendanlegt*	*óendanlegt*
≡	*eins og*	*eins og*
±	*plús mínus*	*plús mínus*
≥	*stærra en eða samasem*	*stærra en eða samasem*
≤	*minna en eða samasem*	*minna en eða samasem*
≈	*næstum jafnt og*	*næstum jafnt og*
°	*gráða*	*gráða*
•	*depill*	*depill*
·	*hækkaður punktur*	*hækkaður punktur*
√	*kvaðratrót*	*kvaðratrót*
²	*í öðru veldi*	*í öðru veldi*
	autt sæti	*autt sæti*

Table 6 Other non-alphanumeric characters

Special case:
km², **mm²**, **cm²** and **m²** are read as *ferkílómetri* or *ferkílómetrar*, *fermillimetri* or *fermillimetrar* etc. The forms ending in **-i** are used after the digit **1** (or digit strings ending in **1**). Those ending in **-ar** are used in all other cases.

km, **mm** and **cm** are also read as abbreviations, see 8.1, but **m** on its own is read as the name of the letter; to have it read differently it is necessary to enter it in the user lexicon.

3.5 Control characters

In LM (only) most control characters are read out. In most cases they are read as *stýring* + the appropriate letter, e.g. **^T** is read as *stýring 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>	*strok aftur*
^I <TAB>	*dálkhak*
^K <VERTICAL TAB>	*lóðrétt dálkhak*
^L <LINE FEED>	*nýlínustafur*
^M <RETURN>	*vending*
^? <DELETE>	*strok áfram*

Table 7 Control characters with special names read in letter mode

3.6 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 Case and gender of Icelandic numerals
- 4.3 Year reading
- 4.4 Leading zero
- 4.5 Decimal numbers
- 4.6 Monetary amounts
- 4.7 Arithmetic operators
- 4.8 Mixed digits and letters

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

Note that the system does not handle formats for time of day, dates and monetary amounts.

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 (comma).

Numbers denoting thousands, millions and billions (numbers larger than 999) may be grouped in groups of three digits starting from the end of the number or from a decimal marker. Full stop (not space) is used as separator. If a number string containing full stops does not consist of groups of exactly three digits (one, two or three digits in the case of the group preceding the first full stop) the string will be broken up into separate groups of numbers. Each group will be read according to the ordinary number rules and the full stops will be read as *punktur*.

The highest number read is 999.999.999.999.999 (fifteen digits). Numbers higher than that are read digit by digit.

Number	**Reading**
2425	*tvöþúsundfjögurhundruðtuttuguogfimm*
1000000005	*einn milljarður*
123456789012	*eitthundraðtuttuguogþrír milljarðar fjögurhundruðfimmtíuogsex milljónir sjöhundruðáttatíuogníuþúsund og tólf*
2.425	*tvöþúsundfjögurhundruðtuttuguogfimm*
22 000	*tuttuguogtveir núll núll núll*
1.2988	*einn punktur tvö þúsundníuhundruðáttatíuogátta*
1.1988	*einn punktur eittþúsundníuhundruðáttatíuogátta*
1.198.238000	*einn punktur eitthundruðníutíuogátta punktur tvöhundruðáttatíu þúsund*

4.2 Case and gender of Icelandic numerals

The gender and case of the numbers 1, 2, 3 and 4 are not correctly processed by the Icelandic system. When gender and case cannot be determined, the system uses the masculine nominative singular as the default form for the number.

Input	**Reading**
`hann keypti 2 banana og 3 snúða`	*hann keypti tveir banana og þrír snúða*
`hér eru 23 konur og 34 menn`	*hér eru tuttuguogþrír konur og þrjátíuogfjórir menn*

4.3 Year reading

A four digit number between 1100 and 1999 is read as hundreds (year reading). If the digit string contains a full stop or a comma, the digit string is read according to the number processing rules above.

