

Two firsts for Faroese Speech Recognition and Standard Orthography

Iben Nyholm Debess & Annika Simonsen, TALUTØKNI, the Faroe Islands
Peter Juel Henriksen, Dansk Sprognævn, Denmark

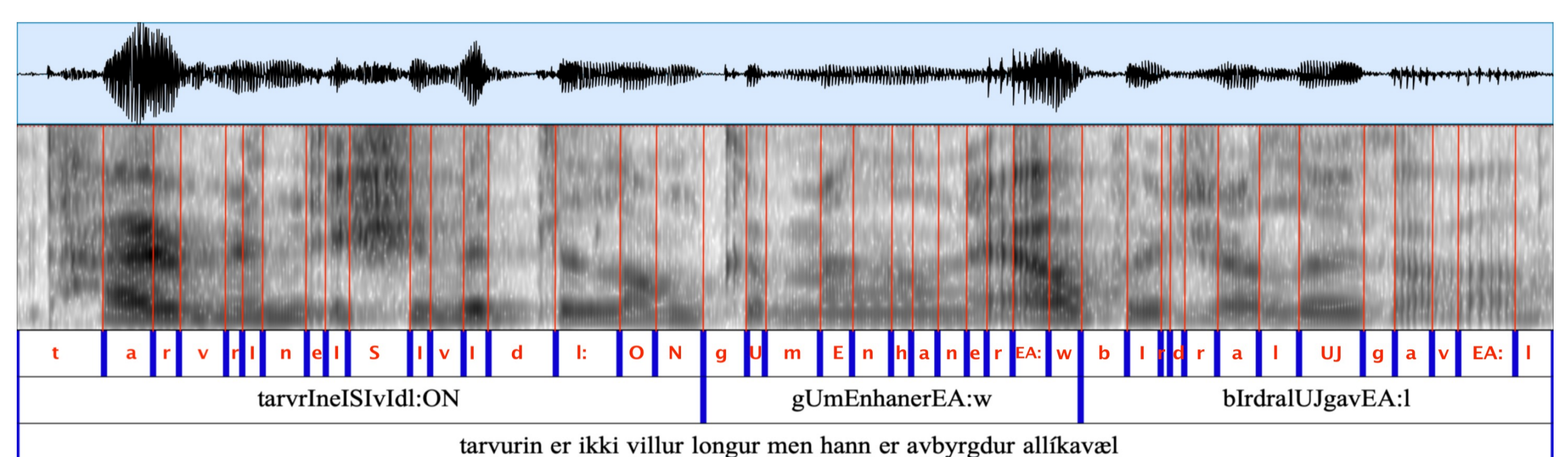
INTRODUCTION

A new ecosystem for Faroese language technology is under development: formal standards, state-of-the-art resources, all-purpose applications

THE COMPONENTS					
SAMPA	IPA compatible; includes stress and length.				
PoS	PAROLE compatible; full morphology.				
Dictionary	Includes pronunciation, PoS, and frequency. Current size: 19,500 complete paradigms. Aiming at 25,000 complete paradigms.				
Background Corpus	Text and speech. Current size: 23M 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				
Transcript Corpus	Orthographic and phonetic transcription, time coded. Current size: <table border="1"> <tr> <td>Transcribed</td> <td>420.000 running words</td> </tr> <tr> <td>Manually transcribed</td> <td>80.000 running words</td> </tr> </table>	Transcribed	420.000 running words	Manually transcribed	80.000 running words
Transcribed	420.000 running words				
Manually transcribed	80.000 running words				
Sound	Current size: 127 hours of speech (354 speakers) in WAV-files. Aiming at 150 hours (200 male and 200 female speakers).				
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:Zl
ORTO:sekkjar	PPOS:NCMSG==IO	PHON:s%EHd:Zar
ORTO:sekkjarin	PPOS:NCMSN==DU	PHON:s%EHgUrn
ORTO:sekkjarin	PPOS:NCMSA==DU	PHON:s%EHd:Zln
ORTO:sekkjarinum	PPOS:NCMSD==DU	PHON:s%EHd:ZlnUn
ORTO:sekkjarins	PPOS:NCMSG==DO	PHON:s%EHd:Zarlns
ORTO:sekkir	PPOS:NCMP[AN]==IU	PHON:s%EHd:Zlr
ORTO:sekkjum	PPOS:NCMPD==IU	PHON:s%EHd:Zun
ORTO:sekkja	PPOS:NCMPG==IO	PHON:s%EHd:Za
ORTO:sekkimir	PPOS:NCMPN==DU	PHON:s%EHd:Zlrn
ORTO:sekkimar	PPOS:NCMPA==DU	PHON:s%EHd:Zlrmar
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

