Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4637–4643
Marseille, 20-25 June 2022

Creating a Basic Language Resource Kit for Faroese

Annika Simonsen[1], Sandra Saxov Lamhauge[2], Iben Nyholm Debess[2], Peter Juel Henrichsen[3]
[1]Grunnurin Talutøkni, [2]University of the Faroe Islands, [3]Dansk Sprognævn
[1]Íslandsvegur 10A, 100 Tórshavn, Faroe Islands [2]V. U. Hammershaimbs gøta 22, 100 Tórshavn, Faroe Islands,
[3]Adelgade 119B, 5400 Bogense, Denmark
annika.simonsen@hotmail.com, sandrasl@setur.fo, ibennd@setur.fo, pjh@dsn.dk

Abstract

The biggest challenges we face in developing LR and LT for Faroese is the lack of existing resources. A few resources already exist for Faroese, but many of them are either of insufficient size and quality or are not easily accessible. Therefore, the Faroese ASR project, Ravnur, set out to make a BLARK for Faroese. The BLARK is still in the making, but many of its resources have already been produced or collected. The LR status is framed by mentioning existing LR of relevant size and quality. The specific components of the BLARK are presented as well as the working principles behind the BLARK. The BLARK will be a pillar in Faroese LR, being relatively substantial in both size, quality, and diversity. It will be open-source, inviting other small languages to use it as an inspiration to create their own BLARK. We comment on the faulty yet sprouting LT situation in the Faroe Islands. The LR and LT challenges are not solved with just a BLARK. Some initiatives are therefore proposed to better the prospects of Faroese LT. The open-source principle of the project should facilitate further development.

Keywords: low-resource language, BLARK, Faroese

1. Introduction

This article presents the Basic Language Resource Kit (BLARK) for Faroese, collected by the authors in the course of the ongoing Faroese ASR project, Ravnur. The project was founded in January 2019 by a number of public and private initiators and investors, including the Faroese government. The project is expected to be finished by summer of 2022. The main purpose of the project is to create all necessary constituents for developing high quality Automatic Speech Recognition (ASR) for Faroese. The biggest challenge of making ASR (as well as Text-to-Speech (TTS) and other language technology) for small languages is the lack of language resources (Nikulasdóttir et al., 2018), which makes the development of such resources a vital part of the project.
Section 2 introduces the existing language technology (LT) for Faroese. In section 3, we go into details about our BLARK, including a presentation of the toolbox created as part of the ASR project. Section 4 concludes with perspectives and prospects for the future.

2. Existing LT for Faroese

There are estimated to be 75.000 native speakers of Faroese (faroeislands.fo, n.d.). Like other low-resource languages (for Icelandic, see Nikulásdóttir et al., 2020), Faroese is rarely supported in LT software and applications from international suppliers due to its market size. However, several institutions have taken the initiative to create smaller language resources for their own purposes, but these resources are not open-source and not easily accessible, if at all. In this section, we will give a brief overview of language resources for Faroese, focusing on LT tools and corpora of spoken Faroese. This list will not be exhaustive but will mention the language resources that are of substantial size and have been relevant to us during our process.

2.1 LT tools

Faroese LT tools are sparse in general; there is a grammar and spell checker and speech synthesis. UiT The Arctic University of Norway and The University of the Faroe Islands have collaborated to create the grammar and spell checker. The beta was released in 2019 (Fróðskaparsetur Føroya, 2020; Fróðskaparsetur Føroya, n.d.; Trosterud, Jákupsson and Moshagen, 2021). The speech synthesis was created in 2005 (Helgason, Gullbein and Kass, 2005) by combining efforts from researchers at the University of Stockholm and the University of Uppsala and the University of the Faroe Islands. The speech synthesis and the right to its resources are currently tied in a software company abroad.

2.2 Corpora of spoken Faroese

There are four substantial corpora of spoken Faroese that we are familiar with: the Nordic Dialect Corpus, FADAC Hamburg, Talumálsbankin ('Corpus of spoken Faroese') and the Nordic Word order Database.