Expression	**Reading**
`1088`	*eittþúsundáttatíuogátta*
`1900`	*nítjánhundruð*
`1988`	*nítjánhundruðáttatíuogátta*
`1.988`	*eittþúsundníuhundruðáttatíuogátta*
`1988,0`	*eittþúsundníuhundruðáttatíuogátta komma núll*
`1988.0`	*eittþúsundníuhundruðáttatíuogátta punktur núll*

4.4 Leading zero

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

Number	**Reading**
`09253`	*núll níu tveir fimm þrír*
`0210`	*núll tveir einn núll*

4.5 Decimal numbers

Decimal amounts are properly indicated with a comma in the Icelandic system. The full number part to the left of the decimal marker is read according to the number processing rules above.

Number	**Reading**
`16,234`	*sextán komma tvöhundruðþrjátíuogfjórir*
`3,1415`	*þrír komma eittþúsundfjögurhundruðogfimmtán*
`1251,04`	*eittþúsundtvöhundruðfimmtíuogeinn komma núll fjórir*
`2,50`	*tveir komma fimmtíu*
`2.50`	*tveir punktur fimmtíu*
`3.141`	*þrjúþúsundeitthundraðfjörutíuogeinn*

4.6 Monetary amounts

No special arrangements are made for the processing of monetary amounts in the current version of Icelandic. Thus monetary amounts are read as ordinary numburs. Still the abbreviation **kr** is read *krónur*, see section 8.

Number	**Reading**
`kr 2988,45`	*krónur tvöþúsundníuhundruðáttatíuogátta komma fjörutíuogfimm*
`1988,45 kr`	*eittþúsundníuhundruðáttatíuogátta komma fjörutíuogfimm krónur*

4.7 Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below. See also section 3.2. Note that neither the time of day format **hours:minutes** nor the date format **day/month/year** are supported in the current version of the Icelandic system.

Expression	**Reading**
-12	*strik tólf*
+24	*plús tuttuguogfjórir*
2*3	*tveir sinnum þrír*
25%	*tuttuguogfimm prósent*
3,4%	*þrír komma fjórir prósent*
,05%	*komma núll fimm prósent*
10.30	*tíu punktur þrír núll*

4.8 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	*sjötíuogsjö B áttatíuogfjórir Z þrír*
0092B87-B	*núll núll níu tveir B áttatíuogsjö* (pause) *B*
208kr	*tvöhundruðogátta krónur*
0092B87-B	*null null ni to B áttisju strek B*

5 Icelandic 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 Icelandic voice Snorri). 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 Icelandic system uses a phonetic alphabet based on standard 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 A note on dialect

It should be noted that the system approximates the northern dialect of Icelandic and thus the transcription reflects the phonetics of the northern dialect. This means that it may prove difficult to render correctly certain aspects of the southern dialect in the phonetic transcription.

The differences between the dialects can be summed up in two points. First, plosives spelled **p**, **t** and **k** are aspirated in the northern dialect whereas in the southern dialect they are not. To render the southern pronunciation, simply use the phonetic symbols /**b**/, /**d**/ and /**g**/ instead of /**p**/, /**t**/ and /**k**/.

Spelling	**Northern dialect**	**Southern dialect**
leita	`/l eI4 t a/`	`/l eI4 d a/`
flaka	`/f l a4 k a/`	`/f l a4 g a/`

The other difference between the dialects involves the voicing of nasals before plosives spelled with **p**, **t** or **k**. In the northern dialect the nasals are voiced before **p**, **t** and **k**, but in the southern dialect they are unvoiced. It is not possible to correctly render the unvoiced nasals before plosives in the current version of the Icelandic system.