1633092143.688 OK	42 Segði tú nakað	[blset]
1633092146.335 OK	43 Hann fekk eitt gott hugskot	[blset]
1633092149.411 OK	44 Ber svarið hjá tær saman við tað hjá Helgu	[blset]
1633092154.328 OK	45 Ert tú giftur	[blset]
1633092156.795 OK	46 Høvdu tit kunnað stokt kjøtið eitt sindur meira	[blset]
1633092160.962 OK	47 Kann eg læna súkkkluna hjá tær í nakrar dagar	[blset]
1633092165.007 OK	48 Kann eg ringja aftur	[blset]
1633092167.546 OK	49 Ári er tvey ár eldri enn eg	[blset]
1633092170.623 OK	50 Grenj broytir einki	[blset]
1633092173.789 OK	51 Kann tað gerast bligari	[blset]
1633092176.955 fail	52 Minnst tú enn tá vit møttust á fyrsta sinni	[blset]
1633092180.836 fail	52 Minnst tú enn tá vit møttust á fyrsta sinni	[blset]
1633092186.189 OK	52 Minnst tú enn tá vit møttust á fyrsta sinni	[blset]
1633092192.454 OK	53 Kann eg leggja eini boð eftir meg	[blset]
1633092195.802 OK	54 Hon veit sikkurt ikki hvar eg eri	[blset]