In 2008, researchers from the University of Oslo held a dialect workshop in the Faroe Islands, during which recordings were made of three Faroese informants. These recordings became part of **the Nordic Dialect Corpus** (Johannesen et al., 2009). Currently, there are recordings of 20 native speakers of Faroese in the corpus. The corpus is manually transcribed and is hosted at the Text Laboratory in Oslo, which is part of CLARIN[1].

FADAC Hamburg is a corpus of spoken Faroese and Faroe-Danish (Debess, 2019). The corpus contains 469,000 manually transcribed words from 92 informal interviews with 56 speakers belonging to different generations and dialects. The speakers are native speakers of Faroese and speak Danish as a second language. The University of Hamburg is hosting the corpus.

The Faculty of Faroese Language and Literature at the University of the Faroe Islands began creating a **Corpus of Spoken Faroese** in 2015 (Føroyskur talumálsbanki). The corpus is manually transcribed, and at the time of writing, there are 471,178 tokens (incl. punctuation). The

[1] CLARIN : *Common Language Resources and Technology Infrastructure*, a European research infrastructure working with digital language data for research in humanities and social sciences.

University of Bergen is hosting the corpus of spoken Faroese on their corpus management system, Corpuscle, and is a part of CLARIN.

The Nordic Word order Database contains elicited production data from speakers of all of the North Germanic languages, including Faroese (Lundquist et at., 2019). The recordings of the Faroese informants were made during a field trip by researchers from the University of Oslo and UiT The Arctic university of Norway in 2018. In total, 58 native speakers of Faroese participated in the elicited production experiment. The database is hosted by the Text Laboratory in Oslo.

2.3 Other LT tools and materials

Other language materials and tools have been developed for other purposes, such as language analysing tools (Hafsteinsson, 2020; Trosterud and Jákupsson, 2012) and various text corpora and treebanks (Hansen, 2014; Sigurðsson et al., 2012; Tyers et al., 2018; FTS, GiellaLT korpus for færøysk), but the efforts have not been coordinated, and the resources are therefore scattered. Many of these resources are insufficient in size and/or quality for an ASR task, and some are not made freely available. The Ravnur project therefore set out to form a linguistic database from scratch that would be available open-source and could be used to develop LT tools (e.g. ASR) and for research in general. This came with the advantages of establishing rational and explicit principles for all aspects of data collection, annotation, and processing (Debess, Saxov and Henrichsen, 2019).

3. Faroese BLARK

In developing ASR for Faroese, we had to identify the language resources already available for Faroese and language resources still needing to be developed. Therefore, we decided to develop a BLARK (Basic Language Resource Kit) for Faroese. A BLARK is determined as the minimal set of language resources needed to develop language and speech technology in a specific language (Krauwer, 2003; Maegaard et al., 2006). Although the definition is meant to be language independent, the components of the BLARK might differ in different languages, as languages are different (Krauwer, 2003).

Here, we list the components of the Faroese BLARK, developed as part of the ASR project, but intended to be of use in other research and LT areas as well. As recommended by Krauwer (2003), and in line with the purpose of the ASR project, the BLARK will be open-source and freely accessible, inviting other researchers and developers to use the resources[2]. The BLARK can act as an inspiration as well for other small languages who wish to create their own BLARK.

3.1 SAMPA

Inspired by the SAMPA initiative, the Faroese SAMPA is a computer-readable phonetic inventory developed by the authors. It is IPA-compatible and includes the most common and distinctive phones and diacritics for Faroese (Adams and Petersen, 2014; Árnason, 2011; Petersen, 2021; Rischel, 1961; Thráinsson et al., 2012). Table 1 presents a part of the Faroese SAMPA, the monophthongs. Although the original SAMPA project is closed and no longer updated, we put forward our suggestion of a Faroese SAMPA, as suggested by Wells (p.c.), the founder of the SAMPA initiative. In line with the purpose of the BLARK, the SAMPA will be made freely available for other projects to use.

