Description of the Phonological Corpus of Czech

[last updated 28/07/20]

1 Phonological Analysis

The Corpus consist primarily of lexemes (in the Lexical Sub-corpus) or sentences (in the Textual Sub-corpus) and their phonological representation (transcription). The transcription follows the phonological analysis of Modern Czech worked up in the book *Phonotactics of Czech* by Aleš Bičan (Peter Lang, 2013). The mentioned book contains more details and information.

A phonological description accounts for two types of properties of linguistic data: segmental phonetic properties and suprasegmental phonetic properties. The selected phonological theory (Axiomatic Functionalism, Mulder 1989) provides phonological models for accounting for these sets of these properties. Three phonological models are relevant here: *phoneme*, *phonotagm*, *para-phonotactic feature*.

- The **phoneme** has three equivalent definitions: a) an unordered set of distinctive features; b) a minimal phonotactic entity; c) a set of allophones. It accounts for segmental phonetic properties of speech sounds. See 1.1.
- The **phonotagm** is defined as a self-contained (autonomous, complete) group of phonemes. One of the phonemes is a nuclear, compulsory element; the others are peripheral or non-nuclear (and usually optional) elements within the group. Phonotagms account for segmental phonetic properties of syllables. See 1.2.
- The **para-phonotactic features** are additional phonological properties which either distinguish one phonological form from other another (e.g. tones in tone languages) or group phonotagms into higher-level units such as accent groups. They thus account for suprasegmental phonetic properties of utterances. See 1.3.

1.1 Phonemes of Czech

According to phonotactic function three classes of phonemes are recognized for Czech: **vowels**, **consonants** and **semiconsonants**. The vowels function as identity elements of phonotagms (i.e. are nuclei of phonotagms), while the consonants are dependent on vowels for their function and occurrence. The semiconsonants are phonemes which have both the nuclear and non-nuclear function.

1.1.1 Vowels

The vowels are the phonemes which are always nuclear; the classification of the Czech vowels according to distinctive features together with the range of their allophones is given in table 1. The diphthongs [ɛu̯] (*euro*), [au̯] (*auto*) and [ou̯] (*louka*) are interpreted as single phonemes (see Bičan 2013: 37–40).

	Front		Central	Back	
	high	mid		high	mid
Short	/i/ [ɪ]	/e/ [ɛ]	/a/ [a]	/u/ [u]	/o/ [o]
Long	/ī/ [iː]	/ē/ [ɛː]	/ā/ [aː]	/ū/ [uː]	/ō/ [oː]
Diphthongal	/ë/ [ɛu̯]		/ä/ [au̯]	/ö/ [ou̯]	

Table 1: Vowels of Czech with their allophones

1.1.2 Consonants

The consonants are the phonemes which are always peripheral (non-nuclear); the classification of the Czech consonants according to distinctive features together with the range of their allophones is given in table 2.

	Labial		Alveolar		Palatal		Velar	
	v'less	v'ed	v'less	v'ed	v'less	v'ed	v'less	v'ed
Occlusive	/p/ [p]	/b/ [b]	/t/ [t] [c]	/d/ [d] [ɟ]	/ť/ [c]	/ď/ [ɟ]	/k/ [k]	/g/ [g]
Fricative	/f/ [f]	/v/ [v]	/s/ [s]	/z/ [z]	/š/ [ʃ]	/ž/ [ʒ]	/x/ [x]	/h/ [ɦ]
Nasal	/m/ [m] [ɱ]		/n/ [n] [ɲ] [ŋ]		/ň/ [ɲ]			
Outside the proportional system: /j/ [j] and /ř/ [r̝] [r̝̊]								

Table 2: Consonants of Czech in the context of relevance and their allophones

Affricates

The affricates [t͜s], [d͜z], [t͜ʃ] and [d͜ʒ] are interpreted as combinations of two phonemes (see Bičan 2013: 33–7 for the defense of this analysis). Their interpretation is shown in table 3. They are subject to neutralization (see 1.1.4)