5.2 Consonants

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

Consonant symbol	Example	Transcription
b	**b**o**bb**i	/b O4 b b I/
	he**f**la	/h e4 b l a/
d	**D**i**dd**a	/d I4 d d a/
	ha**ll**a	/h a4 d l a/
	by**r**la	/b I4 r d l a/
	sei**nn**i	/s eI4 d n I/
	þo**r**na	/T O4 r d n a/
D	ba**ð**a	/b a4 D a/
f	**f**ótur	/f OU4 t U r/
	só**f**i	/s OU4 f I/
g	**g**a**gg**a	/g a4 g g a/
G	sa**g**a	/s a4 G a/
h	**h**eitur	/h eI4 t U r/
	pa**kk**a	/p a4 k k a/ or /p a4 h g a/
	e**pl**i	/e4 p l I/ or /e4 h p l I/
j	**j**æ**j**a	/j aI4 j a/
	se**g**i	/s eI4 j I/
k	**k**a**k**a	/k a4 k a/
x	sa**x**a	/s a4 x s a/
	sa**g**t	/s a4 x d/
	lo**k**sins	/l O4 x s I n s/
	dö**kk**t	/d 94 x d/
l	**l**ú**ll**a	/l u4 l l a/
l_0	hal**t**ur	/h a4 l_0 d U r/
m	**m**a**mm**a	/m a4 m m a/
n	**n**e**nn**a	/n e4 n n a/
N	ha**ng**a	/h aU4 N g a/
p	**p**i**p**ar	/p I4 p a r/
r	**r**ý**r**a	/r I4 r a/
r_0	kva**r**ta	/k v a4 r_0 d a/
s	**s**ó**s**a	/s OU4 s a/
t	**t**æ**t**a	/t aI4 t a/
T	**þ**eyta	/T eI4 t a/
v	**v**ita	/v I4 t a/
	ley**f**a	/l eI4 v a/

Table 8 Icelandic consonant symbols in SAMPA

5.3 Comments on phonetic symbols for consonants

5.3.1 Duration

In phonetic transcription of Icelandic, a long consonant is denoted simply by writing the two consonants after each other separated by space. The following phonetic symbols for consonants have long variants: **/b**, **d**, **f**, **g**, **j**, **l**, **m**, **n**, **r**, **s/**.

Example	**Luma**	/l U4 m a/
	Lumma	/l U4 m m a/

5.3.2 G and x

The pronunciation of the letter **g** varies depending on which letters precede and follow it. If a vowel precedes it and a vowel or either **r** or **đ** or no letter follows it, it is pronounced as in the words **saga**, **fegra**, **bragða** and **lag**. To transcribe this sound the system uses the symbol **/G/**. Thus the correct transcription of these words would be the following (see also Table 8).

Examples	**Saga**	/s a4 G a/
	Fegra	/f e4 G r a/
	Bragða	/b r a4 G D a/
	Lag	/l a4 G/

In some cases a **g** (and **k**) is pronounced as in the words **sagt** and **dokkt**. In these cases it is transcribed as **/x**/, see Table 8.

5.3.3 Preaspiration

If a double plosive (i.e. **pp**, **tt** or **kk**) occurs in the spelling, the phonetic symbol **/h/** is inserted before the plosive. This is known as preaspiration and is exemplified by words such as **klappa**, **stokkva** and **hitta** (see Table 8).

Preaspiration also occurs if either a single or double plosive is followed by either **l**, **n** or **m**. Thus words like **epli**, **kaupmaður**, **botnar** and **okkli** contain preaspiration (see also table 8).

If you wish to make the system pronounce these sequences without the preaspiration, insert a hyphen, < **-** >, between the two letters in question in the word in your text. Note that the hyphen must be preceded by an underscore in SAMPA transcriptions. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
Grjóttunna	/g r j OU4 d U n a/	Grjót-tunna	/g r j OU4 t _- t U n n a/
Hriplak	/h r I4 h b l a k/	Hrip-lak	/h r I4 p _- l a1 k/

5.3.4 Devoicing

When **r** occurs before **p**, **t** or **k** in spelling, it becomes devoiced, as in the words **karpa** and **skurkur**. Also, when **l** occurs before **t**, **r** becomes devoiced as in the words **haltur** and **gult**. To transcribe these sounds the system uses the symbols **/l_0/** and **/r_0/** (see table 8).

If you wish to make the system pronounce these sequences without devoicing, insert a hyphen < **-** > between the two letters in question in the text or underscore followed by hyphen, < _- >, in your transcription. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
stálteinn	/s d aU4 l_0 d eI d n/	stál-teinn	/s d aU4 l _-t eI d n/
bjórkippa	/b j OU4 r_0 g I h p a/	bjór-kippa	/b j OU4 r _- k I h b a/

