Grapheme to Phoneme models for Norwegian Bokmål

Version: 2.0 (2024-02-09)

This repo contains G2P models for Norwegian bokmål[^1], which produce phonemic transcriptions for *close-to-spoken* pronunciations (such as in spontaneous conversations: `spoken`) and *close-to-written* pronunciations (such as when reading text aloud: `written`) for 5 different dialect areas:

1. East Norwegian (`e`)
2. South West Norwegian (`sw`)
3. West Norwegian (`w`)
4. Central Norwegian (Trøndersk) (`t`)
5. North Norwegian (`n`)

[^1]: Bokmål is the most widely used written standard for Norwegian. The other written standard is Nynorsk. Read more on Wikipedia.

Setup

Follow installation instructions from Phonetisaurus. You only need the steps "Next grab and install OpenFst-1.7.2" and "Checkout the latest Phonetisaurus from master and compile without bindings".

Data

The pronunciation lexica that were used to train the G2P-models are free to download and use from Språkbanken's resource catalogue: NB Uttale

For more information about the lexica, see the Github repo: Sprakbanken/nb_uttale

Content

- `models/`: contains the models, as well as auxiliary files used by Phonetisaurus
  - `nb_*.fst`: model files to use with `phonetisaurus-apply`. The expansion of `*` is a string of a dialect and pronunciation style, e.g. `e_spoken` or `t_written`.
  - `nb_*.o8.arpa`: 8-gram-models for phoneme sequences that Phonetisaurus uses during training.
  - `nb_*.corpus`: aligned graphemes and phonemes from the lexica.
- `data/`: contains various lexica used for training and testing, including predictions from the models on the test set
  - `NB-uttale_*_train.dict`: training data for `models/nb_*.fst`. Each file contains 543 495 word-transcription pairs (WTP), and makes up 80% of all unique WTPs in the lexicon.
  - `NB-uttale_*_test.dict`: test data for `models/nb_*.fst`. Each file contains the remaining 20% of the WTPs in the lexicon, i.e. 135 787 WTPs.
  - `predicted_nb_*.dict`: The words from the testdata with the model's predicted transriptions.

  - `wordlist_test.txt`: The orthographic words from the test data, which the models predict transcriptions for.
- `evaluate.py`: script to evaluate the models. The method for calculating WER and PER were re-implemented.
- `g2p_stats.py`: script to evaluate the models from V1.0, which can be used to compare results between these models and the NDT models (with and without tone and stress markers) from version 1.
- `LICENSE`: The license text for CC0, which this resource is distributed with.

Usage

```
phonetisaurus-apply --model models/nb_e_spoken.fst --word_list
data/wordlist_test.txt  -n 1  -v  > output.txt
```

- Input data (`--word_list`) should be a list of newline-delimited words. See the file `data/wordlist_test.txt` for an example.
- The trained G2P-models are `.fst` files located in the `models` folder. The same folder also contains aligned `.corpus` files and phoneme 8-gram models (`.arpa` files), also from the `Phonetisaurus` training process.
- `-n` lets you adjust the number of most probable predictons.

Evaluation

There are 2 scripts to calculate WER and PER statistics, which give slightly different results.

`evaluate.py`

Calculates stats for all the provided models by default. You can give a pronunciation variant (e.g. `-l e_spoken`) to calculate stats for specific models.

- The WER score is calculated as the count of all mismatching transcriptions (1 error = 1 mismatching word) divided by the count of all words in the reference, i.e. a `*_test.dict` file.
- PER is calculated as the count of all errors (1 error = a mismatching phoneme) divided by the total count of phonemes in the reference file.

```
python evaluate.py
```

Model	Word Error Rate	Phoneme Error Rate
nb_e_written.fst	13.661238654564869	1.9681178920293207
nb_e_spoken.fst	13.72501038144391	1.9832518286152074
nb_sw_written.fst	13.240048644480037	1.8396612197218096
nb_sw_spoken.fst	16.422702734768936	2.426312206336983

Model	Word Error Rate	Phoneme Error Rate
nb_w_written.fst	13.240048644480037	1.8396612197218096
nb_w_spoken.fst	16.892833837574894	2.5064155890730686
nb_t_written.fst	13.736133357062347	1.98774986044724
nb_t_spoken.fst	16.47992288013051	2.5809178688066843
nb_n_written.fst	13.736133357062347	1.98774986044724
nb_n_spoken.fst	17.22590930999963	2.8209779677747715

`g2p_stats.py`

Calculates WER and PER for two input files.

1. The reference file (e.g. `data/NB-uttale_e_spoken_test.dict`)
2. The model prediction file (e.g. `output.txt` from the command in Usage, or `data/predicted_nb_e_spoken.dict`).

