Language Manual

HQ and HD Czech

Language Manual
Czech
Sabrine and Eliska
15 March 2011

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
Immeuble Les Erables
Rue du Lac
BP 98153
31681 Labège Cedex
France

Tel: +33 (0)5 62 24 71 00
Fax: +33 (0)5 62 24 71 01

www.acapela-group.com

Table of Contents

List of Tables

Chapter 1 General

This document discusses certain aspects of text-to-speech processing for the Czech 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 Sabrine and the High Quality voice Eliska.

Please note that the *User's Guide*, mentioned several times in the manual, is called *Help* in some applications.

Note: This language manual is general and applies to all Acapela Group HQ Czech voices specified above. One or more of the voices may be included in a certain Acapela Group product.

Note: For efficiency reasons, the processing described in this document has a different behaviour in some Acapela Group products. Those products are:

- Acapela TTS for Windows Mobile
- Acapela TTS for Linux Embedded
- Acapela TTS for Symbian

For these products, the default processing of numbers, phone numbers, dates and times has been simplified for the low memory footprint (LF) voice formats. Developers have the possibility to change the default behaviour from simplified to normal preprocessing by setting corresponding parameters in the configuration file of the voice. Please see the documentation of these products for more information.
In the following chapters, each simplification will be described by the indication *[not SP]* following the description of the standard behaviour. The SP in the indication stands for *Simplified Processing*.

Chapter 2 Letters in orthographic text

Characters from **A-Z**, **a-z** (as well as **á**, **Á**, **ä**, **Ä**, **č**, **Č, ď, Ď**, **é, É, ě**, **Ě, í, Í, ň**, **Ň, ó**, **Ó, ö, Ö, ř**, **Ř,** š, **Š**, **ů, Ů, ú**, **Ú**, **ü**, **Ü**, **ť**, **Ť, ý, Ý, ž, Ž**) may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, i.e. " ł, õ, ą , ç ". These letters are not pronounced as in their native languages though, they are pronounced as regular " l, o, a, c " when occurring in a word.

Characters outside of these ranges, i.e. numbers, punctuation characters and other non-alphanumeric characters are not considered as letters.

Chapter 3 Punctuation characters

Punctuation marks appearing in a text affect both rhythm and intonation of a sentence. The following punctuation characters are permitted in the normal input text string:
, : ; “ ” . ? ! ( ) [] {}

3.1 Comma, colon and semicolon

Comma < **,** >, colon < **:** > and semicolon < **;** > cause a brief pause to occur in a sentence, accompanied by a small rising intonation pattern just prior to the character.

3.2 Quotation marks

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

3.3 Full stop

A full stop < **.** > is a sentence terminal punctuation mark which causes a falling end-of-sentence intonation pattern and is accompanied by a somewhat longer pause. A full stop may also be used as a decimal marker in a number (see chapter 5) and in abbreviations (see chapter 8).

3.4 Question mark

A question mark <**?**> ends a sentence and causes question-intonation, first rising and then falling.

3.5 Exclamation mark

The exclamation mark <!> behaves in a similar manner to the full stop, causing a falling intonation pattern followed by a pause.

3.6 Parentheses

Parentheses ' **( )** ', brackets ' **[ ]** ' and braces ' **{ }** ' appearing around a single word or a group of words cause a brief pause before and after the bracketed text.

Chapter 4 Other non-alphanumeric characters

4.1 Non-punctuation characters

The characters listed in Table 4-1 are processed as non-letter, non-punctuation characters. Some are pronounced at all times and others are only pronounced in certain contexts, which are described in the following sections of this chapter.

Symbol	Reading
/	lomeno
+	plus
$	dolar
€	euro
<	menši než
>	větši než
%	procento
^	vokaň
|	vertikalni zavorka
~	Tilda
@	Zavinač
=	rovna se
²	(see below)
³	(see below)
-	(see below)
*	(see below)

Table 4-1 Non-punctuation characters

4.3 Symbols with pronunciation variable depending on the context

4.3.1 The ² and ³ signs

The expressions with ² and ³ are read as accordingly "čtvereční" or " kubický ". The accompanying unit measure words: such as mm, cm, m, km or l, are expanded in Nominative Case in number agreement with the preceding number.

