Braxen 1.0

Christina Tånnander
Swedish Agency for Accessible Media
KTH Royal Institute of Technology
christina.tannander@mtm.se

Jens Edlund
KTH Royal Institute of Technology
edlund@speech.kth.se

Abstract

With this paper, we release a Swedish pronunciation lexicon resource, Braxen 1.0, which is the result of almost 20 years development carried out at the Swedish Agency for Accessible Media (MTM). The lexicon originated with a basic word list, but has continuously been expanded with new entries, mainly acquired from university textbooks and news text. Braxen consists of around 850 000 entries, of which around 150 000 are proper names. The lexicon is released under the CC BY 4.0 license and is accessible for public use.

1 Introduction

The mission of the Swedish Agency for Accessible Media (MTM) includes the production of accessible materials for individuals with print impairments, such as low vision or dyslexia. This work primarily involves converting books into accessible formats such as Braille and talking books. The talking books are produced through either human narration or text-to-speech synthesis (TTS). MTM uses TTS to produce approximately 1 500 university textbooks annually and more than 120 newspapers on a near-daily basis. While commercial TTS voices are used in this production, the complexity of non-fiction texts often necessitates additional support to ensure accuracy, mainly through pronunciation instructions. The pronunciation dictionary used at MTM, from which Braxen is derived, is referred to as MTM-lex for clarity.

The starting point of MTM-lex was CentLex, a generalised Swedish lexicon for speech technology developed at the academic-industrial centre of excellence CTT in the early 2000s (Jande, 2006). In 2005, as CTT approached the end of its 10-year run, MTM made the decision to develop an in-house TTS system for the production of talking books (Tånnander, 2018). The pronunciation lexicon in the MTM TTS started with 55 000 entries from CentLex, and was supplemented with around 35 000 entries acquired from Svenska språknämndens uttalsordbok (SUO), *67 000 ord i svenskan och deras uttal* (Garlén, 2003). SUO was made publicly available by the Institute for Language and Folklore under the CC-0 license in 2023 (Isof, 2023). MTM has since made significant changes and expansions to the lexicon to meet the substantial demands placed on a pronunciation lexicon used for TTS synthesis of long and information-rich text, such as university textbooks. An in-house format for a phone alphabet used for phonetic transcriptions was developed and inflections of baseforms were added along with hundreds of thousands of new entries, mainly proper names and domain-specific vocabulary.

As part of an active production process, MTM-lex is continuously updated, primarily with words from Swedish newspapers and university textbooks. MTM produces over 120 newspapers in spoken form, read aloud by TTS. The lexicon is updated weekly with the 100+ most frequent news words that are not yet part of MTM-lex. These pronunciations are then forwarded to the TTS system, either as a user lexicon or as SSML insertions in the newspaper document. In addition, MTM produces around 1 500 Swedish and English university textbooks with speech synthesis annually. Frequency lists are computed individually for most books, and new high-frequency words are added to MTM-lex. In this way, the lexicon is kept up-to-date with the current vocabulary of the news world, as well as with vocabularies from specific domains, such as medicine or law.

The sharing of Braxen has been approved by MTM's legal team. The lexicon can be downloaded here: http://www.github.com/sprakbankental/braxen.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 709–713
March 3-4, 2025

2 Initial release: Braxen 1.0

Braxen is not identical to the original MTM-lex and does not include all of its entries or information. Firstly, only 5 of the original 27 fields are included in the first release (see section 4). The remaining fields are excluded for one of the following reasons: they are internal to MTM, unavailable for most of the lexicon entries, lacking in quality or consistency, or simply mere placeholders for future information.

Secondly, not all entries are included in the release. English proper names are included, but approximately 35 000 general English words are not. These words were originally transcribed to match the English variety of a specific Swedish speaker, which was incorporated into the Swedish TTS system mentioned in section 1. As a result, the current pronunciations differ from more established transcription conventions of English.

As with any lexicon, it is virtually impossible to guarantee complete accuracy.

For example, all entries do not have complete PoS information, partially due to the purpose of the resource. Features that are less important in speech-oriented dictionaries, such as whether a word is an adjective or a perfect participle, have been given less attention. We are also aware that there are a small number of incorrect entries.

To the best of our knowledge, this remains the best Swedish resource of its kind available by some margin.

The release includes full documentation and Perl scripts for conversion between the native transcription format and IPA, as well as validation scripts.

3 Statistics

This section presents statistics on a selection of features of general words (852 000 entries, Table 1) and proper names (151 000 entries, Table 2).