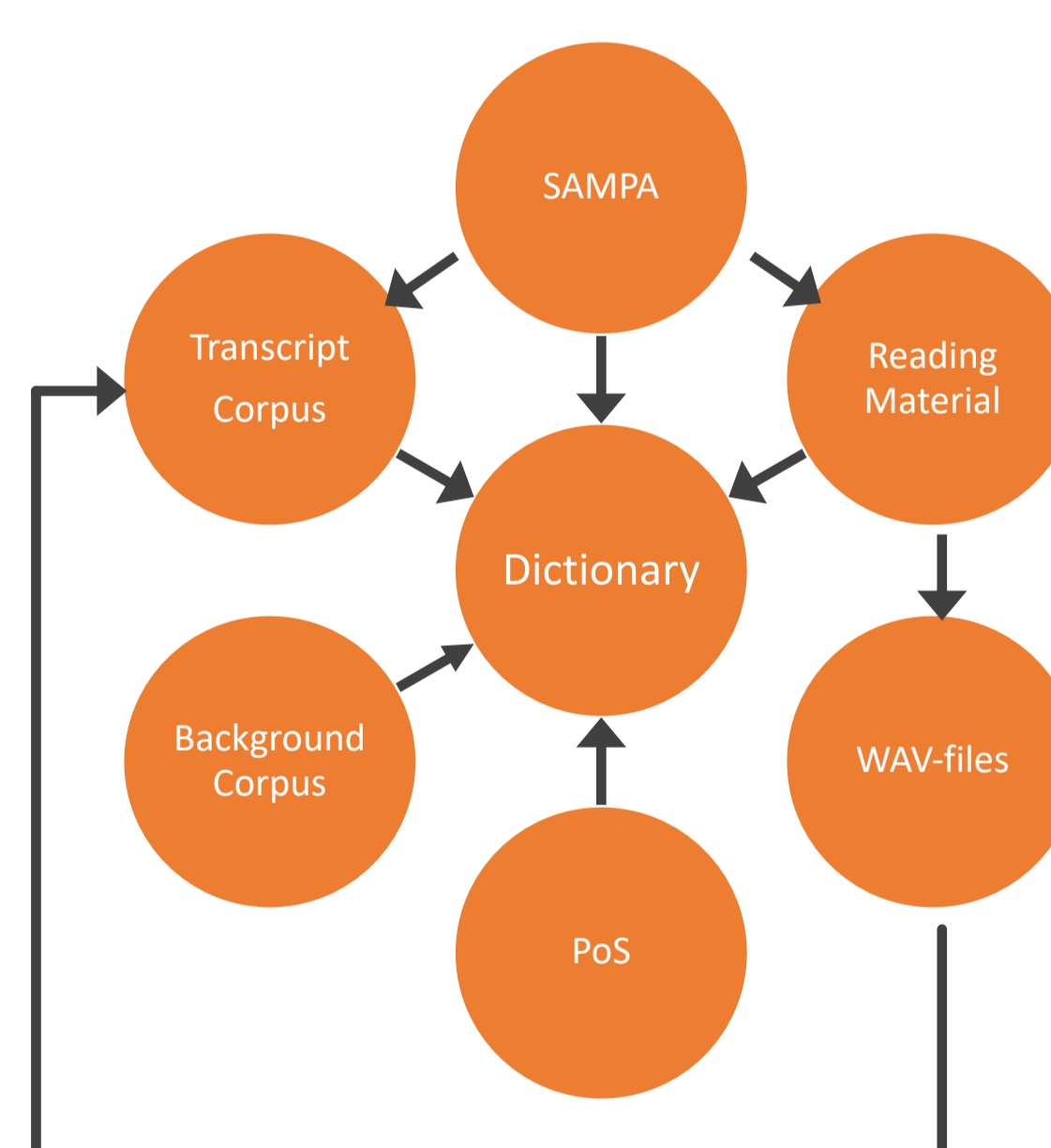
Excerpt from transcription

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

EVERYTHING DOCUMENTED

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



CONSISTENCY PRINCIPLE

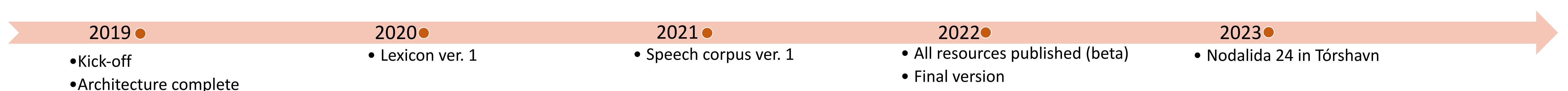
The individual resource components feed off and grow from each other in an iterative process.

OTAL

The OTAL project or "Orðaskráin" (The Word Register) is a GUID-style index system used to cross-reference all the linguistic resources in the current ASR project. The index system is developed bearing in mind working towards implementing the system into as many Faroese language resources as possible, most notably the new orthographic dictionary, when it has been completed. OTAL takes its main inspiration from the ongoing Danish project COR or Det Centrale Ordregister (the Central Danish Word Register). The name "OTAL" refers to the Faroese civil registration number, "P-tal".

EXAMPLES

Word form	OTAL	Word form	OTAL
tráður	OTAL.RAVN.285043.02231	bátur	OTAL.RAVN.207686.02231
tráð	OTAL.RAVN.285043.02225	bát	OTAL.RAVN.207686.02225
tráð	OTAL.RAVN.285043.02227	báti	OTAL.RAVN.207686.02227
tráði	OTAL.RAVN.285043.12227		
træðri	OTAL.RAVN.285043.22227		
tráðar	OTAL.RAVN.285043.02229	báts	OTAL.RAVN.207686.02229



Literature

Iben Nyholm Debess, Sandra Saxov Lamhauge and Peter Juel Henriksen. 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.