Creating a Basic Language Resource Kit for Faroese

Annika Simonsen, TALUTØKNI, Faroe Islands
Sandra Saxov Lamhauge, Iben Nyholm Debess, University of the Faroe Islands, Faroe Islands
Peter Juel Henrichsen, Dansk Sprognævn, Denmark

INTRODUCTION

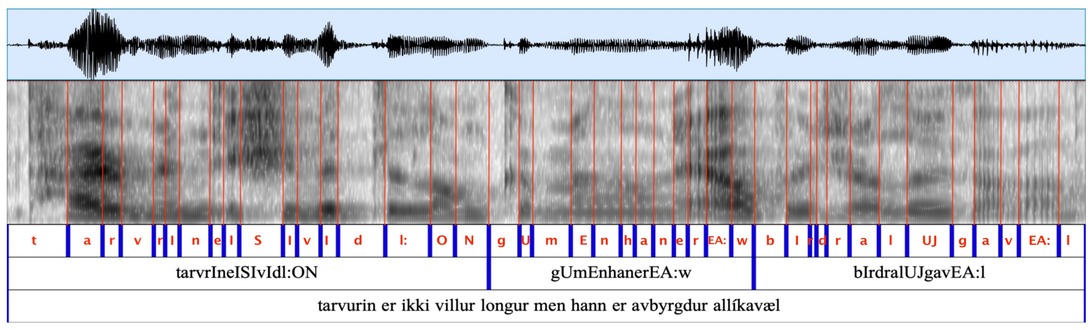
A BLARK for Faroese language technology is under development with formal standards, state-of-the art resources and all-purpose applications.

THE COMPONENTS

SAMPA	IPA compatible; includes stress and length.
Sound	Current size: 150 hours of speech (431 speakers) in WAV-files. 252 female speakers and 170 male speakers.
Transcript Corpus	Orthographic and phonetic transcription, time coded. Current size: Transcribed: 790.000 running words Manually transcribed: 80.000 running words
Reading Material	Word lists, closed vocabularies (e.g. numerals), phrase lists (eliciting intonation patterns, etc.) Short texts, spontaneous speech (monologue) Program that displays a variety of sentences to the reader
Dictionary	Includes pronunciation, PoS, and frequency. Current size: 23,000 complete paradigms. Aiming at 25,000 complete paradigms.
PoS	PAROLE compatible; full morphology.
Background corpus	Text and speech. Current size: 25M words.
Tools	The text and speech tools developed in the project will be available.

ORTO:sekkur	PPOS:NCMSN==IU	PHON:s%EHgUr
ORTO:sekk	PPOS:NCMSA==IU	PHON:s%EHg
ORTO:sekki	PPOS:NCMSD==IU	PHON:s%EHd:ZI
ORTO:sekkjar	PPOS:NCMSG==IO	PHON:s%EHd:Zar
ORTO:sekkurin	PPOS:NCMSN==DU	PHON:s%EHgUrIn
ORTO:sekkin	PPOS:NCMSA==DU	PHON:s%EHd:ZIn
ORTO:sekkinum	PPOS:NCMSD==DU	PHON:s%EHd:ZInUn
ORTO:sekkjarins	PPOS:NCMSG==DO	PHON:s%EHd:ZarIns
ORTO:sekkir	PPOS:NCMP[AN]==IU	PHON:s%EHd:ZIr
ORTO:sekkjum	PPOS:NCMPD==IU	PHON:s%EHd:ZUn
ORTO:sekkja	PPOS:NCMPG==IO	PHON:s%EHd:Za
ORTO:sekkirnir	PPOS:NCMPN==DU	PHON:s%EHd:ZIrnIr
ORTO:sekkirnar	PPOS:NCMPA==DU	PHON:s%EHd:ZIrnar
ORTO:sekkjunum	PPOS:NCMPD==DU	PHON:s%EHd:ZUnUn
ORTO:sekkjanna	PPOS:NCMPG==DO	PHON:s%EHd:Zana