Affricate	Phonological interpretation			
	In contexts of relevance	Example	In contexts of neutralization	Example
[t͜s]	/Ts/	/TseSta/ *cesta*	/TS/	/peTS/ *pec* /klaTSki/ *klacky*
[d͜z]	/Tz/	/Tzinkaťi/ *dzinkati*	/TS/	/leTSKgo/ *leckdo*
[t͜ʃ]	/Tš/	/TšaS/ *čas*	/TŠ/	/mīTŠ/ *míč* /poTŠti/ *počty*
[d͜ʒ]	/Tž/	/Tžem/ *džem*	/TŠ/	/lēTŠba/

Table 3: Interpretation of the affricates

The affricates are distinguished from the sequences of pre-alveolar stops (realized with an explosion) and pre-alveolar or post-alveolar fricatives [t.s], [d.z], [t.ʃ], [d.ʒ] (cf. *práce* × *prát se*, *počít* × *podšít*). The stop-fricative sequences are in this case interpreted as signals of phonological boundaries because this pronunciation is invariably found across grammatical boundaries only (see 1.3.2).

1.1.3 Semiconsonants

The semiconsonants are the phonemes which can be both nuclear entities (like the vowels) and peripheral entities (like the consonants). Two such phonemes are recognized for Czech: the vibrant /r/ realized as syllabic [r̩] and non-syllabic [r], and the lateral /l/ realized as syllabic [l̩] and non-syllabic [l].

Notation of nuclear /r/ and /l/

For the sake of convenience, the nuclear semiconsonants are transcribed /R/, /L/ in the Corpus (e.g. /pLnī/ *plný*).

Syllabic nasals

The syllabic bilabial nasal [m̩] may be pronounced in *sedm*, *osm* and similar words, but since it is always freely substitutable for the sequence [um] (i.e. [sɛdum] instead of [sɛdm̩]), these words are analyzed as containing the sequence /um/.

1.1.4 Neutralization and archi-phonemes

Neutralization is the contextual irrelevancy of a difference between two or more phonemes which is relevant in other contexts (= contexts of relevance). The phonemes occurring in contexts of neutralization are archi-phonemes defined as the intersection of the distinctive features of two or more phonemes (i.e. the phoneme between which the neutralization takes place). Neutralization accounts for the fact that a phonological difference which distinguishes the meaning under some conditions does not distinguish it under others.

Two neutralizations are recognized for Czech: **neutralization of voicing**, and **neutralization of the place of articulation for nasals**.

Neutralization of voicing

The neutralization of voicing is the contextual irrelevancy of the difference between the voiceless and voiced consonants. It results in **voicing archi-phonemes**; their classification together with the range of realization is given in table 4.

Neutralization of voicing takes place:

1) At the end of phonological words: /peS/ *pes*, /beS-ūTšelnī/ *bezúčelný*;
2) Before a voiceless or voiced consonant (with the exception of /v/): /StāT/ *stát*, /Sdar/ *zdar* (cf. /svāT/ *svát* × /zvāT/ *zvát*);
3) Before a voicing archi-phoneme: /FSpřīmiT/ *vzpřímit*, /xePSkī/ *chebský*;
4) Before /ř/ if the latter is followed by a voiceless or voiced consonant (with the exception of /v/): /Třťina/ *třtina*, /XřbeT/ *hřbet* (cf. /břve/ *Břve*);
5) Before /ř/ if the latter occurs at the end of a phonological word: /pePř/ *pepř*, /dovniTř/ *dovnitř*.

	Labial	**Alveolar**	**Palatal**	**Velar**
Occlusive	/P/ [p] [b]	/T/ [t] [d]	/Ť/ [c] [ɟ]	/K/ [k] [g]
Fricative	/F/ [f] [v]	/S/ [s] [z]	/Š/ [ʃ] [ʒ]	/X/ [x] [ɦ]
Nasal	/m/ [m] [ɱ]	/n/ [n] [ɲ] [ŋ]	/ň/ [ɲ]	
Outside the proportional system: /j/ [j] and /ř/ [r̝], [r̝̊]				

Table 4: Consonants of Czech in the context of neutralization of voicing with their allophones

The voicing archi-phonemes are phonologically neither voiceless nor voiced. Their realizational voicing is completely predictable from the context they occur in and is thus non-phonological. Hence they are transcribed with a special letter.

