Pronunciation Lexicon for Norwegian Nynorsk

This pronunciation lexicon for Nynorsk was originally developed by Lingit AS to be used in their Text-to-Speech voices which were first released in 2008.

The resource consists of a set of lexical items, each associated with an inflected word form, a single pronunciation, lexical features and lemma. It does not contain names.

An accompanying helper script provides means for expanding information like tags into feature maps. It also computes various stats and it serves as a practical example of how to load and use the resource.

Resource

Kind	Count
Items	570390
Unique (inflected form,tag) pairs	561132
Unique inflected forms	422528
Unique pronunciations	415827

Files

- `lgt_pronlex_nn.py` -- a helper script.
- `lgt_pronlex_nn_20210315.txt.gz` -- the main resource file.
- `lgt_pronlex_nn_20210315_info.json` -- metainfo generated by the helper script
- `lgt_pronlex_nn_20210315_stats.json` -- stats generated by the helper script

Lexicon File Format

The resource is stored as a basic, UTF-8 encoded CSV-file with a three line header and non-significant blank lines between lemma groups. Each content line defines a lexical item with 5 TAB separated fields. For automatic processing, all blank lines and all lines starting with "#" can be ignored.

Below is a small but complete example of a resource file:

```
# coding: utf8
# separator: \t
#

bil     "bi:l        noun.sg.ind.masc   bil   8472
bilane  ""bi:$lA$n@  noun.pl.def.masc   bil   8472
bilar   ""bi:$lA4    noun.pl.ind.masc   bil   8472
bilen   "bi:$l@n     noun.sg.def.masc   bil   8472
```

Fields

Logically, the resource has 6 fields. The first five are stored in file, the sixth is derived automatically.

Field	Name	Example	Comment
1	Inflection	bilar	
2	Pronunciation	“”bi:lA4	X-SAMPA
3	Tag	noun.pl.ind.masc	May be converted to the expanded form in field 6.
4	Lemma	bil	
5	Group ID	8472	Identifies a group of items that share lemma and base tag.
6	Features	form:indefinite gender:masculine number:plural pos:noun	Automatically derived from Tag, not stored in file.

Helper Script

The helper script provides means for computing verbose feature maps, for computing stats and it can dump information about tags and feature mappings. It requires Python 3 and have not been tested on Windows. There is no need to unarchive the resource file before you run the script.

Use the help system to see details about what you can do:

```
>>> lgt_pronlex_nn.py --help
```

Examples

Show tag and feature definitions:

```
>>> lgt_pronlex_nn.py show-features
```

Expand alias tags, add verbose features and print the result:

```
>>> lgt_pronlex_nn.py expand-lex
```

Compute basic stats:

```
>>> lgt_pronlex_nn.py show-stats
```

Compute basic stats from the expanded form:

```
>>> lgt_pronlex_nn.py --expand-lex show-stats
```

Basic stats with some modifications

```
>>> lgt_pronlex_nn.py show-stats --strip-tag-flags --countsort --reverse
--unique-pronunciations
```

Pronunciations

Pronunciations are represented as X-SAMPA (https://en.wikipedia.org/wiki/X-SAMPA) and include primary and secondary stress markers and syllable boundaries. The symbol set in use resembles the traditional Norwegian SAMPA (https://www.phon.ucl.ac.uk/home/sampa/norweg.htm), but some additional symbols that are useful for Text-to-Speech applications have been added. All symbols are listed in the table below.

The general transcription guideline is as follows:

- Prefer near-orthographic pronunciation.
- Retroflexion at morpheme boundaries is generally allowed even though this phenomenon does not occur in all Norwegian dialects.
- It is OK to use a rich symbol set, and preference is towards using symbols that are possible to reduce to a more restricted set. For instance, the use of `/r`/` as in `/"br`{ik/` can always be reduced to `/"bl{ik/`. The opposite transformation is not always possible. Similarly, transcriptions with English symbols can often be automatically reduced to a Norwegian approximation.

Syllable Boundaries

The general guideline for syllable boundary placement is:

- Word boundaries generally overrides other rules
- Maximum onset
- Retroflex consonants are accepted as onset, e.g. `/v{$d`i:/` is preferred over `/v{d`$i:/`

Boundaries were initially computed automatically and later edited manually when needed from a TTS application performance point of view. Academic correctness has not been the ultimate goal and the resource has some inconsistencies, particularly related to compound word boundaries which the initial automatic approach did not take into consideration.

Stress Markers

Transcriptions are marked with primary and secondary stress markers. Two primary stress symbols (for toneme 1 and toneme 2) and one secondary stress symbol are used, as shown in the symbol table. Secondary stress placement has been somewhat guided by TTS application performance.

General rules are:

- One unique primary stress per transcription.
- A multi-word transcription (which contains whitespace) may have more than one primary stress.

X-SAMPA Symbols

Symbol	**Word**	**Transcription**	**Comment**
Non-segmentals			
$	lete	""le:$t@	Syllable boundary
"	laget	"lA:$g@	Primary stress, toneme 1
""	lage	""lA:$g@	Primary stress, toneme 2
%	T-banestasjon	"te:$bA:$n@stA%Su:n	Secondary stress
			Word boundary (whitespace character)
Short vowels			
A	hatt	"hAt	
@	håpe	""hO:$p@	Always unstressed
i	litt	"lit	
e	tett	"tet	
u	pukk	"puk	
O	topp	"tOp	

