Nasjonalbiblioteket

NB Uttale - Norwegian pronunciation lexicon

The Norwegian Language Bank

Version: 1.0

About the lexicon

NB Uttale is a pronunciation lexicon for Norwegian Bokmål made by *Språkbanken*, the Norwegian Language Bank at the National Library of Norway. It consists of 785 000 words which are phonemically transcribed into 5 dialects. It is mainly intended for developing speech technology: ASR, forced alignment, grapheme-to-phoneme conversion, speech synthesis. The lexicon is based on the NST lexicon, a pronunciation lexicon from the late 90-ies. Dialect transcriptions are added by means of string replacements, and the lexicon is supplemented with 25 500 new words. The word forms in the lexicon are in Norwegian Bokmål.

What you need to know

Dialect areas and written vs. spoken transcription

There are transcriptions for 5 dialect areas in the pronunciation lexicon:

- east (Oslo, Viken, Innlandet, Vestfold og Telemark)
- south-west (Agder, Rogaland)
- west (Vestland, southern parts of Møre og Romsdal)
- Trøndelag (northern parts of Møre og Romsdal, Trøndelag)
- north (Nordland, Troms og Finnmark)

Each dialect area has two lexicon csv files: one called "spoken", with pronunciation variants typical for unplanned speech such as dialect inflections and dialectal forms of words, and one called "written", with only the most common dialectal phenomena, which is closer to manuscript-read Bokmål. Since the "spoken" transcriptions contain only the most general dialectal phenomena, they don't vary a lot across dialects. For example, the csvs with written transcriptions are exactly the same for Trøndelag and north, and for west and south-west. The spoken transcriptions include a lot more dialectal phenomena, so the transcriptions differ significantly across the lexicon csvs.

The transcription choices in the different dialects are documented in the documents *NB_Uttale_Dialektarbeid.pdf* (Norwegian) and *NB_Uttale_Work_on_Dialects.pdf* (English).

Files

This release contains two zip archives: *nb_uttale_leksika.zip*, containing the lexicon csv files, and *nb_uttale_tillegg.zip*, containing various files from the lexicon project that may be of use. In addition, there are documentation pdfs.

Content of *nb_uttale_leksika.zip*

- *e_spoken_pronunciation_lexicon.csv:* eastern dialect region, spoken-style transcriptions
- *e_written_pronunciation_lexicon.csv:* eastern dialect region, written-style transcriptions
- *w_spoken_pronunciation_lexicon.csv:* western dialect region, spoken-style transcriptions
- *w_written_pronunciation_lexicon.csv:* western dialect region, written-style transcriptions
- *sw_spoken_pronunciation_lexicon.csv:* south-western dialect region, spoken-style transcriptions
- *sw_written_pronunciation_lexicon.csv:* south-western dialect region, written-style transcriptions
- *t_spoken_pronunciation_lexicon.csv:* Trøndelag dialect region, spoken-style transcriptions
- *t_written_pronunciation_lexicon.csv:* Trøndelag dialect region, written-style transcriptions
- *n_spoken_pronunciation_lexicon.csv:* northern dialect region, spoken-style transcriptions
- *n_written_pronunciation_lexicon.csv:* northern dialect region written-style transcriptions

Content of *nb_uttale_tillegg.zip*

- *original_nst_lexicon_nofabet.csv*: The original NST lexicon with NOFABET transcriptions
- *nor030224NST_utf8.pron*: The original NST lexicon file, converted to utf-8. The “unique_id” field corresponds to the “wordform_id” in the dialect lexicon files.
- *rules_v1.py:* the string transformation rules producing NB Uttale from the transcriptions in *original_nst_lexicon_nofabet.csv* (more on this below)
- *exemptions_v1.py*: words which are exempt from certain string transformation rules in *rules_v1.py*
- *newwords_2022.csv*: New words added to the lexicon
- *conversion.py:* Python code for converting from the NOFABET standard to X-SAMPA and IPA (see below)

Documentation

- *NB_Uttale_v_1_0_documentation.pdf*: This document
- *NB_Uttale_Dialektarbeid.pdf:* Detailed documentation of the dialect transcriptions (in Norwegian)
- *NB_Uttale_Work_on_Dialects.pdf:* An overview of the dialect transcriptions in English

The documentation of the original NST lexicon can be found here (in Norwegian).