Neutralization of the place of articulation of nasals
The neutralization of the place of articulation of nasals is the contextual irrelevancy of the difference between /m/, /n/ and /ň/. It results in the **nasal archi-phoneme /M/** always realized as [m]. The neutralization is a consequence of the fact that no nasal other than [m] is found in certain contexts; the nasal's place of articulation is thus completely predictable there.

Neutralization of the place of articulation of nasals takes place phonotagm-initially before any consonant or a semiconsonant: /Mdlo/ *mdlo*, /MhöřiT/ *mhouřit*; /sMlöva/ *smlouva*, /zMRznöT/ *zmrznout*.

Note: The analysis of the Corpus has shown that no other context for this neutralization cannot be consistently postulated.

1.2 Phonotagms

The phonotagms are self-contained combinations of phonemes. In Czech a phonotagm is almost always realized as a single syllable, but it can be a more complex unit in other languages. The syllable is understood here as a *phonetic* entity definable by phonetic theory, while the phonotagm is a *phonological* entity; it could also be called a phonological syllable.

In Czech there may be a discrepancy between phonotagms and syllables with words like *stárl*, *Karl* or *umrlčí*. Some users of Czech perceive the phoneme /l/ as syllabic in these words. From the phonotactic perspective, however, the phoneme is not nuclear because it is dependent for its occurrence of the preceding vowel. Hence, the mentioned words are phonologically transcribed as /Stārl/, /karl/, /umRlTšī/ (Bičan 2013: 140ff.), i.e. the /l/ is not nuclear here.

1.2.1 "Syllabification"

Phonological words containing more phonotagms are divided into constituent phonotagms ("syllabified") according to the rules given in table 5. The rationale behind these rules is explained in Bičan (2017).

Phoneme sequence	**Division rule**		**Example**	**Comment** (R = sonant, O = obstruent, i.e. occlusive or fricative, A = affricate group, v = /v/)
VV	0	V.V	/du.āl/	the only possible division
VCV	1	V.CV	/te.le.vi.ze/	the only possible division
VCCV	2-1	VC.CV	/mīS.to/ /mal.ta/	default division except for the combinations of types OR, Ov
			/loď.mi/ /ob.vaS/	exceptions to rule 2-2 (/ďm/ and /bv/ are not allowed initial combinations)
	2-2	V.CCV	/pā.dlo/ /ko.tva/	division for types OR, Ov (due to neutralization of voicing) syllabication for the “affricates” A)
	Division VCC.V is not used anywhere; all combinations can be divided with the previous rules.			
VCCCV	3-1	VC.CCV	/koS.tra/ /pol.Sko/	default division except for types RRO, ORO, RRR, ORR, OvR; division for types OA, RA
			/ob.Mňena/	exceptions to rules 3-2, 3-3 (/bmň/ is not an allowed initial combination)
	3-2	V.CCCV	/ro.zvrā.ťiT/ /vi.svlē.kaT/	division for types ORR and OvR (cf. rule 2-2); division for types Av, AR

	3-3	VCC.CV	/Štern.ber.Skī/
			/do.vňiTř.ku/
				division for types RRO and ORO;
				division for type AO;
				exceptions to type AR

			/drŠŤ.ka/	exception to rule 3-1 (/Ťk/ is not an allowed initial combination)
	Division VCCC.V is not used anywhere; all combinations can be divided with the previous rules.			
VCCCCV	4-1	VCC.CCV	/o.sa.zenS.tvo/	default division except for types RORR, OORO, ORRR; division for types AA, AOO, AOR; exceptions to type OOA
	4-2	VC.CCCV	/dvoj.sMňer.nī / /	division for types RORR, OORO (cf. 2-2); division for types OAO, OAR, OOA, RAO, RAR
			/boŠ.Stvī/	exceptions to rule 4-1 (/ŠS/ is not an allowed final combination)
	4-3	V.CCCCV	/za.hřMňe.lo/	division for type ORRR; division for type ARR
	4-4	VCCC.CV		exceptions to types RAO, RAR
	Division VCCCC.V is not possible because phonotagms do not end in four peripheral phonemes.			

