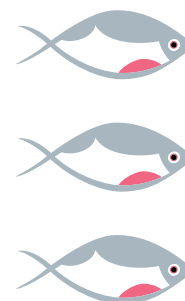


Table of Contents

Braxen 1.0 Documentation	1
Structure	1
Entries	2
Field structure	2
Field 0: Orthography	2
Field 1: Pronunciation	2
Stress	3
Boundaries	3
/e/ Sounds	4
Xenophones (foreign speech sounds)	4
Field 2: Part of speech and morphology	4
Field 3: Language code	5
Field 16: Case	5
Field 26: ID	5
References	5
Appendix A: Field information	5
Appendix B: Phoneme table - Base and IPA	6
Appendix C: Part of speech	8
Appendix D: Language codes	9



Braxen 1.0 Documentation

Braxen 1.0 is a Swedish pronunciation dictionary for speech technology, developed by Swedish Agency for Accessible Media (MTM). It is made public in cooperation with Språkbanken Tal.

Structure

The underlying data for Braxen 1.0 is a straight-forward two-dimensional table, with entries for full-form words, abbreviations, and acronyms, and columns (Fields) for various attributes of these entries.

Entries

An entry in Braxen 1.0 is defined by its full-form orthography, and several entries with the same orthography is permissible, but only if they differ in another field.

Field structure

An entry in Braxen 1.0 consists of 27 fields. In its raw, tab-separated form, each field occupies a column, and the order is meaningful, so that columns 0 to 26 have semantics.

In the current version, only a subset of these fields are shared publicly. The structure of the public dictionary remains the same, however, resulting in a number of empty columns (see [Appendix A: Field information](#)).

The most important fields are described in the following.

Field 0: Orthography

The orthography field shows the spelling of the word.

A considerable amount of pragmatics goes into the orthography field. For example, the most frequent casing is used, as different casings can sometimes indicate different pronunciations (see [Field 16: Case](#)).

Table 1. Examples

Orthography	Part of Speech	Transcription
björn	Noun	/b j 'oe: rn/
Björn	Proper name	/b j 'oe: rn/
BrB	Proper name (abbreviation)	Brottsbalken
BRB	Proper name (acronym)	B R B

Field 1: Pronunciation

The format for the phonetic-phonological transcriptions covers Swedish and common foreign phonemes. Here, this format is referred to as **Base**.

[Appendix B: Phoneme table - Base and IPA](#) provides a conversion table between Base and IPA. Tools for automatic conversion are included in the [/p5m/scripts](#) directory).

Table 2. Conversion Tools

Script	Task	Call
braxen_conversion.pl	Braxen to IPA(encode) or IPA to Braxen (decode)	perl p5m/scripts/braxen_conversion.pl <encode decode> <infile> <outfile>
validate_braxen.pl	Validates Braxen	perl p5m/scripts/validate_braxen.pl <infile> <outfile>

Script	Task	Call
validate_ipa.pl	Validates IPA	perl p5m/scripts/validate_ipa.pl <infile> <outfile>

Note that the conversion tools are written in Perl. Using e.g. Docker, they can be used without installing a Perl interpreter.

Example of conversion using Docker

Use "encode" for Braxen to IPA format and "decode" for IPA to Braxen format.

```
cd /braxen/repo/root
docker run -v "$PWD":/work -it perl -c "/work/p5m/scripts/braxen_conversion.pl
ENCODE|DECODE /work/INFILE /work/OUTFILE"
```

Example of validation using Docker

```
cd /braxen/repo/root
docker run -v "$PWD":/work -it perl -c "/work/p5m/scripts/validate_braxen.pl
/work/INFILE /work/
OUTFILE"
```

Stress

All Braxen 1.0 words have exactly one main stress, which can be either accent 1 or accent 2. Words with accent 2 also have secondary stress. Stress is marked immediately before the stressed vowel.

Table 3. Stress notation examples

Stress	Notation	Example	Transcription
Main stress accent 1	'	boll	/b 'o l/
Main stress accent 2	”	bollar, dalbana	/b ”o . l , a r / , /d ”a : l - b , a : . n a /
Secondary stress	,	bollar, dalbana	/b ”o . l , a r / , /d ”a : l - b , a : . n a /

Boundaries

Word boundaries allow multiple main stresses within an expression.

Table 4. Word boundary examples

Boundary	Notation	Example	Transcription
Word		berg- och dalbana	/b 'ae rj 'o : d ”a : l - b , a : . n a /
Compound	-	dalbana	/d ”a : l - b , a : . n a /
Morpheme*	~	transalpin	/t r a n s ~ a l . p 'i : n /
Syllable	.	alpin	/a l . p 'i : n /

*The morpheme boundary is optional but can be included when needed.

/e/ Sounds

Braxen 1.0 distinguishes between four /e/ sounds, mainly reflecting a central Swedish pronunciation.

Table 5. /e/ pronunciations

Phoneme	Description	Example
Transcription	e	Semi-open /e/ sound
sett (/s 'e t/)	ä	Semi-open /e/ sound (dialectal variation)
sätt	/s 'ä t/	eh
Before stressed syllable in unstressed, open syllable	betona	/b eh . t 'u: . n a/
ex	Schwa, used in unstressed syllables	bollen

Xenophones (foreign speech sounds)

The phoneme inventory includes foreign phonemes, so-called xenophones, primarily from English.