SAMPA	Length	IPA	Example	Example SAMPA
i	-/+	i	linur, hugnaligur	l%i:nUr, h%u:nalijUr
I	-	ɪ	lint, vunnið	l%Ixd, v%Un:I
e	-/+	e	Elisa, frekur	el%i:sa, fr%e:gUr, gl %e:a
E	-	ɛ	frekt	fr%EHgd
a	-/+	a	japanskur	jap%a:nsgUr
y	+	y	mytiskt	m%y:dIsd
Y	-	ʏ	mystiskt	m%YsdIsd
2	+	ø	høgur	h%2:vUr
9	-	œ	høgt	h%9Hgd
u	-/+	u	gulur, nuansa	g%u:lUr, nu%ansa
U	-	ʊ	gult, vinnur	g%ULd, v%In:Ur
o	-/+	o	tola, apotek	t%o:la, abot%e:g, f%o:a
O	-	ɔ	toldi	t%OldI

Table 1: A part of the Faroese SAMPA inventory. Here, the monophthongs inventory is presented.

3.2 Part-of-Speech (PoS)

The PoS-tagset for Faroese is inspired by and compatible with the Pan-European PAROLE meta-tagset (Bilgram and Keson, 1998). Our tagset is heavily inspired by the Danish PAROLE meta-tagset (Bilgram and Keson, 1998; Keson, 1998), but as the Faroese language differs substantially from the Danish language, considerable changes and modifications have been made in the Faroese PAROLE PoS-tagset. When the project started, we were not aware of any PoS-tagger for Faroese, and we knew of little to no PoS-tagged Faroese text. Creating a tagger of our own and tagging text has not been a part of our project to date; however, the decision to create a detailed Faroese PoS-tagset and adding these PoS-tags to the dictionary was made with that in mind that it would open up the door for others to use our tagset to create a Faroese tagger - or to map these tags onto other well-known tagsets, such as the Universal Dependencies PoS-tags if needed. In our project, we did not make use of the treebank annotations, but an overview over the UD treebanks and their token count is below as a footnote[3].

3.3 Dictionary

At present, the dictionary holds about 21,700 entries. As a minimum, it includes all words from our transcription

[2] By June 2022 the BLARK 1.0 for Faroese LT will be accessible through www.maltokni.fo

[3] The treebank created by Tyers et al (2018) as part of a Universal Dependencies project contains 10.002 tokens. GiellaLT korpus for færøysk is also a treebank using dependencies trees; however, is still in the making. Currently it contains 10.56 mio tokens. Links to these resources can be found in our Language Resource Reference list.

corpus and the most frequent lemmas beyond that. While traditional Faroese dictionaries only include a subset of inflectional forms, each entry in our dictionary includes all forms associated with that entry. Each form is given a PAROLE PoS-tag, a pronunciation written in SAMPA, and frequency information. The dictionary was manually created in a partly automated process (read about the MakeLemma tool in 3.11.4). All automated output was manually checked for quality. Table 2 presents an example of an entry from the dictionary.

ORTO:hundur	PPOS:NCMSN==IUU	PHON:h%UndUr
ORTO:hund	PPOS:NCMSA==IUU	PHON:h%Und
ORTO:hundi	PPOS:NCMSD==IUU	PHON:h%Undl
ORTO:hunds	PPOS:NCMSG==IOU	PHON:h%Unds
ORTO:hundurin	PPOS:NCMSN==DUU	PHON:h%UndUrIn
ORTO:hundin	PPOS:NCMSA==DUU	PHON:h%UndIn
ORTO:hundinum	PPOS:NCMSD==DUU	PHON:h%UndInUn
ORTO:hundsins	PPOS:NCMSG==DOU	PHON:h%UndsIns
ORTO:hundar	PPOS:NCMP[AN]==IUU	PHON:h%Undar
ORTO:hundum	PPOS:NCMPD==IUU	PHON:h%UndUn
ORTO:hunda	PPOS:NCMPG==IOU	PHON:h%Unda
ORTO:hundarnir	PPOS:NCMPN==DUU	PHON:h%UndarnIr
ORTO:hundarnar	PPOS:NCMPA==DUU	PHON:h%Undarnar
ORTO:hundunum	PPOS:NCMPD==DUU	PHON:h%UndUnUn
ORTO:hundanna	PPOS:NCMPG==DOU	PHON:h%Undana