VCCCCCV	5-1	VCC.CCCV	/nerF.Stvo/	default division;
				division for types AOOO, OAOO, AOOR;
				exceptions to type RAOO

	5-2	VC.CCCCV	/přeT.FStřiK/	exceptions to rule 5-1 (/TF/ is not an allowed final combination)
	5-3	VCCC.CCV		division for types RAA, RAOR, ROAO, ROAR, RRAO, RRAR, RAOO; exceptions for type AOOR
	Division V.CCCCCV is not used anywhere. Divisions VCCCC.CV, VCCCCC.V are not possible (see above).			
VCCCCCCV	6	VCC.CCCCV	/adjunK.TStvī/	only for combinations with “affricates”
Combinations with more than six peripheral phonemes are not attested.				

Table 5: Division rules (“syllabification rules”)

1.3 Para-phonotactic features

In Czech para-phonotactic features determine the groupment of phonotagms, i.e. they gather phonotagms into higher-level unit. Three types of groupment are recognized for Czech: **phonological word**, **accent group** and **tone group**. In addition, orthographic word is recognized in the Corpus.

The phonological transcription makes use of several boundary-signaling symbols to indicate the extension of some unit. See table 6 for the notation used.

Symbol	**Marks the boundary of**	**Corresponds grammatically roughly to the boundary of**	**Example**
#	tone group	clause	/köpil_si+kňihu#Kterö_si+Fždi+přāl/ *Koupil si knihu, kterou si vždy přál.*
+	accent group	word or stem in compound words	/hoďʼiT+sebö/ *hodit sebou* /TšeSko+slovenSkī/ *česko-slovenský*
=	phonological word	word	/nuďʼiT=se/ *nudit se*
-	phonological word	morpheme	/poT-ūředňīK/ *podúředník*
_	orthographical word	word	/pēTsi_se/ *péci se*
.	phonotagm	–	/po.lēF.ka/ *polévka*

Table 6: Boundary-signaling marks used in the Corpus

1.3.1 Accent group

The accent group is a group of phonotagms gathered together by features of accent. The group is marked by internal coherence and melodic contour (Palková 2013). The same sequence of phonotagms (syllables) may correspond to different accent groups the difference between which is determined by the melodic contour (e.g. *od ní mají* × *odnímají*: /od_ňī+majī/ × /odňīmajī/).

In the Phonological Corpus the concept of accent group is accordance with the theory of Zdena Palková, and the groupment of phonotagms into accent groups follows her rules formulated and applied to automatic speech synthesis of Czech (Palková 2004).

1.3.2 Phonological word

The phonological word is recognized as a constituent of accent groups in the case of the occurrence of the features marking their internal organization. In general, those phonetic and phonological features which can or must be understood as boundary signals are interpreted as interpreted as the signals of phonological word boundaries. For more details see Bičan (2014).

Glottal stop

Since the glottal stop and its equivalents occur only before vocoids at what can be recognized as morphological boundaries (prefix–stem boundaries, composite boundaries, word boundaries), they are taken as signals of phonological words. Accordingly, the following accent groups are analyzed as two phonological words (see above for the notation):

[potʔokɛm] *pod okem*	→ /poT=okem/
[fʔaktɛx] *v aktech*	→ /F=aKteX/
[vɛlkoʔopxot] *velkoobchod*	→ /velko-oPxoT/
[bɛsʔokiː] *bezoký*	→ /beS-okī/

Alveolar stop – sibilant sequences

Since the sequences of an alveolar stop and an alveolar or post-alveolar fricative (sibilant) are possible only at morphological boundaries, they are understood as the signals of phonological words. Hence:

