Braxen 1.0

Documentation

# License

CC BY 4.0

<https://creativecommons.org/licenses/by/4.0/legalcode.en>

# Background

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

# Fields

The dictionary consists of 27 fields, enumerated from 0 to 26. In Braxen 1.0, only a selection of the fields is shared (see Appendix A).

## 0 Orthography

The orthography shows the spelling of the word. The intention is to use the mote frequent casing of the word. Different casings can sometimes show different pronunciations (see 16 Case).

Example:

|  |  |  |
| --- | --- | --- |
| **Orthography** | **Part of speech etc.** | **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 |

## 1 Pronunciation

The phonetic-phonological transcriptions in Braxen follows a format developed to cover Swedish, but also foreign speech sounds that frequently occur in Swedish, such as certain English phonemes. This format is here referred to as *Base*.

Appendix B shows a conversion table between *Base* and IPA, and a tool for automatic conversion between the two formats is included in the GitHub repository.

|  |  |  |
| --- | --- | --- |
| **Script** | **Task** | **Example** |
| convertBase2IPA.exe | Converts *Base* to *IPA* | /b j ’oe: rn/ → /bjˈœːɳ/ |
| convertIPA2Base.exe | Converts *IPA* to *Base* | /bjˈœːɳ/ → /b j ’oe: rn/ |
| validateBase.exe | Validates *Base* |  |
| validateIPA.exe | Validates |  |

This section lists the most important principles for the transcriptions.

### Stress

* All words must have exactly one main stress, which can be accent 1 or accent 2.
* If the word has accent 2, it should also have secondary stress.
* Stress is located immediately before the stressed vowel.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 an expression to have multiple main stresses.
* Compounds with ’-’ have accent 2 and secondary stress.
* Syllables can contain only one vowel.

|  |  |  |  |
| --- | --- | --- | --- |
| **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/ |

\*optional, but can occur in the lexicon to mark boundaries that do not follow common syllabification principles.

### /e/ sounds

Braxen has four notations for different /e/ sounds. The dictionary reflects mainly a central Swedish pronunciation, with an overlap between /e/ and /ä/. This has led to some inconsistency in how /ä/ is used in the pronunciations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Phoneme** | **Description** | **Example** | **Transcription** |
| **e** | semi-open /e/ sound | s**e**tt | /s ’**e** t/ |
| **ä** | semi-open /e/ sound (differ from the above in some dialects) | s**ä**tt | /s ’**ä** t/ |
| **eh** | before stressed syllable in unstressed, open syllalbe | b**e**tona | /b **eh** . t ’u: . n a/ |
| **ex** | ”schwa”. In unstressed syllable only, e.g., inflections | boll**e**n | /b ’o . l **ex** n/ |

### Xenophones: foreign speech sounds

The phoneme inventory contains several foreign speech sounds, most of the from English.

# Compatibility with speech technological systems

Different speech technology applications accept different input. Therefore, it is important to adjust the level of detail in the transcriptions according to the context in which you use the lexicon.

## 2 Part of speech and morphological information

Part of speech and morphological information follow mainly the principles of the Stockholm-Umeå Corpus (SUC) (Gustafson-Capková & Hartmann, 2006). However, *UO* (foreign word) is rarely used, since Braxen often contain a language code showing if the word is Swedish or not. All words do not have full morphological information. Appendix C shows a part of speech overview. For more detailed information, see Gustafson-Capková & Hartmann (2006).

Different part of speech and/or morphological information for the same orthography can have different pronunciations:

|  |  |  |
| --- | --- | --- |
| **Orthography** | **Language code** | **Pronunciation** |
| 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/ |

## 3 Language code

The language codes follow the ISO standard 639-2 (Library of Congress, 2017) and consists of three letter combinations. The language code shows which language the orthography was supposed to belong to when the pronunciation was created. When in doubt, *unk* (unknown) has been used. Consequently, a word can occur in several languages and have different pronunciations:

Different language code for the same orthography can involve different pronunciations:

|  |  |  |
| --- | --- | --- |
| **Orthography** | **Language code** | **Pronunciation** |
| Anne | swe | /’a n/ |
| Anne | eng | /’ae n/ |

Appendix D shows a selection of the language codes used.

## 16 Case

This field shows if the entry is case sensitive (1) or not (0), as explained in 1 Orthography above.

## 26 ID

An internal ID for the entry.

# References

Gustafson-Capková, S., & Hartmann, B. (2006). *Manual of the Stockholm Umeå Corpus version 2.0* [Dataset].