5.3.5 Spelling considerations

A rule applies to the orthographic clusters **RN**, **NN**, **RL** and **LL** which inserts a **/d/** between the two letters in most cases (and deletes the first **N** in **NN**). This is exemplified by the pronunciation of words such as **halla**, **byrla**, **seinni** and **þorna** (see table 8). To make the system pronounce the clusters without the **/d/** a hyphen < **-** > must be inserted between the two letters in question. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
Carnaby	/k a4 r d n a _- b I/	Car-na-by	/k a4 r _- n a _- b i1/
Seinnuminn	/s eI4 d n U m I n/	Sein-numinn	/s eI4 n _- n U m I n/
Carla	/k a4 r d l a/	Car-la	/k a4 r _- l a/
Gullitaður	/g U4 d l I t a D U r/	Gul-litaður	/g U4 l _- l I t a1 D U r/

The system will pronounce a non-initial **fl** cluster as **/b l/**. To make the system pronounce the cluster without the **/b/** there are two alternatives: a hyphen < **-** > can be inserted between the two letters if the cluster is to be pronounced **/v l/**; or a hyphen can be inserted before the two letters if the cluster is to be pronounced **/f l/**. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
Aflima	/a4 b l I m a/	Af-lima	/a4 v _- l I m a/
Kvennaflangs	/k v e4 n n a b l aU N s/	Kvenna-flangs	/k v e4 n n a _- f l aU1 N s/

5.4 Vowels

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

Vowel symbol	Example	Transcription
a	s**a**t	/s a4 t/
e	b**e**r	/b e4 r/
I	v**i**l	/v I4 l/
i	st**í**l	/s d i4 l/
O	b**oð**	/b O4 D/
U	sk**u**t	/s g U4 t/
u	st**ú**t	/s d u4 t/
9	f**ö**l	/f 94 l/
aI	h**æ**l	/h aI4 l/
eI	r**ei**m	/r eI4 m/
9y	str**au**m	/s d r 9y4 m/
aU	h**á**r	/h aU4 r/
OU	st**ór**	/s d OU4 r/
j e	v**é**l	/v j e4 l/ (see section 6.3.2)

Table 9 Icelandic vowel symbols in SAMPA

Note that /**j e**/ is not a proper SAMPA symbol but it may still be used when making transcriptions. There is a corresponding symbol in RULSYS, see next chapter.

5.5 Comments on phonetic symbols for vowels

5.5.1 Long and short vowels

It is not necessary to discriminate between long and short vowels when writing Icelandic phonetic text since the system can determine the length by rule.

5.5.2 The treatment of É and é

The text-to-speech system converts the characters **E** and **e** to **'E**. Therefore **E** and **e** can always be replaced by **'E** in the orthographic text without affecting the system's pronunciation. In the phonetic representation this is changed to /**j e**/. As a consequence, it is usually also possible to write words with **JE** instead of **E** or **e**. Examples:

Input	System text	Phonetic text
Pétur	P'ETUR	/p j e4 t U r/
frétta	FR'ETTA	/f r j e4 t t a/
fr'jetta	FR'ETTA	/f r j e4 t t a/
frjetta	FRJETTA	/f r j e4 t t a/

5.5.3 Diphthongs caused by glide insertion

When a **gi** or a **gj** follows a vowel in the orthography, the glide **/j/** emerges following the vowel. These sequences are often regarded as diphthongs but the text-to-speech system treats them as vowel+consonant sequences, not as diphthongs.

Examples	**bogi**	/bO4 j j I/
	hugi	/h U4 j j I/

5.6 Extra symbols for phonetic details

In the current version of the Icelandic 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.

5.6.1 Stress

Stress is used to indicate the level of prominence of a syllable in a word (word level stress), or of a word in a sentence (phrase level stress), see section 5.6.2 on emphasis and reduction.

On word level, two types of stress are available within a word. The strongest stress, primary stress, is represented by a < **4** > placed after the vowel to receive stress. In Icelandic the primary stress almost always falls on the first syllable of a word and therefore in most cases the stress mark < **4** > should be placed immediately after the first vowel in a word.

Examples	**Maður**	/m a4 D U r/
	Leita	/l eI4 t a/

In some cases, however, you may wish to stress some other syllable in the word, in which case you should place the stress mark immediately after the vowel that you wish to stress.