[pot.ʃiːt] *podšít*	→ /poT-šīT/
[praːt.sɛ] *prát se*	→ /prāT=se/

Non-syllabic realization of /r/ and /l/

Since /r/ and /l/ are realized as syllabic between two non-nuclear phonemes, their non-syllabic realization is taken as a boundary signal, which coincides with a morphological boundary. Hence:

[vrtɛx] *v rtech*	→ /v=rteX/
[podlhuːtɲiː] *podlhůtní*	→ /pot-lhūtňī/

Neutralization of voicing

Since neutralization of voicing of the final occlusives and fricatives takes place before sonants and /v/ across boundaries of orthographic words, a phonological word boundary is postulated here:

[naːrot jɛ vɛlkiː] *národ je velký* → /nāroT=je+velkī/
[vlak vaːm ʔujɛl] *vlak vám ujel* → /vlaK=vām+ujel/

Nasals

Instead of [m], the nasal /m/ may be realized as [ɱ] before /f/ and /v/; there is free variation here (e.g. [tramvaj] and [traɱvaj] *tramvaj*). Since this realization is not orthoepic across boundaries of orthographic words, the absence of the free variation necessitates the postulation of a phonological word boundary.

The nasal /n/ is always realized as [ŋ] before /k/, /g/ and /K/, but not across boundaries of orthographic words. The sequences [nk] and [ng] are therefore taken as a phonological word boundary.

Accordingly, the following examples are analyzed as two phonological words:

[tam vaːm nɛbudɛ] *tam vám nebude* → /tam=vām+nebude/
[jɛn kos spiːval] *jen kos zpíval* → /jen=koS+Spīval/

Non-palatal realization of /T/ and /n/

Before palatal occlusives and nasals, the palatal occlusives /t/, /d/, /T/ may be realized as [c] or [ɟ] instead of [t] or [d] and the palatal nasal /n/ as [ɲ] instead of [n]; there is free variation (e.g. [kotɲiːk] and [kocɲiːk] *kotník*). Since these realizations are not orthoepic across boundaries of orthographic words, the absence of the free variation necessitates the postulation of a phonological word boundary. Hence:

[nɛvjɛɟɛt ɲɪt͜s spraːvɲɛ] *nevedět nic správně* → /nevjeďeT=ňiTS+Sprāvňe/
[jan ciːm tr̩pjel] *Jan tím trpěl* → /jan=ťīm+tRpjel/

1.3.3 Tone group

In accordance with the approach by Palková (2004), the tone group is a defined as a sequence of accent groups between two punctuation marks, the second of which may also be the conjunction *a* “and”. Tone units are used in the Textual Sub-corpus only.

1.4 Allophonic transcription

Phonemes can be defined as classes of allophones, which show how they may be phonetically realized in some contexts. The Corpus therefore contain also allophonic transcription. Because it is convenient (for searching and for computer processing) to have a transcription where one symbol stands for one allophone, a special set of symbols are used. See table 7.

Phonological Corpus	**IPA**	**Used for example in**
a	a	*strofa*
e	ɛ	*triolet*
i	ɪ	*lyrika*
o	o	*trochej*
u	u	*metrum*
á	a:	*báseň*
é	ɛ:	*cézura*
í	i:	*rýmovník*
ó	o:	*óda*
ú	u:	*půlverš*
O	ou̯	*dvouverší*
A	au̯	*klauzule*
E	ɛu̯	*eufonie*
m	m	*madrigal*
M	ɱ	*amfibrach*
n	n	*rondel*
N	ŋ	*blankvers*
ň	ɲ	*píseň*
p	p	*poezie*
b	b	*balada*
t	t	*tercína*
d	d	*daktyl*
ť	c	*šestiverší*
ď	ɟ	*předěl*
k	k	*kasída*
g	g	*elegie*
?	ʔ	*ʔanapest*
c	t͜s	*siciliána*
Z	d͜z	*Špicberky*
č	t͜ʃ	*časomíra*
Ž	d͜ʒ	*lučba*
f	f	*ferekratej*
v	v	*versifikace*
s	s	*sonet*
z	z	*gazel*
š	ʃ	*verš*
ž	ʒ	*syžet*
x	x	*choliamb*
X	ɣ	*Bar-Kochba*
h	ɦ	*hexametr*
l	l	*limerik*
r	r	*rým*
ř	r̝	*řádek*