Excerpt from dictionary: ORTO, PPOS, PHON



Excerpt from manual transcription

EA	+	ɛa	spakur	sb%EA:gUr
OA	+	ɔa	vátur	v%OA:dUr
UJ	-/+	ʊi	hvítur, hvítt	kv%UJ:dUr, kv%UJHd:
EJ	+	ɛi	deyður	d%EJ:jUr
aJ	-/+	ai	feitur, feitt	f%aJ:dUr, f%aJHt:
aW	-/+	au	Havn, august	h%aW:n, aWg%Usd
OJ	-/+	ɔi	gloyma, gloymdi	gl%OJ:ma, gl%OJmdI
OW	+	ɔu	tómur	t%OW:mUr
3W	+	ʉu	kúla	k%3W:la
EW	-/+	eu	nevnd, nevna	n%EWnd, n%EW:na
9W	-/+	œu	svøvntur, nøvn	sv%9WxdUr, n%9W:n
9J	-/+	œi	leikutoy, floyal, skoyta	l%aJ:gUt9J, fl%9J:al, sg %9J:da

Excerpt from SAMPA alphabet

The Faroese LT Toolbox

MakeWdList

Creates phonetically complete (random) lists of words that collectively cover all the SAMPA-phones (based on their lexicalized pronunciations)

EvalPhonetics

This tool compares the phonetic forms appearing in a transcription (of a reading session) to the corresponding phonetic forms in the dictionary.

ScrambleText

Makes random scramblings of sentences. Scrambled texts are useful as reading materials for voice recordings (e.g. for avoiding effects of priming and monotony).

MakeLemma

Expands a single wordform (inserted by the user along with PoS-value and phonetic form) into a fully-fledged lemma derived from the existing Dictionary by analogous reasoning.

PushPrompt

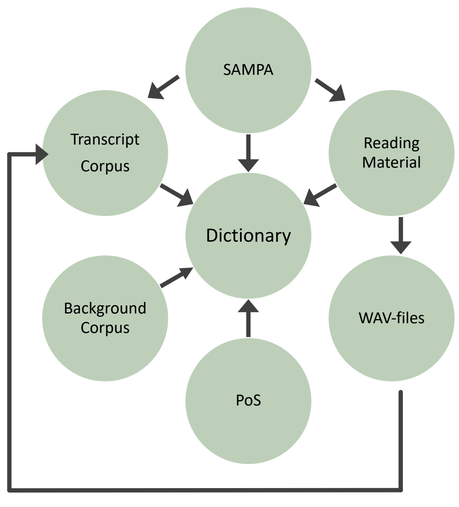
Is used for reading sessions (voice recordings). When the reading session is completed, a log file (with time stamps for each production) is written as a data table compliant with the TextGrid-format.

EvalBlark

Takes BLARK resources as and returns a list of formal inconsistences annotated for location, frequency, kind and degree of inconsistency, and level of significance.

CONSISTENCY PRINCIPLE

The individual resource components work as an eco-system with the resources depending on, feeding off and growing from each other.



EVERYTHING DOCUMENTED

All the resources of the BLARK are documented in both Faroese and English for future work.

OPEN SOURCE, CODE AND FORMAT

All resources are to be freely available, and thereby we will be solving the matter of copyright infringement and GDPR during production. Only non-proprietary file formats are used

Literature

Iben Nyholm Debess, Sandra Saxov Lamhauge and Peter Juel Henrichsen. 2019. Garnishing a phonetic dictionary for ASR intake. In *Proceedings of the 22nd Nordic Conference of Computational Linguistics, NODALIDA 2019.*

Acknowledgments

We wish to personally thank Karin Kass for her never-failing entrepreneurship and diligence. We also wish to thank a number of investors from the Faroese society.

Further information

See https://www.maltokni.fo to access the Faroese BLARK (due in june 2022) and LREC 2022 paper for details about the Faroese BLARK.