Examples	**Descartes**	/d e k a4 r_0 d/
	Jamaica	/d j a m aI4 k a/

A slightly weaker level of stress, secondary stress, is represented by a < **1** >. Vowels not followed by either a primary stress mark < **4** > or a secondary stress mark < **1** > receive no stress at all.
It is important to have stress marks in a sentence written in phonetic text. Generally there should be only one primary stress mark per word, although more than one secondary stress mark may occur in a word. If no stress marks at all appear in a phonetically transcribed sentence, the system may produce a monotone reading of the sentence.

Examples	**Domur**	/d OU4 m U r/
	Hæstarettardomur	/h aI4 s d a r j e1 h d a r d OU1 m U r/

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

5.6.2 Emphasis and reduction

It is also possible to emphasise or reduce the stress on a particular word. In the input text string, this is done by placing < **_X** >, where **X** represents a single digit between **0** and **9**, within a PRN-tag (see User's Guide) immediately before the word whose prominence is to be altered. The emphasis mark can also be used in transcriptions in a user lexicon. (See examples on next page).

_2	normal stress for most words
_0	makes a word non-stressed
_1	gives stress to a normally unstressed word
_3 - **_9**	gives levels of emphatic stress

Compare how the meaning is changed when the emphatic stress is varied in the sentence below. Note that slashes are used instead of the necessary PRN-tag (see User's Guide) in the examples below.

Examples
```
Hann keypti átján rauðar rósir handa mér.
Hann keypti /_7/ átján rauðar rósir handa /_0/ mér.
Hann keypti átján rauðar rósir handa /_7/ mér.
```

These emphatic stress marks may also appear in a phonetic text string and must be placed at the beginning of the string.

Example:
```
Hann keypti átján rauðar /6 h i4 j a s I1 n t U r/ handa mér.
```

5.6.3 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.

5.6.4 Hyphen

In phonetic text, hyphen (in SAMPA underscore + 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 Icelandic 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
- there are symbols for blocking and unstressed /E/

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

Consonant symbol	Example	Transcription
B	**b**o**bb**i	#B"OBBI#
	he**f**la	#H"EBLA#
D	**D**i**dd**a	#D"IDDA#
	ha**l**la	#H"ADLA#
	by**r**la	#B"IRDLA#
	sei**n**ni	#S"EIDNI#
	þo**r**na	#th"ORDNA#
dh	ba**ð**a	#B"AdhA#
F	**f**ótur	#F"'OTUR#
	só**f**i	#S"'OFI#
G	**g**a**gg**a	#G"AGGA#
GH	sa**g**a	#S"AGHA#
H	**h**eitur	#H"EITUR#
	pa**kk**a	#P"AKKA# or #P"AHGA#
	e**pl**i	#"EPLI# or #"EHPLI#
J	**j**æ**j**a	#J"[JA#
	se**g**i	#S"EIJI#
K	**k**a**k**a	#K"AKA#
KH	sa**x**a	#S"AKHSA#
	sa**g**t	#S"AKHD#
	lo**k**sins	#L"OKHSINS#
	dö**kk**t	#D"\KHD#
L	**l**ú**ll**a	#L"'ULLA#
l	hal**t**ur	#H"AlDUR#
M	**m**a**mm**a	#M"AMMA#
N	**n**e**nn**a	#N"ENNA#
NG	ha**ng**a	#H"'ANGGA#
P	**p**i**p**ar	#P"IPAR#
R	**r**ý**r**a	#R"IRA#
r	kva**r**ta	#KV"ArDA#
S	**s**ó**s**a	#S"'OSA#
T	**t**æ**t**a	#T"[TA#
th	**þ**eyta	#th"EITA#
V	**v**ita	#V"ITA#
	ley**f**a	#L"EIVA#

Table 10 RULSYS consonants

6.2 Comments on phonetic symbols for consonants

6.2.1 Duration

In phonetic transcription of Icelandic, a long consonant is denoted simply by doubling the consonant. The following phonetic symbols for consonants have long variants: #**B**, **D**, **F**, **G**, **J**, **L**, **M**, **N**, **R**, **S**#.

