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 permissable, 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, hwoever, 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 [_16_case]).

Table 1. Examples

Orthography	Part of Speech	Pronunciation
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 fomat 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
convertBase2IPA.pl	Converts Base to IPA
convertIPA2Base.pl	Converts IPA to Base
validateBase.pl	Validates Base transcription
validateIPA.pl	Validates IPA transcription

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

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 (/b ’o . l ex n/)

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	Pronunciation
slutet	NN	/s l ’uu: . t ex t/
slutet	JJ	/s l ”uu: . t ,ex t/

Orthography	Part of Speech	Pronunciation
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	Pronunciation
Anne	swe	/’a n/
Anne	eng	/’ae n/

Appendix D: Language Codes (Examples) 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 ”ae: . r ,u . rn a s
2	posmorph	NN UTR PLU DEF GEN

Field	Name	Example
3	lang	swe
16	case	0
26	id	0060097

Appendix B: Phoneme Table (Base and IPA)

Base	IPA	Example
p	p	pil
i:	i□	sil
y:	y□	syl
ä:	□□	säl
ö:	ø□	rön

Appendix C: Part of Speech

Following (Gustafson-Capková & Hartmann, 2006).

Tag	Description	Example
AB	Adverb	inte
JJ	Adjective	glad
NN	Noun	pudding
VB	Verb	kasta

Appendix D: Language Codes (Examples)

Code	Language
swe	Swedish
eng	English
fre	French
ger	German
rus	Russian