Uttaleleksikon fra Norsk lyd- og blindeskriftbibliotek

(FOR ENGLISH VERSION, SEE BELOW)

Dette uttaleleksikonet er utviklet av Norsk lyd- og blindeskriftbibliotek (NLB) for bruk i Filibuster, NLBs tekst-til-tale-system. Leksikonet bygger på uttaleleksikonet til Nordisk språkteknologi (NST). For en beskrivelse av NSTs uttaleleksikon, se filene nst_leksdat_no.pdf og transcription_conventions_nst.pdf.

Transkripsjonsretningslinjene i NLBs leksikon er beskrevet i dokumentet transkripsjon_nlb.pdf. Transkripsjonene er gjort i SAMPA, se https://www.phon.ucl.ac.uk/home/sampa/index.html.

I tillegg til en ortografisk og en fonetisk representasjon av ordene, inneholder leksikonet flere kolonner med annen type informasjon, f.eks. grammatiske tagger. Se dokumentasjonen til NST-leksikonet for en beskrivelse av dette.

Leksikonet består av tre deler for bokmål: et hovedleksikon, et navneleksikon og et leksikon med et utvalg engelske ord. For nynorsk finnes et leksikon med automatisk genererte transkripsjoner, der man har tatt utgangspunkt i Norsk ordbank. Tanken bak dette leksikonet var å forbedre opplesningen av nynorsk for stemmen til NLBs talesyntese, Brage.

Leksikonfilene er i TSV-format og kan enkelt importeres til ulike programmer som håndterer strukturerte data. Filene har filekstensjonen .lex.

Filoversikt

- nlb_nob_20181129.lex
  Frekvensbasert liste over ord på bokmål, basert på NSTs leksikon. NLB har oppdatert listen med flere ord. Inneholder 744 442 poster.

- nlb_nob_p_20181129.lex
  Liste over egennavn basert på NSTs leksikon. NLB har oppdatert listen med flere navn. Inneholder 54 270 poster.

- nlb_eng_20181129.lex
  Et utvalg engelske ord transkribert med en "fornorsket" uttale (se transkripsjon_nlb.pdf). Inneholder 13 446 poster.

- nlb_nno_20190712.lex
  Uttaleleksikon med automatisk genererte transkripsjoner for nynorsk med utgangspunkt i Norsk ordbank for nynorsk. Inneholder 352 788 poster

I tillegg til leksikonfilene finnes dokumentasjon og en katalog med ulike lister og diverse som vi dessverre ikke har en nærmere beskrivelse av.

Dokumentasjon

- nst_leksdat_no.pdf - beskrivelse av NSTs uttaleleksikon (norsk)
- transcription_conventions_nst.pdf - transkripsjonsretningslinjer for NSTs uttaleleksikon (engelsk)
- transcription_nlb.pdf - transkripsjonsretningslinjer for NLBs leksikon (norsk/engelsk)
- lesmeg.pdf - pdf-versjon av denne filen
- lesmeg.txt - denne filen

Nasjonalbiblioteket, 2019-10-16

Pronunciation lexicon from The Norwegian Library of Talking Books and Braille (NLB)

This pronunciation lexicon has been developed by NLB for use in Filibuster, NLB's system for text-to-speech. The lexicon is built on the pronunciation lexicon developed by Nordic Language Technology (NST). For a description of NST's pronunciation lexicon, see the files nst_leksdat_no.pdf (in Norwegian) and transcription_conventions_nst.pdf.

The transcription guidelines for NLB's lexicon are described in the document transcription_nlb.pdf. The transcriptions are done using SAMPA, see https://www.phon.ucl.ac.uk/home/sampa/index.html.

In addition to an orthographic and a phonetic representation of the words, the lexicon contains several columns with additional information, e.g. POS and grammatical tags. See the documentation for the NST lexicon for a closer description of this.

The lexicon concists of three parts for Norwegian Bokmål: a main lexicon, a lexicon containing proper names and a lexicon containing a selection of English words.

For Norwegian Nynorsk there is a lexicon with automatically generated transcriptions. This lexicon is based on Norsk ordbank, and no manual corrections have been made to the output. The idea behind this lexicon was to improve the Nynorsk pronunciation of NLB's speech synthesis Brage.

The lexcion files are in TSV format and can easily be imported into applications which can handle structured data. The files have the file extension .lex.

File overview

- nlb_nob_20181129.lex
  Frequency based list of words in Norwegian Bokmål based on NST lexicon. NLB has updated the list with additional words. Contains 744 442 items.

- nlb_nob_p_20181129.lex
  List of proper names based on the NST lexicon. NLB has updated the list with additional names. Contains 54 270 items.

- nlb_eng_20181129.lex
  A selection of English words transcribed with a Norwegian sounding pronunciation (see transcription_nlb.pdf). Contains 13 446 items.

- nlb_nno_20190712.lex
  Pronunciation lexicon for Norwegian Nynorsk based on Norsk ordbank. The transcriptions are automatically generated. Contains 352 788 items.

In addition to the lexicon files, there are documentation files and a catalogue containing various lists and other things which unfortunately lack a description.

Documentation

- nst_leksdat_no.pdf - description of NST pronunciation lexicon (Norwegian)
- transcription_conventions_nst.pdf - transcription guidelines for the NST lexicon (English)
- transcription_nlb.pdf - transcription guidelines for the NLB lexicon (Norwegian/English)
- lesmeg.pdf - pdf version of this file
- lesmeg.txt – readme file (this file)

The National Library of Norway, 2019-10-16