Expression	Reading
1 mm²	jeden čtvereční milimetr
2 cm²	dva čtvereční centimetri
5 m²	pět čtvereční metri
80 km²	osumdesát čtvereční kilometri
1 mm³	kubický milimetr
1 cm³	kubický centimetr
1 m³	kubický metr
1 km³	kubický kilometr

4.3.2 Hyphen

A hyphen < - > is pronounced “mínus” in two cases:

a) if followed by a digit and no other digit is found in front of the hyphen

b) if followed by a digit and an equals sign. If there is no equals sign, it is pronounced “pomlčka”.

In certain date formats, in between days or years, the hyphen is pronounced “až”.
In compounds or between words, the hyphen is not pronounced. Example: T-Mobile.

Expression	Reading	
-3	mínus tři	
44.3	čtyřicet čtyři pomlčka tři	
44-3=41	čtyřicet čtyři mínus tři rovná se čtyřicet jedna	
15-20 října	patnácteho až dvacáteho října	[not SP]
6-10 listopadu	šestého až desáteho listopadu	[not SP]
1998-2004	tisíc devjetset devadesát osum až dva tisíce čtyři	[not SP]
02-02-2002	druhého února dva tisíce dvě	[not SP]

4.3.3 Asterisk

Asterisk < * > is only pronounced as “krát” if enclosed by digits and followed by equals sign. In other cases it is pronounced as “hvězdička”.

Expression	Reading
2*3	dva hvězdička tři
2*3=6	dva krát tři rovná se šest
*bc	hvězdička b c

Chapter 5 Number processing

Strings of digits that are sent to the text-to-speech converter are processed in several different ways, depending on the format of the string of digits and the immediately surrounding punctuation or non-numeric characters. To familiarise the user with the various types of formatted and non-formatted strings of digits that are recognised by the system, a brief description of the basic number processing is provided below, along with examples.

Number processing is subdivided into the following categories:

Full number pronunciation
Leading zero
Decimal numbers
Currency amounts
Ordinal numbers
Arithmetic operators
Mixed digits and letters
Time of day
Year
Dates
Phone numbers

5.1 Full number pronunciation

Full number pronunciation is given for the whole number part of the digit string.

Examples:

2425	full number
2.425	full number
2 425	full number
24,25	24 is a full number, 25 is the decimal part

Numbers denoting thousands, millions and billions (numbers larger than 999) may be grouped using space or full stop. In order to achieve the right pronunciation the grouping must be done correctly.

The rules for grouping of numbers are the following:

- Numbers are grouped in groups of three starting at the end.
- The first group in a number may consist of one, two, or three digits.
- If a group, other than the first, does not contain exactly three digits, the sequence of digits is not interpreted as a full number.
- The highest number read is 999999999999 (twelve digits). Numbers higher than this are read as separate digits.

Number	Reading
2580	dva tisíce pět set osumdesát
2 580	“
2.580	“
25800	dvacet pět tisíc osum set
25 800	“
25.800	“
2580350	dva milióny pět set osumdesát tisíc tři sta padesát
2 580 350	“
2.580.350	“
1000000000	jedna miliarda
1234567890123	jedna dva tři čtyři pět šest sedum osum devět nula jedna dva tři
23 456 789 012	dvacet tři miliarda čtyři sta padesát šest milióny sedum set osumdesát devět tisíc dvanáct

5.2 Leading zero

Numbers that begin with 0 (zero) are read as a whole number, with a leading zero read out separately.

Number	Reading
09253	nula devět tisíc dvě stě padesát tři
020	nula dvacet

5.3 Decimal numbers

Comma or full stop may be used when writing decimal numbers.

The full number part of the decimal number (the part before comma or full stop) is read according to the rules in 5.1. The decimals (the part after comma or full stop) are read as separate digits if there are more than 3 digits after the comma. When the decimals are read as a whole number, the words "desetin", "setin", and "tisícin" are added after the decimal, depending on the number of digits in the decimal. Note: A number containing a period followed by exactly three digits is not read as a decimal number but as a full number, following the rules in 5.1.