Examples	**Luma**	`#L"UMA#`
	Lumma	`#L"UMMA#`

6.2.2 GH and KH

The pronunciation of the letter **g** varies depending on which letters precede and follow it. If a vowel precedes it and a vowel or either **r**, **ð** or no letter follows it, it is pronounced as in the words **saga**, **fegra**, **bragða** and **lag**. To transcribe this sound the system uses a combination of two characters, `#GH#`. Thus the correct transcription of these words would be the following (see also Table 10).

Examples	**Saga**	`#S"AGHA#`
	Fegra	`#F"EGHRA#`
	Bragða	`#BR"AGHdhA#`
	Lag	`#L"AGH#`

In some cases a **g** (and **k**) is pronounced as in the words **sagt** and **dokkt**. In these cases it is transcribed as `#KH#` (see Table 10).

6.2.3 Preaspiration

If a double plosive (i.e. **pp**, **tt** or **kk**) occurs in the spelling, an `#H#` is inserted before the plosive. This is known as preaspiration and is exemplified by words such as **klappa**, **stokkva** and **hitta** (see Table 10). Preaspiration also occurs if either a single or double plosive is followed by either **l**, **n** or **m**. Thus words like **epli**, **kaupmaður**, **botnar** and **okkli** contain preaspiration (see also Table 10).

If you wish to make the system pronounce these sequences without the preaspiration, insert a hyphen < **-** > between the two letters in question in the word in your text. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
`Grjóttunna`	`#GRJ"'OHDUNA#`	`Grjót-tunna`	`#GRJ"'OT-TUNNA`
`Hriplak`	`#HR"IHBLAK#`	`Hrip-lak`	`#HR"IP-L`AK#`

6.2.4 Devoicing

When **r** occurs before **p**, **t** or **k** in the spelling, it becomes devoiced, as in the words **karpa** and **skurkur**. Also, when **l** occurs before **t, r** becomes devoiced as in the words **haltur** and **gult**. To transcribe these sounds the system uses the lower-case version of the letters, i.e. `#l#` and `#r#` (see Table 10).

If you wish to make the system pronounce these sequences without devoicing, insert a hyphen < **-** > between the two letters in question in the text or in your transcription. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
`stálteinn`	`#SD"'AlDEIDN#`	`stál-teinn`	`#SD"'AL-TEIDN#`
`björkippa`	`#BJ"'OrGIHPA#`	`björ-kippa`	`#BJ"'OR-KIHBA#`

6.2.5 Spelling considerations

A rule applies to the clusters **rn**, **nn**, **rl** and **ll** which inserts a #D# between the two letters in most cases (and deletes the first **n** in **nn**). This is exemplified by the pronunciation of words such as **halla**, **byrla**, **seinni** and **þorna** (see Table 10). To make the system pronounce the clusters without the #D# a hyphen < **-** > must be inserted between the two letters in question. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
Carnaby	#K"ARDNA-BI#	Car-na-by	#K"AR-NA-B`'I#
Seinnuminn	#S"EIDNUMIN#	Sein-numinn	#S"EIN-NUMIN#
Carla	#K"ARDLA#	Car-la	#K"AR-LA#
Gullitaður	#G"UDLITAdhUR#	Gul-litaður	#G"UL-LIT`AdhUR#

The system will pronounce a non-initial **fl** cluster as #**BL**#. To make the system pronounce the cluster without the #**B**#there are two alternatives: a hyphen < **-** > can be inserted between the two letters if the cluster is to be pronounced #**VL**#*;* or a hyphen can be inserted before the two letters if the cluster is to be pronounced #**FL**#. Examples:

Word	**System transcription**	**Altered text**	**Corrected transcription**
Aflima	#"ABLIMA#	Af-lima	#"AV-LIMA#
Kvennaflangs	#KV"ENNABL'ANGS#	Kvenna-flangs	#KV"ENNA-FL`'ANGS#

6.3 RULSYS Vowels

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

Vowel symbol	Example	Transcription
A	s**a**t	#S"AT#
E	b**e**r	#B"ER#
I	v**i**l	#V"IL#
'I	st**í**l	#SD"'IL#
O	b**o**ð	#B"Odh#
U	sk**u**t	#SG"UT#
'U	st**ú**t	#SD"'UT#
\ or Ö	f**ö**l	#F"\L#
[or Æ	h**æ**l	#H"[L#
EI	r**ei**m	#R"EIM#
AU	str**au**m	#SDR"AUM#
'A	h**á**r	#H"'AR#
'O	st**ó**r	#SD"'OR#
'E	v**é**l	#VJ"EL#
		(see Section 6.2.2)

Table 11 RULSYS vowels

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

6.3.1 Long and short vowels

It is not necessary to discriminate between long and short vowels when writing phonetic text since the system can determine the length by rule.