Library of Congress. (2017, December 21). *ISO 639-2 Language Code List—Codes for the representation of names of languages*. https://www.loc.gov/standards/iso639-2/php/code\_list.php

# Appendix A: Field information

Bold fields are shared publicly.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Field name** | **Comment** | **Example** |
| **0** | **orth** | **orthography** | **bjärornas** |
| **1** | **pron** | **transcription** | **b j ”ae: . r ,u . rn a s** |
| **2** | **posmorph** | **part of speech + morphology** | **NN UTR PLU DEF GEN** |
| **3** | **lang** | **language (orthography)** | **swe** |
| 4-15 | - | misc |  |
| **16** | **case** | **casing** | **0** |
| 17-25 | acr | misc |  |
| **26** | **id** | **Id** | **0060097** |

# Appendix B: Phoneme table, Base and IPA

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Base** | **IPA** | **Exempel** |  | **Base** | **IPA** | **Exempel** |
| p | p | **p**il |  | i: | iː | s**i**l |
| b | b | **b**il |  | i | ɪ | s**i**ll |
| t | t | bo**t** |  | ih | ɪ̯ | rad**i**o |
| rt | ʈ | bo**rt** |  | y: | yː | s**y**l |
| d | d | bo**d** |  | y | ʏ | s**y**ll |
| rd | ɖ | bor**d** |  | e: | eː | s**e**l |
| k | k | **k**al |  | e | e | s**e**tt |
| g | ɡ | **g**al |  | eh | e̝ | b**e**tona |
| f | f | **f**al |  | ex | ə | papp**e**r |
| v | v | **v**al |  | ä: | ɛː | s**ä**l |
| s | s | **s**al |  | ä | ɛ | s**ä**tt |
| rs | ʂ | fo**rs** |  | ae: | æː | n**ä**r |
| h | h | **h**al |  | ae | æ | sp**ä**rr |
| x | ɧ | **sj**al |  | ö: | øː | r**ö**n |
| c | ɕ | **tj**at |  | ö | ø | r**ö**nn |
| m | m | **m**att |  | oe: | œː | f**ö**r |
| n | n | **n**att |  | oe | œ | f**ö**rr |
| rn | ɳ | ba**rn** |  | u: | uː | b**o**t |
| ng | ŋ | ri**ng** |  | u | u | b**o**tt |
| r | r | **r**iv |  | oh | o | b**o**hem |
| l | l | **l**iv |  | o: | oː | s**å**t |
| rl | ɭ | so**rl** |  | o | ɔ | s**å**tt |
| j | j | **j**ag |  | uu: | ʉː | s**u**r |
| w | w | **w**ay (eng) |  | uu | ɵ | s**u**rr |
| sh | ʃ | **sh**e (eng) |  | uuh | ʉ | b**u**tik |
| zh | ʒ | mea**s**ure (eng) |  | uw: | ʊː | y**ou** |
| z | z | **z**ebra (eng) |  | uw | ʊ | w**ou**ld |
| dh | ð | **th**is (eng) |  | a: | ɑː | m**a**t |
| th | θ | **th**ick (eng) |  | a | a | m**a**tt |
| rh | ɾ | **r**oad (eng) |  | aa: | aː | Zl**a**tan |
| r0 | - | fathe**r** (br. eng) |  | au | aʊ | p**au**s |
| rx | ʀ | **r**ouge (fre) |  | eu | ɛʊ | **eu**ro |
| tc | t͡ʃ | **ch**icken (eng) |  | ei | eɪ | w**ay** (eng) |
| dj | d͡ʒ | fu**dg**e (eng) |  | ai | aɪ | l**ie** (eng) |
| xx | x | Ba**ch** (ger) |  | oi | ɔɪ | b**oy** (eng) |
|  |  |  |  | ou | əʊ | l**ow** (eng) |
|  |  |  |  | eex | eə | wh**ere** (eng) |
|  |  |  |  | iex | ɪə | h**ere** (eng) |
|  |  |  |  | uex | ʊə | p**ure** (eng) |
|  |  |  |  | an | ã | s**an**s (fre) |
|  |  |  |  | en | ɛ̃ | p**ai**n (fre) |
|  |  |  |  | on | õ | pard**on** (fre) |
|  |  |  |  | un | œ̃ | l**un**di (fre) |

# Appendix C: Part of speech

From (Gustafson-Capková & Hartmann, 2006).

|  |  |  |
| --- | --- | --- |
| **Ordklass** | **Förklaring** | **Exempel** |
| AB | Adverb | inte |
| DT | Determinerare | denna |
| HA | Frågande/relativt adverb | när |
| HD | Frågande/relativ determinerare | vilken |
| HP | Frågande/relativt pronomen | som |
| HS | Frågande/relativt possessivt pronomen | vars |
| IE | Infinitivmärke | att |
| IN | Interjektion | ja |
| JJ | Adjektiv | glad |
| KN | Konjunktion | och |
| NN | Substantiv | pudding |
| PC | Particip | utsänd |
| PL | Partikel | ut |
| PM | Egennamn | Mats |
| PN | Pronomen | hon |
| PP | Preposition | av |
| PS | Possessivt pronomen | hennes |
| RG | Grundtal | tre |
| RO | Ordningstal | tredje |
| SN | Subjunktion | att |
| UO | Utländskt ord | the |
| VB | Verb | kasta |

# Appendix D: Language codes (examples)

<https://www.loc.gov/standards/iso639-2/php/code_list.php>

|  |  |
| --- | --- |
| **Code** | **Language** |
| swe | Swedish |
| nob | Norwegian bokmål |
| nno | Norwegien 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 |
| ara | Arabic |

|  |  |
| --- | --- |
| **Code** | **Language group** |
| smi | Sami languages |
| sla | Slavic languages |
| afa | Afro-asian languages |
| aus | Australian languages |
| unk | Unknown language |