Table 2: An example of a lemma from the dictionary. Here, the noun *hundur* 'dog' is presented with its PoS-tags and phonetic transcription. As an example, the tag NCMSN==IUU is explained: N = noun, C = common, M = masculine, S = singular, N = nominative, I = indefinite, U = unmarked

Currently, no official orthographic dictionary has been published in the Faroe Islands, but the Orthographic Dictionary is forthcoming. Project Ravnur has collaborated with the editorial staff of the Orthographic Dictionary regarding spelling.

3.4 Background Corpus

The background corpus consists of a text corpus and a speech corpus. The text corpus currently contains 23M words and consists of both formal and informal styles. The speech corpus includes interviews from UiO (Johannesen, 2009; Johannesen et al., 2009) and audiobooks. The speech corpus, however, is not open-source, but part of the text corpus will be made freely accessible.

3.5 Reading Material

The reading material consists both of texts written by us and of texts collected and modified by us to be used in the ASR project. The texts have been read aloud and recorded. In the first part of the ASR project, the reading material consisted of a word list, a phrase list, a closed vocabulary reading (e.g. numerals), two short texts and spontaneous speech. All the texts were constructed in a way so as to elicit all possible phones in Faroese. In the later part of the project, we have deviated from this set up, and the focus has been to get more volume and many kinds of text genres.

3.6 Sound

Each speaker session produces sound recordings in WAV-format of the reading material and lasts for about 20 min. At the moment, we have collected 135 hours of speech (414 speakers). The speakers come from all major dialect areas in the Faroe Islands (Thráinsson et al., 2012). All ages and genders are covered.

3.7 Transcript Corpus

All recordings are orthographically transcribed and time-coded according to the conventions developed for the ASR project. Some of the recordings have been phonetically transcribed by trained phoneticians. The transcript corpus currently holds 450,000 automatically transcribed running words and 80,000 manually transcribed running words.

3.8 Documentation

All the work done in course of the ASR project and all of the resources developed have been thoroughly documented in both Faroese and English. Furthermore, manuals have been written about the PAROLE PoS-tagset, the orthographic and phonetic transcription convention, the recording sessions, and the dictionary.

3.9 Consistency Principle

All language resources in our BLARK work as an eco-system with the resources depending on, feeding off and growing from each other. For example, all words occurring in the reading texts and transcript corpus must also be included in the dictionary, and all words in the dictionary must be phonetically transcribed with phones from our SAMPA and tagged with PoS-tags from our PAROLE tag-system, see figure 1.

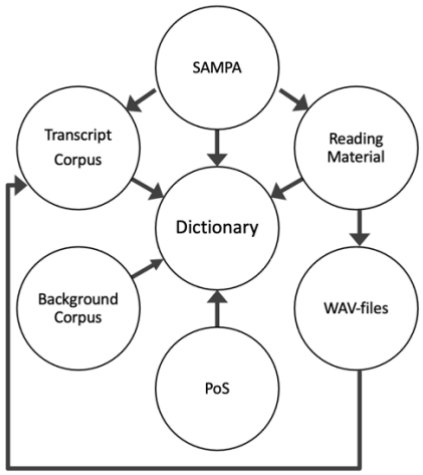

Figure 1: Inter-dependency of BLARK-components.