Ř	r̝̊	*přízvuk*
j	j	*jamb*
R	r̩	*pentametr*
L	l̩	*mlha*

Table 7: Allophonic transcription used in the Corpus

2 Lexical Sub-corpus

The Lexical Sub-corpus consists of the main database mostly containing appellative vocabulary and a number of additional lexical databases (see section 2.4).

2.1 Format

The Lexical Sub-corpus is available in comma-separated value format (csv). This format allows for storing tabular data in plain-text form. It can be opened and edited in applications designed for editing csv files or editors such as Microsoft Excel, Microsoft Access etc. The data are separated by the separator “;”, i.e. the semicolon. Once opened or imported, the data will be displayed as a table.

2.2 Heading

The first row of the table is the heading consisting of the six sections shown in table 8.

Orthographic form	Phonological representation	Allophonic transcription	Phonological properties	Parts of speech	Occurrence in dictionar-ies

Table 8: Sections of the heading

2.3 Columns

The content of the respective columns is described below. Standard editors such as MS Excel allow searching and filtering the data according to criteria specified for each column.

Ortho
This column lists lexical items in their standard orthographic form as recording in the dictionaries.

PhRep
This column provides the phonological representation of the lexical item, i.e. the phonological form (PhF) of the item. The transcription follows the analysis outlined in section 1. Initially, the transcription was gained by automatic conversion of the orthographic form. This is possi-

ble since Czech orthography to a large extent reflects the expected orthoepic pronunciation. The conversion was then manually checked and corrected by Aleš Bičan. Some errors might have remained, though. They are expected to be corrected with new updates for the corpus.

Syllab
This column provides the syllabification of the lexical item. Syllabification rules are stored in external files and can be freely replaced with alternative rules. See section 1.2.1 for the rules.

Alloph
This column provides the allophonic transcription of the entry. See section 1.4 for the symbols used. Orthoepic pronunciation is assumed.

Length
This column contains numbers corresponding to the length of a PhF. The length equals to the number of phonemes within the PhF. Boundary-signaling symbols are not counted.

Phtagms
This column contains numbers corresponding to the number of phonotagms within a PhF. It equals the number of nuclear phonemes within the PhF.

CVStr
This column reproduces the structure of a PhF according to the membership of the constituent phonemes into the class of non-nuclear entities and nuclear entities. The symbols used are given in table 9. The boundary-signaling symbols are not included in the transcription.

Symbol	Stands for
C	consonants and non-nuclear semiconsonants
V	vowels
W	nuclear semiconsonants

Table 9: Symbols used in column CVStr

Place
This column reproduces the structure of a PhF according to the place of articulation of the constituent consonants (see table 2). The symbols used are given in table 10. The category “place” is not relevant for vowels and nuclear semiconsonants, and they are thus transcribed with minuscule letters.

Symbol	Stands for
L	labials
A	alveolars
P	palatals
K	velars
I	isolated consonants and semiconsonants (/M/, /j/, /ř/, /r/, /l/)
v	vowels
w	nuclear semiconsonants

Table 10: Symbols used in column Place

Manner

This column reproduces the structure of a PhF according to the manner of articulation of the constituent consonants (see table 2). The symbols used are given in table 11. The category “manner” is not relevant for vowels and nuclear semiconsonants, and they are thus transcribed with minuscule letters.

Symbol	Stands for
O	occlusives
F	fricatives
N	nasals
R	sonants (/j/, /ř/, /r/, /l/)
v	vowels
w	nuclear semiconsonants

Table 11: Symbols used in column Manner

Voicing

It reproduces the structure of a PhF according to the voicing of the constituent consonants (see table 2). The symbols used are given in table 12. The category “voicing” is not relevant for vowels and nuclear semiconsonants, and they are thus transcribed with minuscule letters.