Columns in the lexicon csv files

- *wordform*: The orthographic word form
- *pos:* part of speech tag, following the same convention as in the NST lexicon
- *feats:* morphological (inflectional) features, following the same convention as in the NST lexicon
- *wordform_id*: The unique identifier of the wordform and its grammatical information. New entries that have been added from *newwords_2022.csv* have the prefix “NB”.
- *update_info*: The “update_info” column from the original NST lexicon. For the new words added in this project, the column specifies whether the word comes from Målfrid (value: *Målfrid)* or the Norwegian Newspaper Corpus (value: *NB-aviskorpus*)
- *nofabet_transcription*: Phonetic transcription in the NOFABET standard (see below). This is the standard used in the NB Uttale project.
- *ipa_transcription:* Transcriptions in IPA, automatically converted from NOFABET
- *sampa_transcription*: Transcription in X-SAMPA, automatically converted from NOFABET

Transcription standards

The original transcriptions of the NB Uttale are in the NOFABET standard, developed for a previous project at the Norwegian Language Bank. The transcriptions are also converted to X-SAMPA and IPA. We will describe the NOFABET standard in detail, since it is less commonly used. For the two other standards, we refer to existing descriptions, but point out some peculiarities of the transcriptions in this lexicon.

The NOFABET standard

NOFABET is an ARPABET-inspired standard. Each phoneme is written as a sequence of one or more uppercase ASCII letters. Phonemes are split by a whitespace. Vowels and syllabic consonants are followed by a number between 0 and 3, where 0 indicates that the syllable is unstressed, 1 that it is stressed and has tone 1, 2 that it is stressed and has tone 2, and 3 that it has secondary stress. The eastern pronunciation of “barnehelsen” (‘ the children’s health’) may be transcribed as follows: *B AA2 RN AX0 H EH3 L S NX0. AA2* is a long “a” with tone 2. *RN* is a retroflex “n”. *AX0* is an unstressed schwa. *EH3* is a short “e” with secondary stress. *NX0* is a syllabic “n”. There is no marking of syllable or compound boundaries in NOFABET.

Equivalence table

The following table gives the NOFABET phone transcription and the corresponding transcription in the X-SAMPA and IPA transcriptions used in the lexicon.

NOFABET	X-SAMPA	IPA
AE	{:	æː
AEH	{	æ
AEJ	{*I	æ͡ɪ
AEW	E*u0	æ͡ʉ
AH	A	ɑ
AJ	A*I	ɑ͡ɪ
AX	@	ə
B	b	b
D	d	d
EE	e:	eː
EH	E	ɛ
F	f	f
G	g	g
H	h	h
IH	I	ɪ
II	i:	ɪː
J	j	j
K	k	k
KJ	C	ç
L	l	l
LX	l=	l̩
M	m	m
MX	m=	m̩
N	n	n
NG	N	ŋ
NX	n=	n̩
OA	o:	oː
OAH	O	ɔ

OAJ	O*Y	ɔ͡ʏ
OE	2:	øː
OEH	9	œ
OEJ	9*Y	œ͡ʏ
OH	U	ʊ
OO	u:	uː
OU	@U	o͡ʊ
P	p	p
R	r	r
RD	d`	ɖ
RL	l`	ɭ
RLX	l`=	ɭ̩
RN	n`	ɳ
RNX	n`=	ɳ̩
RS	s`	ʂ
RT	t`	ʈ
RX	r=	r̩
S	s	s
SJ	S	ʃ
SX	s=	s̩
T	t	t
UH	u0	ʉ
UU	}:	ʉː
V	v	v
W	w	w
YH	Y	ʏ
YY	y:	yː
AA	A:	ɑː

The X-SAMPA standard

The X-SAMPA standard used in the NB Uttale follows mostly the one used in the NST lexicon, except that compound boundaries and phrase accents are not marked.

Secondary stress is marked with *%* at the beginning of the syllable. Primary stress and tone 1 and 2 are marked with " and "" respectively at the beginning of the syllable. Syllable boundaries are marked by *$*. The transcription of “rapporten” (‘the report’) is *rA$"pO$t`n`=*

IPA

The IPA transcriptions are mostly standard, except that tone 1 and 2 are marked with ′ and ″ respectively.

Reproducing our work

For users who wish to reproduce our work, we have set up a Github repository with code and instructions for how to do so.

How to get in touch with us or contribute

If you have any questions or suggestions for changes, do not hesitate to get in touch with us by sending an email to sprakbanken@nb.no. We are also interested in bug reports or suggestions for improvements in the form of issues and pull requests in the Github repository.