3.10 Quality Assessment

To ensure the quality of our BLARK resources, we have created a system for quality assessment. Our development team consists of a project leader, a technical leader (senior researcher, data linguist and author of this paper), three native speaking junior linguists (authors of this paper), an IT assistant, five student assistants, as well as external advisors. Every resource created by the student assistants is quality assessed by the linguists. The technical quality of the resources created by the linguists is assessed by the IT assistant, while the entire project is overseen by the project leader. The evaluation tool EvalBlark (cf. 3.10.6) serves as a continuous assessment of technical quality and consistency. General questions about language use are brought to advisers from the Faroese Language Council (Málráðið) and professors and linguists from the

Department of Language and Literature at the University of the Faroe Islands.

3.11 The Ravnur Toolbox

As explained above, the BLARK resources are interoperable by design. This formal property, convenient for future LT projects, was obtained by compiling and proofreading the various databases using a toolbox of dedicated programs.

Here, we present some of the basic tools in the BLARK toolbox. All programs can be run from a Linux prompt (for use on a pc) or as cgi-applications (for browser-based use). Most of the code is written in perl (ver. 5 or later).

3.11.1 MakeWdList

The MakeWdList script is used for creating (random) lists of words, phonetically complete in the sense that the words collectively cover all the SAMPA-phones (based on their lexicalized pronunciations). Such word lists are convenient as reading materials for speech technology (speech recordings for TTS and ASR training data).

Input: Dictionary; SAMPA definition table.

Output: A phonetically complete wordlist.

3.11.2 ScrambleText

ScrambleText makes random scramblings of sentences. The scrambling procedure can be controlled and restricted in various ways. Scrambled texts are useful as reading materials for voice recordings (e.g. for avoiding effects of priming and monotony).

Input: Dictionary; a text of words occurring in Dictionary

Output: The input text scrambled (randomly or otherwise)

3.11.3 EvalPhonetics

This tool compares the phonetic forms appearing in a transcription (of a reading session) to the corresponding phonetic forms in the dictionary. The discrepancies (i.e. pronunciations deviating from the lexicalized forms) are compiled as data tables and written out in csv-form. The tool has several uses, an important one being verification of 'phonetical completeness' in voice recordings (cf. 3.11.1 above).

Input: TextGrid (transcription); Dictionary.

Output: Pronunciations deviating from lexical form.

3.11.4 MakeLemma

MakeLemma is used for creating new lexical entries. As the BLARK Dictionary is a full-form lexicon (each lemma represented by all its inflected forms), new entries can be cumbersome to type in. 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. MakeLemma may output several alternative suggestions for lemmas (sorted by likelihood). This process ends with a manual quality check of the lemma before entering the lexicon.

Input: A word form (representing a lemma *L*); Dictionary.

Output: Fully developed lemmas (suggestions for *L*)

3.11.5 PushPrompt

PushPrompt is used for reading sessions (voice recordings). PushPromt presents the text items in the reading material to the reader, allowing her to manage the session interactively (adjusting her reading tempo, repeating speech productions at wish, inserting short breaks as needed, etc.). 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.

Input: Reading manuscript; Dictionary.

Output: Data table (time codes etc.); session log.

3.11.6 EvalBlark

This tool is probably the most important of them all. In order to qualify as an "*A-BLARK*" (a formally approved Ravnur repository of language resources), a BLARK must be technically consistent by proof (cf. section 3.9 "Consistency Principle"). Only approved A-BLARKs are shared publicly and are assigned a (single-digit) version number for future reference.

These are some of the defining requirements for A-BLARKs:

- All letter symbols occurring in the text based resources must be defined in the Global Alphabet
- All phone symbols used in the transcription corpus, the Dictionary, and elsewhere must be defined in the Global Phone Table
- All orthographic word forms occurring in the transcription corpus (TextGrids) must be represented in the Dictionary
- All PoS-tags appearing in the text based resources must (i) be defined in the Global PoS Table; (ii) conform to the Dictionary's PoS mapping

EvalBlark is the essential tool for establishing A-BLARK-hood.