Number	Reading
16,234	šestnáct celých dvě stě třicet čtyři
3,1415	tři celé jedna čtyři jedna pět
1251,04	tisíc dvě stě padesát jedna celých nula čtyři
1.251,04	tisíc dvě stě padesát jedna celých nula čtyři
2.50	dva tečka padesát
2,50	dva celé padesát
3.141	tři tisíce sto čtyřicet jedna

5.4 Monetary amounts

The following principles are followed for monetary amounts:

- Numbers with zero or two decimal places preceded or followed by the currency markers **$**, **€**, **CZK**, **SKK**, **HUF**, **PLN**, **GBP**, **CAD**, **DKK**, **EUR**, **JPY**, **USD** are read as monetary amounts (see Table 8-4 for the expansion of the 3-letter abbreviations)
- Numbers with zero or two decimal places preceded or followed by the words "dolar", "euro", "libra", "lib.", "yen", "česká koruna", "slovenská koruna", "forint" or "zlotý" (singular or plural) are read as monetary amounts.
- Accepted decimal markers are comma and full stop.
- The decimal part (consisting of two digits) in monetary amounts is read as "i nn cent".
- If the decimal part is "00" it will not be read.

Expression	Reading
$15.00	patnáct dolarů
15.00 CZK	patnáct českých korun
15.00 euro	patnáct euro
€ 200.50	dvě stě euro i padesát centů
1 000 000 $	jeden milión dolarů

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

$ 1 milión	jeden milión dolarů	[not SP]

5.5 Ordinal numbers

Numbers are read as ordinals in the following cases:

- The number is followed by a full stop and a space, with no capital letter after the space. Examples: 1. , 12.
- The number is followed by a period, then a month name or one of the month name abbreviations and the number is smaller or equal to 31. The number may be preceded by a day or an abbreviation for a day. Examples: 3. Ledna, 3. Led., Sobota 3. Led.

5.6 Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below.

Expression	Reading
-12	minus dvanáct
+12	plus dvanáct
2*8	dva hvězdička osm
2*8=16	dva krát osm rovnáse šestnáct
2/8	dvě osmini
8/2=4	osm děleno dva rovná se čtyři
25%	dvacet pět procento

5.7 Mixed digits and letters

If a letter appears within a sequence of digits, the groups of digits will be read as numbers according to the rules above. The letter marks the boundary between the numbers. The letter will also be read. Examples:

Expression	Reading
77BB84	77 B B 84
0092BC87	0 0 92 B C 87

5.8 Time of day

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

Possible patterns are:

a) **hh:mm (or h:mm)**

b) **hh:mm:ss (or h:mm:ss)**

c) **hhHmm (or hHmm)** ex: 2H45 [not SP]

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

In pattern a): If the “mm”-part is equal to “00”, this part will not be read. “Hodin(i,a)”, is read after the hours, whether there be minutes or not. “Minut” or "minuti" is added after the “mm” part in agreement with the preceding number.

In pattern b): “Sekunt” will be added after the seconds part. If the “ss”-part is equal to “00”, this part will not be read.

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

Expression	Reading
8:15	osum hodin patnáct minut
9:00	devět hodin
1:02	jedna hodina dvě minuti
5:20:45	pět hodin dvacet minut čtiřiset pět sekunt
5:20:00	pět hodin dvacet minut

5.9 Dates

The valid formats for dates are:

1.dd.mm.yyyy, dd-mm-yyy and dd/mm/yyyy
2.dd.mm.yy, dd-mm-yy and dd/mm/yy

where, "**yyyy**" is a four-digit number, "**yy**" is a two-digit number, "**mm**" is a month number between 1 and 12 and "**dd**" a day number between 1 and 31. Full stop, hyphen and slash may be used as delimiters.