6.3.2 The treatment of É and é

The text-to-speech system converts the characters **E** and **e** to **'E**. Therefore **E** and **e** can always be replaced by **'E** in the orthographic text without affecting the system's pronunciation. In the phonetic representation this is changed to #**JE**#. As a consequence, it is usually also possible to write words with **JE** instead of **E** or **e**. Examples:

Input	System text	Phonetic text
Pétur	P'ETUR	#PJ"ETUR#
frétta	FR'ETTA	#FRJ"ETTA#
fr'jetta	FR'ETTA	#FRJ"ETTA#
frjetta	FRJETTA	#FRJ"ETTA#

6.3.3 Diphthongs caused by glide insertion

When a **gi** or a **gj** follows a vowel in the orthography, the glide *#J#* emerges following the vowel. These sequences are often regarded as diphthongs but the text-to-speech system treats them as vowel+consonant sequences, not as diphthongs.

Examples	**bogi**	#B"OJJI#
	hugi	#H"UJJI#

6.4 Extra symbols for phonetic details

6.4.1 Stress marks

Stress is used to indicate the level of prominence of a syllable in a word (word level stress), or of a word in a sentence (phrase level stress, see section 6.4.2 on emphasis and reduction ,).

On word level, two types of stress are available within a word. The strongest stress, primary stress, is represented by a quote < " > placed before the vowel to receive stress. In Icelandic the primary stress almost always falls on the first syllable of a word and therefore in most cases the stress mark < " > should be placed immediately before the first vowel in a word.

Examples	**Maður**	#M"AdhUR#
	Leita	#L"EITA#

In some cases, however, you may wish to stress some other syllable in the word, in which case you should place the stress mark immediately before the vowel that you wish to stress.

Examples	**Descartes**	#DEK"ArD#
	Jamaica	#DJAM"ÆKA#

A slightly weaker level of stress, secondary stress, is represented by a < ` >. Vowels not preceded by either a primary stress mark < " > or a secondary stress mark < ` > receive no stress at all.
It is important to have stress marks in a sentence written in phonetic text. Generally there should be only one primary stress mark per word, although more than one secondary stress mark may occur in a word. If no stress marks at all appear in a phonetically transcribed sentence, the board may emit a tone to signal the error, or may produce a monotone reading of the sentence.

Examples	**Domur**	#D"'OMUR#
	Hæstarettardomur	#H"ÆSDARJ`EHDARD`'OMUR#

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 real < " > or < ` > in phonetic text.

6.4.2 Emphasis and reduction

It is also possible to emphasise or reduce the stress on a particular word. In the input text string, this is done by placing < **_X** >, where **X** represents a single digit between **0** and **9**, within a PRN-tag (see User's Guide) immediately before the word whose prominence is to be altered. The emphasis mark can also be used in transcriptions in a user lexicon.

_2	normal stress for most words
_0	makes a word non-stressed
_1	gives stress to a normally unstressed word
_3 - **_9**	gives levels of emphatic stress

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

Examples

```
Hann keypti átján rauðar rósir handa mér.
Hann keypti #_7# átján rauðar rósir handa #_0# mér.
Hann keypti átján rauðar rósir handa #_7# mér.
```

These emphatic stress marks may also appear in a phonetic text string and must be placed at the beginning of the string. Example:

```
Hann keypti átján rauðar #6H"'IJAS`INTUR# handa mér.
```

6.4.3 Punctuation marks

The punctuation marks < **, . ! ?** > used in phonetic text have the same effect on intonation as when appearing in orthographic text. The characters < " > and < ` > function as stress markers, see section 6.4.1, and can only be used as such in phonetic text.