EvalBlark accepts as input a diverse collection of BLARK resources (corpora, definition tables, dictionaries, texts, transcriptions, and more), returning a list of formal inconsistencies annotated for location, frequency, kind and degree of inconsistency, and level of significance. Where possible, EvalBlark also suggests corrections (optionally).

Input: Any collection of BLARK language resources.

Output: A list of formal inconsistencies.

3.11.7 Wrapping up

The BLARK toolbox is freely available (open-code). At the time of writing, however, no permanent hosting solution has been established, so please contact the authors to obtain a copy of the code and documentation.

The tools are not language specific and can be used without changes for BLARK compilation in other languages, as long as the necessary formal definitions are provided (data formats and formatting, text encoding, defining tables of alphabetic letters, phonetic symbols and PoS-labels, etc.). It is our hope that the toolbox may help other 'small' languages establish their own BLARK (a caveat is in place, though: Some of the tools may be less relevant for polysynthetic languages).

4. Perspectives and Prospects

Faroese is advancing in LT, yet it faces challenges. Apart from the obvious challenges implicit in being a small language (i.e. small number of speakers, language preservation, low resources, little to no commercial base for LT), the challenges also stem from political priorities. The lack of central coordination of LT activity together with the scattering and limited access and usability of present language resources are also pressing issues. The existing Faroese LT applications are valuable and ground-breaking, but the environment around the applications is small, with little to no organised activity regarding maintenance, updating and continuous development. The issues with language preservation have been pressing in the Faroe Islands for centuries. Though Faroese being L1 for a large majority of speakers, the integration of other languages into the Faroese society has been a constant threat for decades (historically with Norwegian and especially Danish, and contemporarily also with English) (Andreasen, 2021; Mitchinson, 2012; Petersen, 2010). A solid foundation for language preservation was established when the written code for Faroese was made in 1846. The next appropriate, and inevitable, step in strengthening that foundation would be to integrate Faroese language fully in the technological domain. Faroese language is currently suffering from the loss or ongoing exclusion of several linguistic domains, the technological domain in particular. Appropriate language resources and open-code applications will facilitate Faroese into this crucial domain. Because of the small numbers of speakers, and thereby LT application users, the commercial LT market base is small, partly making the success of LT a state funding issue and thereby a political issue. The development of proper and suitable language resources and LT in the Faroes depends deeply on political will. This is evident from the fact that the ongoing ASR project, which is the most extensive language resource collection for Faroese LT thus far, is primarily state funded. The challenge of existing language resources being decentralised, and the lack of basic maintenance and development, could be solved with a political initiative to establish a coordinative centre for language resources and LT. An inspiration in this regard is also the exemplar LT Programme for Icelandic (Nikulásdóttir et al., 2020).

The aforementioned challenges are not limited to the Faroes. All the Nordic countries face difficulties in the areas of LT and language resources, due to the relatively small number of speakers compared to the rest of Europe and even beyond. On behalf of The Nordic Council of Ministers, an expert group has stated the necessity for an inter-Nordic collaboration in order to increase the activity, quality, and compatibility of language resources in all the Nordic languages, both official and minority languages (Kvarfordt, 2022). This highly relevant initiative mirrors the urgent need for development in the LT area for Nordic languages and introduces three main areas of focus: 1) digital inclusion and language preservation 2) linguistic equality in Europe 3) joining Nordic forces to urge the inclusion of small languages in the technology of the tech-giants. All three points are key areas of development for Nordic and Faroese language resources and LT. We have already touched upon the first point, and the second point fits right into the momentum with the European ELE Project. The third point is an important task, as the economic base of Faroese LT seems to keep the large corporations away. Any positive result from a joint effort to encourage implementation of small language tools into the widely used applications would be revolutionary. Though facing challenges, we also see an emerging LT community for Faroese. The next step is to establish best practices for Faroese language resources, for both LT and research, and secure the resources, tools, and work force in a centralised unit. The open-source principle of the ASR project should improve the landscape for Faroese LT resources and hopefully serve as a foundation on which more resources and tools can grow.