- The WER score is calculated as the count of errors (1 error = 1 mismatching word) divided by the count of all words in the predictions, i.e. a `predicted_*.dict` file.
- PER is calculated as the sum of phone error rates for each transcription, divided by the total count of words in the predictions.

> **NOTE**: This method doesn't take transcription lengths into account, so a transcription with 2 phonemes where 1 is wrong has a 0.5 PER while a word with length 10 with 1 error has a 0.1 PER, and the average score for the two words would be 0.35.

```
python g2p_stats.py data/NB-uttale_e_spoken_test.dict
data/predicted_nb_e_spoken.dict
# WER: 14.049209423582523
# PER: 2.714882650391985
```

Model	Word Error Rate	Phoneme Error Rate
nb_e_written.fst	13.97114598599277	2.7038190765903214
nb_e_spoken.fst	14.049209423582523	2.714882650391985
nb_sw_written.fst	13.541060631724685	2.5423757844377284
nb_sw_spoken.fst	16.729141964989285	3.34063477772742
nb_w_written.fst	13.541060631724685	2.5423757844377284
nb_w_spoken.fst	17.186475877661337	3.4137304874392114
nb_t_written.fst	14.059519688924565	2.7190289235234104
nb_t_spoken.fst	--	--

Model	Word Error Rate	Phoneme Error Rate
nb_n_written.fst	14.059519688924565	2.7190289235234104
nb_n_spoken.fst	--	--

> **NOTE**: *The t_spoken and n_spoken model predictions are not the same length as the reference file, which causes the script to exit.*

Transcription standard

The G2P models have been trained on the NoFAbet transcription standard which is easier to read by humans than X-SAMPA. NoFAbet is in part based on 2-letter ARPAbet and is made by Nate Young for the National Library of Norway in connection with the development of *NoFA*, a forced aligner for Norwegian. The equivalence table below contains X-SAMPA, IPA and NoFAbet notatations.

X-SAMPA-IPA-NoFAbet equivalence table

X-SAMPA	IPA	NoFAbet	Example
A:	ɑː	AA0	b**a**d
{:	æː	AE0	v**æ**r
{	æ	AEH0	v**æ**rt
{*I	æɪ	AEJ0	s**ei**
E*u0	æʉ	AEW0	s**au**
A	ɑ	AH0	h**a**tt
A*I	ɑɪ	AJ0	k**ai**
@	ə	AX0	b**e**hage
b	b	B	**b**il
d	d	D	**d**ag
e:	eː	EE0	l**e**k
E	ɛ	EH0	p**e**nn
f	f	F	**f**in
g	g	G	**g**ul
h	h	H	**h**es
I	ɪ	IH0	s**i**tt
i:	iː	II0	v**i**n
j	j	J	**j**a
k	k	K	**k**ost

X-SAMPA	IPA	NoFAbet	Example
C	ç	KJ	**k**ino
l	l	L	**l**and
l=	l̩	LX0	
m	m	M	**m**an
n	n	N	**n**ord
N	ŋ	NG	e**ng**
n=	n̩	NX0	
o:	oː	OA0	r**å**
O	ɔ	OAH0	g**å**tt
2:	øː	OE0	l**ø**k
9	œ	OEH0	h**ø**st
9*Y	œy	OEJ0	k**øy**e
U	u	OH0	f***o**rt
O*Y	ɔy	OJ0	konv**oy**
u:	uː	OO0	b**o**d
@U	oʉ	OU0	sh**ow**
p	p	P	**p**il
r	r	R	**r**ose
d`	ɖ	RD	reko**rd**
l`	ɭ	RL	pe**rl**e
l`=	ɭ̩	RLX0	
n`	ɳ	RN	ba**rn**
n`=	ɳ̩	RNX0	
s`	ʂ	RS	pe**rs**
t`	ʈ	RT	sto**rt**
s	s	S	**s**il
S	ʃ	SJ	**sj**u
t	t	T	**t**id
u0	ʉ	UH0	r**u**ss
}:	ʉː	UU0	h**u**s

X-SAMPA	IPA	NoFAbet	Example
v	ʋ	V	**v**ase
w	w	W	**W**ashington
Y	y	YH0	n**y**tt
y:	yː	YY0	n**y**

Unstressed syllables are marked with a 0 after the vowel or syllabic consonant. The nucleus is marked with a *1* for tone 1 and a *2* for tone 2. Secondary stress is marked with *3*.

License

These models are shared with a Creative_Commons-ZERO (CC-ZERO) license, and so are the lexica they are trained on. The models can be used for any purpose, as long as it is compliant with Phonetisaurus' license.