6.4.4 Hyphen

The use of hyphen in phonetic text has been described in section 6.2.3 – 6.2.5. A hyphen can also be used to change the pronunciation in orthographic text, see section 7.1.2.

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 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

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	**aflima**	can be misspelled as	**avlima**
	grutlinur		**grut linur**
	New York		**Nju Jork**
	John Major		**Djonn Meidjor**

7.1.2 Use of hyphen

A hyphen can be used within a word to separate two letters that might otherwise be incorrectly pronounced together or incorrectly shorten the preceding vowel.

Examples	**garnaflækja**	can be written as	**garna-flækja**
	tjasnið		**tja-snið**
	struflanir		**s-truflanir**
	hrorlegur		**hror-legur**

7.2 Using phonetic text

When 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 4 and 5.

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 Icelandic for the text-to-speech converter.

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 Icelandic 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 current version of the Icelandic text-to-speech system recognises some abbreviations that work in all contexts, in SM only. These abbreviations are case-insensitive and do not require a full stop in order to be processed as abbreviations. If a full stop accompanies the abbreviation, the sentence is terminated at the abbreviation.

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

Sections 8.1 and 8.2 list the abbreviations that are recognised by the current version.

8.1 Abbreviations recognised in SM

In the current version of the Icelandic text-to-speech system, the abbreviations in Table 12 are recognised in all contexts in SM.

Abbreviation	LM	SM
cm	*c m*	*sentimetri/sentimetrar*
dr	*d r*	*doktor*
frmh	*f r m h*	*framhald*
hr	*h r*	*herra*
jr	*j r*	*djúnior*
kg	*k g*	*kílogramm*
kl	*k l*	*klukkan*
klst	*k l s t*	*klukkustundir*
km	*k m*	*kílómetri/kílómetrar*
mkr	*m k r*	*milljónir króna*
mm	*m m*	*millimetri/millimetrar*
nr	*n r*	*númer*
sek	*s e k*	*sekúnda*
sr	*s r*	*séra*

Table 12 Abbreviations that work in all contexts in SM

Note: For **cm**, **km**, and **mm** the spoken forms ending in **-i** are used, in SM only, after the digit **1**; in all other cases the forms in **-ar** are used.

8.2 Abbreviations that function in SM only

The abbreviations in Table 13 are read as abbreviations in SM only. In order to be read as abbreviations these must be entered without a full stop between the letters.

Abbreviation	LM	SM
h/f	*h skástrik f*	*há eff*
m a	*m a*	*meðal annars*
þ e	*þ e*	*það er*
þ e a s	*þ e a s*	*það er að segja*
e h	*e h*	*eftir hádegi*
f h	*f h*	*fyrir hádegi*
o s frv	*o s f r v*	*og svo framvegis.*
e-r	*e strik r*	*einhver*
e-n	*e strik n*	*einhvern*
e-m	*e strik m*	*einhverjum*
e-ð	*e strik ð*	*eitthvað*
e-u	*e strik u*	*einhverju*
e-s	*e strik s*	*einhvers*
u þ b	*u þ b*	*um það bil*
t d	*t d*	*til dæmis*

Table 13 Abbreviations that work in all contexts in sentence mode only

The abbreviations of area in Table 14 are also read correctly only in SM:

Abbreviation	LM	SM
m²	*m í öðru veldi*	*fermetri/fermetrar*
cm²	*c m í öðru veldi*	*fersentimetri/ fersentimetrar*
km²	*k m í öðru veldi*	*ferkílómetri/ ferkílómetrar*
mm²	*m m í öðru veldi*	*fermillimetri/ fermillimetrar*

Table 14 Abbreviations of area that work in all contexts in sentence mode only

Note that the forms ending in **-i** are used, in SM only, after the digit **1**; in all other cases the forms in **-ar** are used.

To have the letter **M** on its own read as *metri* after the digit **1** and *metrar* in other cases it should be entered in the user lexicon in the following form:

Look-up word	Transcription
M	#@METR# or /_@ m e t r/

When the user lexicon is case-sensitive the above entry would only affect a capital **M**, whereas a small **m** would still be read as the name of the letter. Having a small **m** as the look-up word in the case-sensitive mode with the above transcription would cause this to be read as *metri*/*metrar*.