Examining the baseforms of the open part-of-speech classes of general words, we count approximately 129 000 nouns, 8 000 verbs, and 16 000 adjectives, present and perfect participles.

4 Fields

This section describes the five fields included in the initial release.

Words	Number	Example
Baseforms	318 000	lexikon
Inflections	534 000	lexikonen
Swedish	679 000	lexikon
English	(35 000)	lexicon
Latin	3 500	humanitatis
Norwegian	2 600	langrenn
German	2 000	Krankheit
French	1 700	ouvrière
Other	7 800	áhkku

Table 1: Word statistics. Note that the English entries are not part of this initial release.

Proper names	Number	Example
Baseforms	151 000	Stockholm
Inflections	23 000	Stockholms
Swedish	91 000	Göteborg
English	16 000	Gothenburg
Other	44 000	København

Table 2: Proper name statistics.

4.1 Orthography

The orthography is displayed in the letter casing that reflects the most common form of the word.

4.2 Part-of-speech (PoS)

The PoS field contains part-of-speech tags and morphological information, following the SUC standard (Ejerhed and Ridings, 2010).

4.3 Language

The language field generally follows the ISO 639-2 standard (Library of Congress, 2017) and indicates which language the pronunciation refers to. Consequently, the same orthography can occur multiple times and have different pronunciations depending on language. Detailed identification of language properties is not a primary task in this work, but words and proper names are classified as belonging either to a specific language or to a pragmatic placeholder category indicating for example a continent associated with the word, such as 'afr' (Africa) or 'asi' (Asia). These placeholder categories are not linguistically accurate, but pave the way for a more refined classification in a forthcoming edition by providing accessible classes for untrained annotators.

4.4 Pronunciation

The pronunciation field contains the standard pronunciation of the orthography. Only one pronunciation variant is present is this first release of Braxen. The details of the symbol set, stress and boundary information are explained in section 5.

4.5 ID

Finally, the ID field contains a unique identification number.

5 Phonetic-phonemic transcription

This section describes the Braxen transcriptions and the symbol set used to encode them. As Braxen is primarily a pronunciation resource for real-world Swedish speech technology, and particularly for TTS synthesis of long, information-rich texts, much care has been taken to create transcriptions that are *useful* for this purpose. This goal takes precedence over strict adherence to any specific speech or language theory, and even over generality in terms of language independence.

The Braxen transcriptions are encoded using a symbol set based on four main design principles, some of which are language-specific. The symbol set can be converted to its IPA equivalent using tools included with the release.

Principle 1: Programming compatibility
Symbols that complicate programming should be excluded.

Principle 1 primarily rejects characters which often serve as control characters in programming languages, such as the SAMPA symbols /{/ and /@/. It also excludes IPA stress and accent notations such as /ˈ/ and /ˌ/, which can complicate the splitting and parsing of pronunciations. The principle also underpins the decision to separate all phonemes by space, as this facilitates splitting words and longer entities into phonemes and makes pronunciation easier to read.

Principle 2: Keyboard accessibility
All symbols should be easily accessible on a Swedish keyboard without compromising ergonomics.

This principle bluntly excludes most IPA symbols and prohibits keyboard combinations (e.g. combinations involving Shift, Alt, or Ctrl). Consequently, it limits the symbols to lowercase characters but allows the inclusion of Swedish alphabetic characters such as “ä” and “ö”.

Principle 3: Visual transparency
Each symbol should preferably resemble its typical orthographic counterpart or its IPA equivalent.

Principle 3 has various implications, such as using the colon “:” as the vowel length marker and “u” for the closed rounded back vowel.

Principle 4: Internal coherence
Each symbol representation should aim for internal coherence, both within the symbol set and within individual symbols.

This principle is especially important for multi-character symbols, where it favours systematic compositionality and mnemonically sound choices.

5.1 Phones

The symbol set consists of 65 phones, 15 of which are xenophones. These are used for speech sounds that are not inherently Swedish, for example /ð - dh/ or /õ - on/. In this section, we describe the rationale behind the notation but refer the reader to the documentation for a complete list of phones and their IPA counterparts.

Following Principle 4, we aim for a consistent use of multi-character symbols when a single-symbol notation is not feasible using the keyboard alone. The additional characters used are presented in Table 3 and include the following:

- The colon marks long vowels: /i:/.
- ”h” is attached to speech sounds to signal some kind of modification of the single symbol, e.g. /ʃ - sh/ and /ð - dh/. This means that we end up with three-character notations of some English diphthongs, e.g. /ɛə- eeh/.
- Nasal vowels are followed by “n”: /an, on/.
- Retroflex speech sounds are preceded by “r”: /rd, rt, rn, rs, rl/.