Symbol	Stands for
U	voiceless
Z	voiced
X	indifferent (nasals and voicing archi-phonemes)
v	vowels
w	nuclear semiconsonants

Table 12: Symbols used in column Voicing

Horiz

This column reproduces the structure of a PhF according to the horizontal axis of the constituent vowels (see table 3). The symbols used are given in table 13. The category “horizontal axis” is not relevant for consonants and semiconsonants, and they are thus transcribed with minuscule letters.

Symbol	Stands for
Q	front
E	central
B	back
c	consonants and non-nuclear semiconsonants
w	nuclear semiconsonants

Table 13: Symbols used in column Horiz

Vertic

This column reproduces the structure of a PhF according to the vertical axis of the constituent vowels (see table 3). The symbols used are given in table 14. The category “vertical axis” is not relevant for consonants and semiconsonants, and they are thus transcribed with minuscule letters. The category is also not relevant for non-high and non-mid vowels; they are transcribed with a minuscule letter, too.

Symbol	Stands for
H	high
M	mid
v	non-high and non-mid vowels (/a/, /ā/, /ä/, /ĕ/, /ö/)
c	consonants and non-nuclear semiconsonants
w	nuclear semiconsonants

Table 14: Symbols used in column Vertic

Quant

This column reproduces the structure of a PhF according to the quantity (length) of the constituent vowels (see table 3). The symbols used are given in table 15. The category “quantity” is not relevant for consonants and semiconsonants, and they are thus transcribed with minuscule letters.

Symbol	Stands for
S	short
G	long
D	diphthongal
c	consonants and non-nuclear semiconsonants
w	nuclear semiconsonants

Table 15: Symbols used in column Quantity

PoS

This column provides information about the part of speech of a given entry. The digits corresponds to a particular part of speech; see table 16.

Symbol	Stands for
1	nouns
2	adjectives
3	pronouns
4	numerals
5	verbs
6	adverbs
7	prepositions
8	conjunctions
9	particles
0	interjections

Table 16: Symbols used in column PoS

Columns for dictionaries
The remaining columns specify whether a lexical item is recording in the dictionaries and databases of Czech. The zero (0) means it is not recorded; the digit 1 means that it is recorded. The list of the dictionaries included in the corpus is given in table 17.

Abbreviation	Stands for
SSČ	*Slovník spisovné češtiny* (4th edition, 2005)
SSJČ	*Slovník spisovného jazyka českého* (2nd edition, 1989)
PSJČ	*Příruční slovník jazyka českého* (1935–1957)
CSN	*Co v slovnících nenajdete (Novinky v současné slovní zásobě)* (1994)
SN	*Nová slova v češtině. Slovník neologizmů 1, 2* (1998, 2004)
SPr	*Slovesa pro praxi. Valenční slovník nejčastějších českých sloves* (1997)
SVaz	*Slovník slovesných, substantivních a adjektivních vazeb a spojení* (2005)
FSČ	*Frekvenční slovník češtiny* (2004)
ASCS	*Akademický slovník cizích slov A-Ž* (1995)
VSČ	*Výslovnost spisovné češtiny* (1978)

Table 17: Dictionaries included in the Lexical Sub-corpus

2.4 Additional lexical databases

The additional lexical databases are structured as the main Lexical Sub-corpus, but do not contain information about part of speech (PoS; all items are noun or nouns with an adjectival or prepositional attribute) and the occurrence in dictionaries. In contrast to the main Lexical Sub-corpus they contain column Type indicating the type of item.

2.4.1 Municipalities and their parts

This database consists of the names of the Czech municipalities and their parts. Column Type indicates whether the item is a name of a municipality (M) or of its part (P).

2.4.2 Given names and their hypocoristic forms

This database consists of a list of the most common Czech given names and their hypocorisms. Column Type indicates whether the item is a basic name (B) or a hypocorism (H). Column Sex indicated whether it is a male (M) or female (F) name.