}	dugg	"d}g	
y	lytte	""ly$t@	
2	tørr	"t24	
{	herr	"h{4	
V	paste-up	"peIst$%Vp	Extension
Long vowels			
A:	dag	"dA:g	
e:	le	"le:	
i:	si	"si:	
u:	to	"tu:	
O:	tå	"tO:	
}:	ku	"k}:	
2:	rød	"42:	
y:	sy	"sy:	
{:	bær	"b{:4	
Diphthongs			
{i	deig	"d{ig	
A}	au	"A}	
2y	høy	"h2y	
Ai	"hai	"hAi	
eI	play	"pleI	Extension
Oy	joik	"jOyk	
}i	hui	"h}i	
@U	soul	"s@Ul	Extension
aU	layout	"leI$aUt	Extension
Plosives			
p	papp	"pAp	
b	babb	"bAb	
t	tatt	"tAt	
d	gadd	"gAd	
k	kakk	"kAk	
g	tagg	"tAg	
Fricatives			
f	uff	"}f	
s	sess	"ses	

S	tusj	"t}S	
v	lav	"lA:v	
j	jul	"j}:l	
h	ha	"hA:	
C	kje	"Ce:	
Retroflex consonants			
d`	høyrd	"h2yd`	
l`	berlinsk	b{$"l`i:nsk	
n`	barn	"bA:n`	
r`	blå	"br`O:	
t`	hardt	"hAt`	
Sonorant consonants			
m	tam	"tAm	
n	tann	"tAn	
N	sang	"sAN	
l	ball	"bAl	
r\	teamwork	"ti:m$%w2:r\k	Extension
w	teamwork	"ti:m$%w2:r\k	Extension
Other consonants			
4	tørr	"t24	
d__Z	gin	"d__Zin	Extension
t__S	imaget	"imit__S@	Extension
Syllabic consonants			
4=			Edge case
n=	mannen	"mAn$n=	
n`=	fullfaren	""f}l$%fA:$n`=	
l=	middel	"mi$dl=	
l`=	fengsel	"feN$Sl`=	
m=			
s=	pst	"ps=t	Edge case

Tags

Each lexical item is associated with a feature tag that encodes part-of-speech, lexical features and other flags. The general structure is.

```
FEATURE_TAG = MAIN_TAG ( “.” FEAT )* ( “+” FLAG )*
```

As en example, the tag **`noun.pl.def.neut+irr`** corresponds to:

pos	number	form	gender	flag
noun	plural	definite	neuter	irregular

Some special *alias* tags are also defined. These logically expands to multiple feature tags. All aliases, flags and tags in use are described by the accompanying `lgt_pronlex_nn_20210315_info.json` file and in the tables below.

Alias Tags

alias	tags
adj.def_pl	adj.sg.def.fem adj.sg.def.masc adj.sg.def.neut adj.pl
adj.sg.ind	adj.sg.ind.fem adj.sg.ind.masc adj.sg.ind.neut
adj.sg.ind.fem_masc	adj.sg.ind.fem adj.sg.ind.masc

Flags

flag	value	comment
abbr	abbreviation	
irr	irregular	Spelling is not consistent with current norms

Nouns

tag	pos	number	form	gender
noun.sg.ind.fem	noun	singular	indefinite	feminine
noun.sg.ind.masc	noun	singular	indefinite	masculine
noun.sg.ind.neut	noun	singular	indefinite	neuter

noun.sg.ind	noun	singular	indefinite	
noun.sg.def.fem	noun	singular	definite	feminine
noun.sg.def.masc	noun	singular	definite	masculine
noun.sg.def.neut	noun	singular	definite	neuter
noun.sg.def	noun	singular	definite	
noun.pl.ind.masc	noun	plural	indefinite	masculine
noun.pl.ind.fem	noun	plural	indefinite	feminine
noun.pl.ind.neut	noun	plural	indefinite	neuter
noun.pl.ind	noun	plural	indefinite	
noun.pl.def.masc	noun	plural	definite	masculine
noun.pl.def.fem	noun	plural	definite	feminine
noun.pl.def.neut	noun	plural	definite	neuter
noun.pl.def	noun	plural	definite	
noun	noun			

Adjectives

tag	pos	number	degree	form	gender
adj.sg.ind.fem	adjective	singular	positive	indefinite	feminine
adj.sg.ind.masc	adjective	singular	positive	indefinite	masculine
adj.sg.ind.neut	adjective	singular	positive	indefinite	neuter
adj.sg.def.fem	adjective	singular	positive	definite	feminine
adj.sg.def.masc	adjective	singular	positive	definite	masculine
adj.sg.def.neut	adjective	singular	positive	definite	neuter
adj	adjective				
adj.pl	adjective	plural	positive		
adj.comp	adjective		comparative		
adj.sup.ind	adjective		superlative	indefinite	
adj.sup.def	adjective		superlative	definite	

Verbs

tag	pos	type
verb.imp	verb	imperative
verb.inf	verb	infinitive
verb.pres	verb	present

verb.pret	verb	preterite
verb.presp	verb	present_participle
verb.perfp	verb	past_participle
verb.pas	verb	passive

Pronouns

tag	pos	number	form	gender
pron	pronoun			
pron.pl	pronoun	plural		
pron.sg.fem	pronoun	singular		feminine
pron.sg.masc	pronoun	singular		masculine
pron.sg.neut	pronoun	singular		neuter
pron.sg	pronoun	singular		
pron.def	pronoun		definite	

Determiners

tag	pos	number	form	gender
det	determiner			
det.sg.fem	determiner	singular		feminine
det.sg.masc	determiner	singular		masculine
det.sg.neut	determiner	singular		neuter
det.pl	determiner	plural		
det.sg	determiner	singular		
det.def	determiner		definite	

Others

tag	pos
adv	adverb
conj	conjunction
num	numeral
ord	numeral
inf	infinitive_marker
interj	interjection
prep	preposition