Field 2: Part of speech and morphology

Part of speech and morphological data largely follow the Stockholm-Umeå Corpus (SUC) principles (Gustafson-Capková & Hartmann, 2006). The **UO** (foreign word) tag is rarely used, as language codes indicate whether a word is Swedish or foreign.

Different parts of speech or morphological information for the same orthographic form can lead to different pronunciations:

Table 6. Part of speech and pronunciation examples

Orthography	Part of Speech	Transcription
slutet	NN	/s l 'uu: . t ex t/
slutet	JJ	/s l 'uu: . t ,ex t/
planet	NN UTR SIN IND NOM	/p l a . n 'e: t/
planet	NN NEU SIN DEF NOM	/p l 'a: . n ex t/

[Appendix C: Part of speech](#) provides a list of selected PoS codes.

Field 3: Language code

Language codes follow the ISO 639-2 standard (Library of Congress, 2017). The language code indicates the intended language of the orthography at the time of pronunciation creation.

Table 7. Examples of different language codes for the same word

Orthography	Language Code	Transcription
Anne	swe	/ˈa n/
Anne	eng	/ˈae n/

Appendix D: Language codes provides a list of selected language codes.

Field 16: Case

This field indicates case sensitivity:

- 1 = Case-sensitive
- 0 = Not case-sensitive

Field 26: ID

An internal identifier for each entry.

References

- Gustafson-Capková, S., & Hartmann, B. (2006). **Manual of the Stockholm Umeå Corpus version 2.0.**
- Library of Congress. (2017). **ISO 639-2 Language Code List.** https://www.loc.gov/standards/iso639-2/php/code_list.php

Appendix A: Field information

Bold fields are shared publicly.

Field	Name	Example
0	orth	bjärornas
1	pron	b j ˈae: . r , u . rn a s
2	posmorph	NN UTR PLU DEF GEN
3	lang	swe
4-15	internal	-
16	case	0
17-25	internal	-

Field	Name	Example
26	id	0060097

Appendix B: Phoneme table - Base and IPA

Base	IPA	Example
Base	IPA	Exempel
p	p	pil
b	b	bil
t	t	bot
rt	ʀ	bort
d	d	bod
rd	ʀ	bord
k	k	kal
g	ŋ	gal
f	f	fal
v	v	val
s	s	sal
rs	ʀ	fors
h	h	hal
x	χ	sjal
c	tʃ	tjat
m	m	matt
n	n	natt
rn	ʀ	barn
ng	ŋ	ring
r	r	riv
l	l	liv
rl	ʀ	sorl
j	j	jag
w	w	way (eng)
sh	ʃ	she (eng)
zh	ʒ	measure (eng)
z	z	zebra (eng)
dh	ð	this (eng)

Base	IPA	Example
th	θ	thick (eng)
rh	ɾ	road (eng)
r0	-	father (br. eng)
rx	ʀ	rouge (fre)
tc	tʃ	chicken (eng)
dj	dʒ	fudge (eng)
xx	x	Bach (ger)
i:	iː	sil
i	ɪ	sill
ih	ɪ̯	radio
y:	yː	syl
y	ʏ	syll
e:	eː	sel
e	e	sett
eh	e̯	betona
ex	ɐ	papper
ä:	æː	säl
ä	ɐ̯	sätt
ae:	æː	när
ae	æ	spärr
ö:	øː	rön
ö	ø̯	rönn
oe:	œː	för
oe	œ	förr
u:	uː	bot
u	ʊ	bott
oh	o	bohem
o:	oː	såt
o	ɔ̯	sått
uu:	uː	sur
uu	ʊ	surr
uuh	ʊ̯	butik
uw:	uː	you

Base	IPA	Example
uw	ʊ	would
a:	ɑː	mat
a	ʌ	matt
aa:	aː	Zlatan
au	aʊ	paus
eu	œ	euro
ei	eɪ	way (eng)
ai	aɪ	lie (eng)
oi	ɔɪ	boy (eng)
ou	oʊ	low (eng)
eex	eɪ	where (eng)
iex	ɪə	here (eng)
uex	uə	pure (eng)
an	ɑ̃	sans (fre)
en	ɑ̃	pain (fre)
on	õ	pardon (fre)
un	œ̃	lundi (fre)

Appendix C: Part of speech

Following (Gustafson-Capková & Hartmann, 2006).

Tag	Description	Example
AB	Adverb	inte
DT	Determiner	denna
HA	Interrogative/relative adverb	när
HD	Interrogative/relative determiner	vilken
HP	Interrogative/relative pronoun	som
HS	Interrogative/relative possessive pronoun	vars
IE	Infinitive	att
IN	Interjection	ja
JJ	Adjective	glad
KN	Conjunction	och
NN	Noun	pudding
PC	Participle	utsänd

Tag	Description	Example
PL	Verb particle	ut
PM	Proper noun	Mats
PN	Pronoun	hon
PP	Preposition	av
PS	Possessive pronoun	hennes
RG	Numeral	tre
RO	Ordinal	tredje
SN	Subjunction	att
UO	Foreign word	the
VB	Verb	kasta

Appendix D: Language codes

Code	Language
swe	Swedish
nob	Norwegian bokmål
nno	Norwegian nynorsk
dan	Danish
ice	Icelandic
fin	Finnish
eng	English
ger	German
fre	French
spa	Spanish
por	Portuguese
dut	Dutch
gre	Greek
rus	Russian
cze	Czech
chi	Chinese
jap	Japanese
kor	Korean
tha	Thai
swa	Swahili

Code	Language
ara	Arabic