

A Swedish pronunciation lexicon for TTS/ASR development

STTS Södermalms talteknologiservice
Östgötagatan 36
SE-116 25 Stockholm
Sweden
contact@stts.se — <http://stts.se>

November 28, 2008

This document presents the Swedish pronunciation lexicon produced by STTS. It is based on the technical documentation accompanying the lexicon files.

Contents

1	Phoneme set and transcription conventions	1
1.1	Diphthongs	1
1.2	No distinction between /E/ and /e/	1
1.3	Short /y/	1
1.4	Stress/tone	1
1.5	Orthographic characters	1
2	File format	2
3	Morphological tag set	3
4	Size	4
5	Frequencies	4
6	Quality	5
7	Undergeneration of forms	5
8	Overgeneration of forms	5
9	Spelling variants	5
10	Terms of use	5

1 Phoneme set and transcription conventions

The transcriptions follow the SAMPA conventions¹, except for the modifications mentioned below. If a word has several pronunciation variants, the first one is considered the default variant. The transcriptions are “Standard Swedish”, understood as a neutral dialect. Regional accent variants are not included. The transcriptions represent a lexical pronunciation, with a preference for a careful (articulated) pronunciation rather than a reduced one. Example: the adverb ‘alltid’ is transcribed ["" a l t i : d] rather than the reduced ["" a l t I].

The transcriptions include no syllable delimiters, i.e., they have not been syllabified.

1.1 Diphthongs

In the current version of the STTS Swedish lexicon, there are no special diphthong phoneme symbols. If the diphthong is in the stressed syllable, the stress is always assigned to one of the vowels in the diphthong, and the other one is considered unstressed. Example: the noun ‘rauk’ is transcribed [r " a u0 k].

1.2 No distinction between /E/ and /e/

The lexicon does not include the distinction between the Swedish SAMPA /E/ and /e/ phonemes. This is due to fact that this distinction is quite subtle and should, in the “Standard Swedish” dialect, be of no importance. Both phonemes are transcribed /e/.

1.3 Short /y/

The short ‘y’, as in ‘bytt’, is transcribed /y/ (while SAMPA has /Y/).

1.4 Stress/tone

The stress/tone symbols used in the lexicon are:

- " Primary stress (tone 1)
- "" Primary stress (tone 2)
- % Secondary stress in tone 2 words

1.5 Orthographic characters

The characters used in the orthographic word forms are a-z, å, ä, ö, é, ü and -.

Compound numbers are delimited using a dash (sjuttio-sju, sextio-nionde, etc).

¹<http://www.phon.ucl.ac.uk/home/sampa/swedish.htm>

2 File format

The delivery comes in two formats:

TXT A tab-separated format. Example:

LEMMA:ackord	POS:substantiv	GENDER:neu
ackord	sin-ind-nom	a k " o: rd
ackords	sin-ind-gen	a k " o: rd rs
ackordet	sin-def-nom	a k " o: rd @ t
ackordets	sin-def-gen	a k " o: rd @ t s
ackord	plu-ind-nom	a k " o: rd
ackords	plu-ind-gen	a k " o: rd rs
ackorden	plu-def-nom	a k " o: rd @ n
ackordens	plu-def-gen	a k " o: rd @ n s

XML An XML format. Example:

```
<entry gender='neu' lemma='ackord' pos='substantiv'>
  <word orth='ackord' tag='sin-ind-nom'>
    <transcription string='a k &quot; o: rd' />
  </word>
  <word orth='ackords' tag='sin-ind-gen'>
    <transcription string='a k &quot; o: rd rs' />
  </word>
  <word orth='ackordet' tag='sin-def-nom'>
    <transcription string='a k &quot; o: rd @ t' />
  </word>
  <word orth='ackordets' tag='sin-def-gen'>
    <transcription string='a k &quot; o: rd @ t s' />
  </word>
  <word orth='ackord' tag='plu-ind-nom'>
    <transcription string='a k &quot; o: rd' />
  </word>
  <word orth='ackords' tag='plu-ind-gen'>
    <transcription string='a k &quot; o: rd rs' />
  </word>
  <word orth='ackorden' tag='plu-def-nom'>
    <transcription string='a k &quot; o: rd @ n' />
  </word>
  <word orth='ackordens' tag='plu-def-gen'>
    <transcription string='a k &quot; o: rd @ n s' />
  </word>
</entry>
```

3 Morphological tag set

The tag-set used for part-of-speech information is similar to the one used in the Stockholm-Umeå Corpus, SUC.

