README.md

Grapheme-to-Phoneme models for Norwegian

Version

20200601: 1.0

Introduction

This resource contains Grapheme-to-Phoneme (G2P) models for Norwegian, to be used with the G2P engine Phonetisaurus. The G2P models can be used to generate pronunciation lexica from word lists. For more information on how to do that, consult the Phonetisaurus repo.

The models are trained on the Norwegian pronunciation lexicon for ASR, originally made by the defunct company Nordisk språkteknologi (NST), currently distributed by the National Library of Norway.

Two models have been developed. One is trained on a full version of the lexicon, including phones, marking of primary and secondary stress, and tone. The other is trained on a simplified version where tonal markings and markings of secondary stress are removed.

Content

- *train/*: contains the models, as well as auxiliary files used by Phonetisaurus
  - *model-wtone-nob.fst* contains full tone and stress specifications
  - *model-notone-nob.fst* lacks tone and secondary stress
- *lexica/*: contains various lexica used for training and testing
  - *NST-total_train.dict* is the training set for *model-wtone-nob.fst*. It contains 612 366 word-transcription pairs (WTP) and constitutes 90% of the unique WTPs in the NST lexicon.
  - *NST-total_test.dict* is the test set for *model-wtone-nob.fst*. It consists of the remaining 10% of the unique WTPs in the NST lexicon, which have been randomly selected
  - *NST-total-notone_nosecstress_train.dict* is the training set for *model-notone-nob.fst*. It is equal to *NST-total_train.dict*, but markings of tone and secondary stress have been removed
  - *NST-total-notone_nosecstress_test.dict* is the test set for *model-notone-nob.fst*. It is equal to *NST-total_test.dict*, but markings of tone and secondary stress have been removed
  - *NST-total_test_predicted.dict* is the test set with tones and secondary stress with transcriptions predicted by the G2P system
  - *NST-total_test_notone_predicted.dict* is the test set without tones and secondary stress with transcriptions predicted by the G2P system *g2p_stats.py* is the evaluation script used in this project.

Transcription standard

Although the original NST lexicon uses X-SAMPA as a transcription standard, an equivalent standard is used in this project, which is easier to read by humans, *NoFAbet*. 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, soon to be released in Språkbankens resource catalogue.

X-SAMPA-NoFAbet equivalence table

X-SAMPA	NoFAbet	Example
A:	AA0	b**a**d
{:	AE0	v**æ**r
{	AEH0	v**æ**rt
{*I	AEJ0	s**ei**

X-SAMPA	NoFAbet	Example
E*u0	AEW0	s**au**
A	AH0	h**a**tt
A*I	AJ0	k**ai**
@	AX0	b**e**hage
b	B	**b**il
d	D	**d**ag
e:	EE0	l**e**k
E	EH0	p**e**nn
f	F	**f**in
g	G	**g**ul
h	H	**h**es
I	IH0	s**i**tt
i:	II0	v**i**n
j	J	**j**a
k	K	**k**ost
C	KJ	**k**ino
l	L	**l**and
l=	LX0	
m	M	**m**an
m=	MX0	
n	N	**n**ord
N	NG	e**ng**
n=	NX0	
o:	OA0	r**å**
O	OAH0	g**å**tt
2:	OE0	l**ø**k
9	OEH0	h**ø**st
9*Y	OEJ0	k**øy**e
U	OH0	f*o**r**t
O*Y	OJ0	konv**oy**
u:	OO0	b**o**d
@U	OU0	sh**ow**
p	P	**p**il
r	R	**r**ose
d`	RD	reko**rd**
l`	RL	pe**rl**e
l`=	RLX0	

X-SAMPA	NoFAbet	Example
n`	RN	ba**rn**
n`=	RNX0	
s`	SJ	pe**rs**
t`	RT	sto**rt**
r=	RX0	
s	S	**s**il
S	SJ	**sj**u
s=	SX0	
t	T	**t**id
u0	UH0	r**u**ss
u0 j	UH0_J	Anh**ui**
}:	UU0	h**u**s
v	V	**v**ase
w	W	**W**ashington
Y	YH0	n**y**tt
y:	YY0	n**y**

Unstressed syllables are marked with a 0 after the vowel or consonant syllable nucleus. The nucleus is marked with a *1* for tone 1 and a *2* for tone 2. Secondary stress is marked with *3*. In the material without tone and stress marking, all *3*s are replaced by zeros and all *2*s with *1*s.

For compatibility with NoFA, retroflex *s* is rendered as *SJ* instead of *RS*, which means that there is no distinction between postalveolar and retroflex *s* in the transcriptions.

Evaluation

Model	Word Error Rate	Phoneme Error Rate
model-wtone-nob.fst	14.29	2.76
model-notone-nob.fst	10.44	2.00

The PER calculation is borrowed from this tutorial.

Usage

The models created in this project are public property with the license CC0. See also Phonetisaurus' license.