2.4.3 Czech botanical names

This database consists of a list of Czech botanical names. Column Type contains only the value “botanical”.

2.4.4 Czech zoological names

This database consists of a list of Czech zoological names. Column Type provides information about the zoological class of a given item. FCC stands for Fungi, Corrals and Ctenophores.

3 Textual Sub-corpus

The Textual Sub-corpus consists of a selection of Czech texts phonologically transcribed according to their assumed orthoepic pronunciation. Preference was given to the novels not protected by copyright. The goals was to create a prose counterpart of the Corpus of Czech Verse (<http://www.versologie.cz/en/kcv.html>), which contains a phonetic transcription which can be phonologized and thus compared with the Textual Sub-corpus.

3.1 Format and Tags

The data of the Textual Sub-corpus are saved in xml format (Extensible Markup Language). It can be opened and edited in applications designed for editing these files (e.g. internet browsers). The data are enclosed within tags.

Tag ortho

This tag encloses sentences in standard orthographic form as they are written in the source text. The original texts were divided into sentences, so that each sentence is assigned a unique tag identified by an ID number. Paragraphs are ignored.

Arabic and Roman numbers were manually rewritten to words according to their meaning in the respective context (e.g. *10 hodin* to *deset hodin*, *Ludvík XIV.* to *Ludvík Čtrnáctý*).

Abbreviations were converted to their full form (e.g. *např.* to *například*). Acronyms were retained.

Sentences or phrases from foreign languages which were obvious instances of code-switching, not of loanwords were enclosed by square brackets (e.g. *["Sie Einjähriger,"] zařval pan obršt, "kdopak to vykřikl?"*). The text so marked is ignored in the phonological transcription (cf. /zařval_pan+obRŠT#KdopaK_to+vikřikL/). Ignored are also various editorial notes or chapter numberings, which were enclosed by square brackets as well.

Tag phrep

The phonological transcription of the sentences is enclosed within this tag. It follows the same principles as the transcription used for the Lexical Sub-corpus (see above). It also reflects the prosodic organization of the texts. A neutral organization was assumed, and it is based on the rules proposed for the automatic TTS synthesis of Czech (see Palková 2004). However, only prose text originally in the written forms were provided with the prosodic organization. Song lyrics and originally spoken texts were not subjected to the automatic TTS synthesis rules, as their prosodic organization may be quite different.

In the first phase, a list of unique word forms was gained from the texts. The words were automatically converted to phonological transcription in the same way as in the case of the Lexical Sub-corpus. The conversion was then manually corrected by Aleš Bičan, but some errors might have remained to be corrected in the future.

Tag syllab
This tag stores a copy of the phonological transcription where syllable boundaries are marked. See 1.2.1 for the syllabification rules.

Tag alloph
This tags encloses the allophonic transcription of the sentences. Orthoepic pronunciation is assumed.

4 References

Bičan, Aleš. 2013. *Phonotactics of Czech*. Peter Lang.

Bičan, Aleš. 2014. “K pojmu fonologické slovo v češtině”. *Sophia Slavica* (eds. Vít Boček – Bohumil Vykypěl), 13–23. Brno: Tribun EU.

Bičan, Aleš. 2017. “Slabikování”. *Nový encyklopedický slovník češtiny Online*. <http://www.czechency.org/slovnik/SLABIKOV%C3%81N%C3%8D>

Kučera, Henry – Monroe, George. 1968. *A Comparative Quantitative Phonology of Russian, Czech, and German*. Elsevier.

Mulder, Jan. 1989. *Foundations of Axiomatic Linguistics*. Berlin – New York: Mouton de Gruyter.

Palková, Zdena. 2004. “The set of phonetic rules as a basis for the prosodic component of an autonomous TTS synthesis in Czech”. *Phonetica Pragensia* X.33–46.

Palková, Zdena. 2013. “Prozodické vlastnosti češtiny ve vztahu k mezislovnímu sandhi”. *Studie k moderní mluvnici češtiny 5, K české fonetice a fonologii*, 106–118. Univerzita Palackého v Olomouci.