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

The following examples show valid formats and their readings:

Type 1 Examples	Reading
10.02.2003 or 10.2.2003	desátého února dva tisíce tři
10-02-2003 or 10-2-2003	"
10/02/2003 or 10/2/2003	"

Type 2 Examples	Reading
10.02.03 or 10.2.03	desátého února dva tisíce tři
10-02-03 or 10-2-03	"
10/02/03 or 10/2/03	"

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

Expression	Reading
1980-1990	tisíc devět set osmudesát až tisíc devět set devadesát
1939-45	tisíc devět set třicet devět až čtyřicet pět
14-15 Duben	čtrnáct až patnáct Duben
14 až 15 Duben	čtrnáct až patnáct Duben
6. do 10. února	šestého do desátého února
6-10.1	šest až deset Ledna

Other possible formats include: [not SP]

Sobota, 12. Srpen
Sobota, 12. Srpen 2000
30. Leden 1980
Duben 1999

Valid abbreviations for months are: led., ún., bř., břez., dub., květ., červ., červen., srp., zář., říj., list., pros. [not SP]

Valid abbreviations for days are: pon., ut., stř., čtv., pát., sob., ned. [not SP]

5.10 Phone numbers

In this section the patterns of digits that are recognized as phone numbers are described. In the pronunciation of phone numbers, groups of two and three digits are read as normal numbers, groups of four digits are read out digit by digit, with a pause between the groups. [not SP]

5.10.1 Ordinary phone numbers

Sequences of digits in the following formats are treated as phone numbers. [not SP]

The following sequences of digits can be separated by a space, a hyphen or a full stop. The first separation can be a space, a hyphen, or a backslash:

x xxxx xxxx
x xx xx xx xx
xx xxx xxxx
xx xxxx xxx
xxx xx xx xx
xxx xxxxxx

The sequences can also all be preceded by a "0".

The sequence xxx xxx xxx is recognized as a phone format only if preceded by "tel, mob, telefon, mobil, fax", with optional ":" between these words and the phone number.

5.10.2 International phone numbers

All preceding formats can be recognised if preceded by international prefix (space is optional) [not SP]:

00 x	+ x
00 xx	+ xx
00 xxx	+ xxx

These international prefixes can be followed by an optional (0) before the phone number.

Other recognized formats, that must be preceded by "00" or "+" with optional separator :

xx xxxx xxxx
xxx xxx xxxx
xxx xxxx xxx
xxxx xx xx xx
xxx xxxx xxxx
xxxx xxx xxxx
xxxx xxxx xxx
xxxxx xx xx xx
xxxx xxxx xxxx
xxxxx xxx xxxx
xxxxx xxxx xxx
xxxxxx xx xx xx
xxx xxx xxx xxx

Chapter 6 How to change the pronunciation

Words that are not pronounced correctly by the text-to-speech converter can be entered in the user lexicon (see User's guide). In this lexicon, the user enters a phonetic transcription of the word (see chapter 7). Phonetic translations can also be entered directly in the text, using a PRN-tag (see User's guide)

Chapter 7 Czech phonetic text

The Czech text-to-speech system uses symbols based on the SAMPA phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*). The symbols are written with a space between each phoneme.

Only SAMPA may be used in phonetic transcriptions. Symbols not listed here are not valid in phonetic transcriptions and will be ignored if included in the user lexicon or in a PRN tag.

7.1 Consonants

7.1.1 Symbols for the Czech consonants

Sym	Word	Phonetic text	Comment
b	Bota	b o1 t a	
	Dal	d a1 l	
d'	Di	d’ i1	
dz	Leckdy	l e1 dz g d l	
dZ	Lečbych	l e1 dZ b i x	
f	Forma	f o1 r m a	
g	Kde	g d e	
H	Had	H a1 t	
j	Jas	j a1 s	
k	Krk	k r=1 k	
l	Led	l e1 t	
m	Mák	m a:1 k	
n	Noc	n o1 ts	
n'	Nic	n’ i1 ts	
N	Banka	b a1 N k a	
p	Pes	p e1 s	
r	Ret	r e1 t	
r'	Tuř	t u1 r’	
s	Sen	s e1 n	
S	Šabach	S a1 b a x	
t	Tam	t a1 m	
t'	Tati	t a1 t’ i	
tS	Tavič	t a1 v i tS	
ts	Cíl	ts i:1 l	
v	Vak	v a1 k	
x	Chata	x a1 t a	
z	Zub	z u1 p	
Z	Žal	Z a1 l	
R	Buřeň	b u1 R e n'	
?	Glottal stop	? o1 s u m=	