Symbol	Meaning	Example
_:	long	i:
_h	modified	dh, oh
_n	nasalised	an
r_	retroflex	rt
_x	more back	rx
__	modified	uu:
_0	silent	r0

Table 3: Meaning of control characters placed before or after the main part of the phone.

- ”x” involves a more back pronunciation of the original speech sound, e.g. /ʀ- rx/.

- Similar to attached ”h”, a double notation of a symbol signals a similar, but different phoneme, e.g. /u: - uː/ and /uu: - ʉː/.

- ”Silent speech sounds”, such as R.P. English /r/ are followed by ”0”: /r0/.

5.2 Stress

We use three stress and accent symbols (see Table 4): the primary stress with its two accent variations, and secondary stress. Note that all accent 2 words in Braxen are assigned secondary stress. In most Swedish phonetic transcriptions, the secondary stress is assigned compounds only, and left out in simplex words such as /h “o . p a/. Here, we acknowledge that we have violated both Principle 1 (programming compatibility: the single and double quotes) and Principle 2 (keyboard accessibility: e.g., the Shift key is used for typing accent 2). However, these symbols are justified by their clear connection to stress symbols of other symbol sets: /’/ (primary stress, accent 1) and /,/ (secondary stress) are visually similar to the IPA symbols, and /”/ (accent 2) visually resembles two primary stress symbols combined.

5.3 Boundaries

Three types of boundaries are used: word, compound and syllable boundaries, as shown in Table 4. Again, Pinciple 2 is violated, this time by the word boundary ”|”. We find some reassurance in the fact that this symbol is rarely needed in a pronunciation dictionary, although it is more frequently used in input for applications such as speech synthesis. In these cases, word boundaries are typically inserted automatically.

6 Conclusions and future work

This release of Braxen 1.0 marks a step forward for Swedish speech technology in that it provides an accessible and high-quality pronunciation lexicon for Swedish speech technology applications.

With its comprehensive symbol set tailored to Swedish language needs and adherence to practical design principles, Braxen is well-suited for TTS synthesis and other real-world applications. While the current release offers robust functionality, several areas remain open for enhancement and expansion, and a range of activities are already on the list for upcoming releases:

- Consolidate the excluded 35 000 English MTM-lex entries, and/or add entries from an existing English pronunciation dictionary.

- Implement validation that conforms that pronunciations are plausible given their associated orthography (a complement to existing validation).

- Correct and include other MTM-lex fields that might be of interest to others, such as compound decomposition, pronunciation variations, and word origin.

- Establishing procedures for regular updates to the dictionary, in particular automated transfer of valid additions from MTM-lex to Braxen.

- Release the full symbol set specification.

- Release a free-standing conversion tool between the symbol set used in Braxen and other widespread symbol sets (e.g. SAMPA) in addition to the existing IPA conversion.

Symbol	Meaning	Orthography	Pronunciation
’	accent 1	boll	b **’o** l
”	accent 2	fotboll	f **”u:** t - b ,o l
,	secondary stress	fotboll	f ”u: t - b **,o** l
	unstressed	bollen	b ’o . l **ex** n
|	word	7-eleven	s ’e . v ex n | e . l ’e . v ex n
-	compound	fotboll	f ”u: t **-** b ,o l
.	syllable	bollen	b ’o **.** l ex n

Table 4: Stress and boundaries.

Acknowledgments

MTM-lex was developed by the Swedish Agency for Accessible Media, MTM. Its refactoring and partial release as Braxen has partly taken place in collaboration in the Vinnova project Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish (2018-02427). The resource will be maintained and accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

References

Library of Congress. 2017. Iso 639-2. Accessed: 2024-08-22.

Eva Ejerhed and Daniel Ridings. 2010. Suc - parole.

Claes Garlén. 2003. *Svenska språknämndens uttalsordbok, 67 000 ord i svenskan och deras uttal*. Norsteds Förlag AB.

Isof. 2023. Svenska språknämndens uttalsordbok.

Per-Anders Jande. 2006. *Modelling phone-Level pronunciation in discourse context*. KTH Computer Science and Communication.

Christina Tånnander. 2018. Speech synthesis and evaluation at mtm. In *Proc. of Fonetik 2018*.