How the lexicon was developed

The lexicon is in part developed through string transformations on the transcriptions of a legacy lexicon, the NST lexicon, developed by the firm Nordisk språkteknologi in the late 90-ies. In addition, new vocabulary with corresponding transcriptions have been added and run through the same string transformations. Below we will describe the string transformations and then the addition of new vocabulary.

Dialect transcriptions via string transformations

The NST lexicon only had transcriptions in eastern Norwegian. A team of trained linguists wrote string transformation rules to create dialect transcriptions to 4 new dialects from the original transcriptions in the NST. The string transformation rules and lists of transformed words were inspected by a colleague before they were applied to the lexicon. A separate document (in Norwegian) describes in detail the dialect transformations made. All the string transformation rules are also part of the release, as well as lists of words which are exempt from certain rules. These are some main groups of changes:

- **Corrections:** The NST lexicon contains transcriptions which do not reflect the actual pronunciation, as well as some errors and inconsistencies. When the team discovered such cases, they were corrected across all dialects.
- **Systematic phonological differences:** For certain phones and phone clusters, there are systematic differences between the eastern dialects and the other dialects. For example, clusters of “d” and subsequent “n”, “t”, “d”, “l” and “s” fuse into retroflex

consonants in all dialect areas except the western and southwestern dialects. We have rules that transform retroflex consonants into non-retroflex consonants in those dialects.

- **Unsystematic phonological differences:** When phonological differences affect only certain words, string replacement rules are constrained to certain word forms or words with certain grammatical tags.
- **Inflectional differences**: To capture inflectional differences between the eastern dialect and the other dialects, string transformation rules are constrained by POS tags and grammatical features. In some cases, inflectional differences concern only certain inflectional classes, and therefore cannot be constrained by POS tags and grammatical features alone. In such cases, the rules are constrained to lists of frequent words belonging to those inflectional classes.

Dialect words with no official orthography or dialectal forms of words which differ substantially from their written form are considered to be outside the scope of a pronunciation lexicon like this and are therefore not part of this release. You will, for example, not find different forms of the personal pronoun “dere” (‘you’), such as “dokker”, “dokk”, “dekkan” etc.

The NST lexicon allowed for up to 4 pronunciation variants for a word.[1] All of these variants were processed by the string replacement rules and were added as a separate line row in the lexicon csv. No additional variants of words in the NST lexicon were added, however. It may occur that a string replacement rule has removed the difference between variants of a word in one or several of the csv files, but there will still be a separate row for each original pronunciation variant. All variants of the same word will have the same *wordform_id*.

The original transcriptions in NST were in the X-SAMPA standard. They were converted to NOFABET prior to the string transformations, as X-SAMPA is not well suited for regex rules and is not particularly human-readable. The NST transcriptions had markings of compound and syllable boundaries. These were removed when the transcriptions were converted to NOFABET, and, consequently, the string transformation rules are written without regard to such boundaries.

At the end of the project, the NOFABET transcriptions were converted to X-SAMPA and IPA. In the X-SAMPA and IPA transcriptions, syllable boundaries were reintroduced.[2]

[1] I.e. irregular pronunciation variants of the same word in eastern dialects, such as “energi” which can be pronounced as /EH0 N AX0 R G II1/ or /EH0 N AX0 RS II1/. The NST project also produced some regular dialectal pronunciation variants (given in field 27 in *nor030224NST_utf8.pron*). These are not included in this project.

[2] The algorithm reintroducing these boundaries is based on chapter 4 of Kristoffersen, Gjert.2015. En kort innføring i norsk fonologi (4th. ed.). Bergen: University of Bergen. The algorithm disregards compound boundaries, however, as it does not have access to information about them.

New words

The NST lexicon consists of 785 000 distinct words. An additional 25 500 words are added to NB Uttale, with up to 4 pronunciation variants. To find new words, words and frequencies were extracted from the Norwegian Newspaper Corpus, a 1.7B word corpus with news text from 1998-2019, as well as from the Målfrid Corpus, a 4.1B word corpus with text from government websites. Words occurring in the NST lexicon were removed, and the remaining words were tagged with POS tags and morphological features. The linguists on the NB Uttale team extracted word lists from this material and ran them through a grapheme-to-phoneme converter trained on the NST lexicon which predicted transcriptions for the different words. They then corrected the transcriptions and the grammatical tags. Subsequently, the new words were run through the same string transformation rules as the NST vocabulary.