Table 7-2 Czech consonants

7.2 Vowels

7.2.1 Symbols for the Czech vowels

Sy	Word	Phonetic text	Comment
a	Pas	p a1 s	Short
a:	Rád	r a:1 t	Long
e	Les	l e1 s	Short
e:	Lék	l e:1 k	Long
i	Myš	m i1 S	Short
i:	Píbl	p i:1 b l=	Long
o	Rok	r o1 k	Short
o:	Móda	m o:1 d a	Long
u	Kus	k u1 s	Short
u:	Kučův	k u1 tS u: f	Long
ou	Bambous	b a1 m b ou s	Short
au	Braum	b r au1 m	Short
eu	Breu	b r eu1	Short
l=	Bágl	b a:1 g l=	Syllabic l
m=	Cozm	ts o1 z m=	Syllabic m
r=	Crha	ts r=1 H a	Syllabic r
E:	äke	E:1 k e	Foreign vowel
2:	Mälmö	m E:1 l m 2:	Foreign vowel
y:	Blüml	b l y:1 m l=	Foreign vowel
ou:	Kouďův	k ou:1 d' u: f	Long
eu:	Kubeův	k u1 b eu: f	Long

Table 7-3 Czech vowels

7.3 Lexical accent

A lexical accent is used to indicate the level of prominence (or emphasis) of a syllable in a word. Practically all words in Czech have a lexical accent even if it does not always serve to differentiate between two different words. It is therefore important to include stress marks when writing phonetic transcriptions.

In the phonetic transcriptions, the lexical accent is indicated by the symbol "1" placed directly after (no space) the accented vowel.

7.4 Pause

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

7.5 Glottal stops

A glottal stop, represented by the phonetic symbol /?/, is a small sound which is often used to separate two words when the second word starts with a vowel. This sound can be inserted in a transcription in order to improve the pronunciation.

Chapter 8 Abbreviations

In the current version of the Czech text-to-speech system, the abbreviations in table 4 below are recognized in all contexts. These abbreviations are mostly case-insensitive, and some require punctuation such as comma or slash.

As previously mentioned, there are also abbreviations for the days of the week and the months, see chapter 5.9.

Abbreviation	Reading
ap	a podobně
atd.	a tak dále
cal	kalorie
ccm	kubický centimetrů
cm	centimetrů
hl.	hlasovi
hlas.	hlasové
jedn.	jednotka
kb	kilobajti
kg	kilogram
kilometry/h	kilometry za hodinu
km	kilometrů
krychl (.)	krichlové
kW	kilowattů
kWh	kilovat hodin
metry/s	metry za sekundu
mm	milimetrů
míle/h	míle za hodinu
přík (.)	příkas
spr.	spravce
stopy/s	stopy za sekundu
°C	stupeň Celsia
°F	stupeň Fahrenheita
°K	stupeň Kelvina
cad	kanacký dolar
czk	česká koruna
dkk	danská koruna
eur	euro
gbp	bricá libra
huf	maïarský forint
jpy	jen
pln	polskí zlotí
skk	slovenská koruna
usd	americký dolar

Table 8-4 Abbreviations

The single-letter abbreviations, such as “m”, “g”, “s” and “t” are expanded only when appearing after a number. Different expansions are applied in agreement with the number.

Expression	Reading
25 m	dvacet pět metri
30 s	třicet sekundi
45 g	čtyřicet pět grami

Chapter 9 Web-addresses and email

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

- "www" is read as three w's spelled letter by letter.
- Full stops are read as "tečka", hyphens as "pomlčka", underscore ("_") as "podtržník", slash ("/") as "lomeno".
- "cz, fr" and all the other abbreviations for countries are spelled out letter by letter.
- The "@" is read "zavináč".
- Words/strings (including "org", "com" and "edu") are pronounced according to the normal rules of pronunciation in the system and in accordance with the lexicon.

Expression	Reading
www.acapela-group.com	w w w tečka acapela pomlčka group tečka com
http://www.acapela-group.com	h t t p dvojtečka lomeno lomeno w w w tečka acapela pomlčka group tečka com
novak@yahoo.cz	novak zavináč yahoo tečka c z
milos_novak@yahoo.cz	milos dlouha pomlčka novak zavináč yahoo tečka c z