The following part-of-speech categories have been used:

substantiv	noun
verb	verb
deponens	deponent
adjektiv	adjective
adverb	adverb
egennamn	proper noun
interjektion	interjection
konjunktion	conjunction
preposition	preposition
pronomen	pronoun
infinitivmärke	infinitive marker
bokstav	character
grundtal	cardinal number
ordningstal	ordinal number

Other morphological tags:

sin	singular	singular
plu	plural	plural
ind	obestämd	indefinite
def	bestämd	definite
nom	nominativ	nominative
gen	genitiv	genitive
sfo	s-form	s-form
utr	utrum	uter
neu	neutrum	neuter
pos	positiv	positive
sup	superlativ	superlative
mas	maskulinum	masculine
kom	komparativ	comparative
akt	aktiv	active
pas	passiv	passive
pc	particip	participle
prf	perfekt	perfect
prs	presens	present
prt	preteritum	preteritum
inf	infinitiv	infinitive
imp	imperfekt	imperfect
kon	konjunktiv	conjunctive

Pronoun tags:

1	första person	first person
2	andra person	second person
3	tredje person	third person
sub	subjekt	subject
obj	objekt	object
pn	personligt pronomen	personal pronoun
ps	possesivt pronomen	possessive pronoun
ipn	interrogativt personligt pronomen	interrogative personal pronoun
ips	interrogativt possesivt pronomen	interrogative possessive pronoun

4 Size

The following figures describe the size of the lexicon.

WORD FORMS	
Total lemmas	8529
Unique lemmas	8105
Total word forms	58710
Unique word forms	47125

CATEGORIES	
substantiv	3728
verb	1339
egennamn	1300
adjektiv	1170
adverb	356
grundtal	124
ordningstal	120
interjektion	119
pronomen	102
preposition	61
konjunktion	40
deponens	40
bokstav	29
infinitivmärke	1

5 Frequencies

The lexicon is frequency based, but “frequency” is not an unproblematic concept. We have used publically available newspaper frequencies of running words as a starting point. The frequency of a lemma form has been calculated by adding the frequencies of the inflected forms of the lemma. (This means that a lemma of many different inflected forms might win over a lemma with fewer forms.) Since the frequencies are of untagged running words, unusual lemmas that happen to have one or more forms identical to frequent words might be included.

6 Quality

All transcriptions of the lemma forms have been quality checked manually. Most inflected forms have been automatically generated from the lemma form, and many of the automatically generated transcriptions have been manually quality checked. All transcriptions have been automatically validated using an extensive set of tests (including phonetic rules, etc).

We consider STTS' Swedish pronunciation lexicon to be of very high quality.

7 Undergeneration of forms

The lexicon is lacking a few forms in the genitive. These are the proper nouns and the adjective genitive forms. However, these are trivial to automatically generate from the delivered lexicon. (Unless the transcription ends in /s/, just add **s** to the orthographic word, and /s/ to the transcription.)

8 Overgeneration of forms

Since many inflected word forms have been automatically generated from a lemma form using some rule, there will be word forms that might appear peculiar or unlikely. The strangeness of a word form might have to do with semantic, phonologic or other factors. However, there is no general way to determine whether a given word form is “possible” or not. More often than not, it is possible to come up with a context in which an “impossible” inflection makes sense. For example, uncountable nouns, such as ‘mjöl’ (*flour*), is not commonly used in the plural, but when comparing different kinds of flour, one might actually speak about ‘mjöler’ (*flours*), etc. A search on the internet can give an indication of the “possibility” of a particular word form.

We have been liberal in generating inflected word forms.

9 Spelling variants

Our main principle has been not to include spelling variants. The handling of spelling variants can be considered a task for a text pre-processor rather than the lexicon. Words to look out for include *cigarett-cigarrett*, *kafé-café*, *Johnsson-Jonsson-Johnson-Jonson*, etc.

10 Terms of use

If you need information on licensing conditions and terms of use, please contact us.