5. Concluding Remarks

In this article, we have described the background and vision of the Faroese ASR project, Ravnur. The core of the project is developing appropriate and sustainable language resources in the body of a BLARK. The components of the BLARK have been presented as well as the ideology of interoperability regarding both format and content. All components will be released open-source, and this will serve as a strong, new foundation for developing LT and more language resources for Faroese. The Faroese LT environment is changing, yet it needs central, government-initiated coordination to organize even basic maintenance tasks. Further development of LT in Faroese is crucial in securing language domains and for the goal of language preservation.

6. Acknowledgments

We would like to thank Karin Kass for her persistent support and motivation. We would also like to express our gratitude for all financial contributions for the ASR project as well as those donating their speech and text samples. Appreciation is also sent towards the secretariate of The Faroese Language Council (Málráðið) and Department of Language and Literature at the University of the Faroe Islands for professional advice throughout the project.

7. Bibliographical References

Adams, J. and Petersen, H. P. (2014). *A Language Course for Beginners*. Stiðin, Tórshavn, Faroe Islands, 3rd edition.

Andreasen, R. S. (2021). *Útlitini fyri einum málskifti: Fara vit at tosa enskt?* Tórshavn: Fróðskapur, Setursrit 12.

Árnason, K. (2011). *The Phonology of Icelandic and Faroese*. Oxford: Oxford University Press.

Bilgram, T. and Keson, B. (1998). The Construction of a Tagged Danish Corpus. In *Proceedings of the 11th*

Nordic Conference of Computational Linguistics, *NODALIDA 1998*, pages 129-139.

Debess, I. N., Lamhauge, S. S., Henrichsen, P. J. (2019). Garnishing a phonetic dictionary for ASR intake. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, *NODALIDA 2019*, pages 395-399.

Faroeislands.fo (n.d.). The Faroese Language. https://www.faroeislands.fo/arts-culture/language/ [Last accessed on 6 January 2022].

Fróðskaparsetur Føroya (2020). Nýggjur føroyskur rættstavari. https://www.setur.fo/fo/setrid/tidindi/nyggjur-foroyskur-raettstavari/ [last accessed on 6 January 2022].

Fróðskaparsetur Føroya (n.d.). Føroyamálsdeildin hevur í samstarvi við universitetið í Tromsø (UiT) ment ein nýggjan føroyskan rættstava. https://www.setur.fo/fo/setrid/almennar-taenastur-og-grunnar/raettstavarin/ [Last accessed on 8 January 2022]

Hafsteinsson, H. (2020). *A Faroese part-of-speech tagger built with Icelandic methods - Data preparation, training and evaluation.* M.A. thesis in language technology at the University of Iceland.

Hansen, K. D. (2004). FTS - Føroyskt TekstaSavn/færøsk talekorpus. In H. Holmboe (ed.). 2005. *Nordisk sprogteknologi 2004 - Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*, pages 47-50.

Helgason, P., Gullbein, S., and Kass, K. (2005). Færøsk talesyntese: Rapport marts 2005. In H. Holmboe (Ed.), *Nordisk sprogteknologi 2005 - Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*, pages 51-58.

Johannesen, J. B. (2009). A corpus of spoken Faroese. *Nordlyd*, 36(2), pages 25-35.

Keson, B. (1998). Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus. DSL.

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of the International Workshop "Speech and Computer", SPECOM 2003*, pages 8-15, Moscow, Russia.

Kvarfordt, K. (2022). *Nordiskt initiativ för språkteknologi - digital inkludering av små språk i Norden*. Uppsala: Inst. for Språk och Folkminnen, ISOF Working Papers

Lundquist, B., Larsson, I., Westendorp, M., Tengesdal, E., and Nøklestad, A. (2019). Nordic Word order Database: motivations, methods, material and infrastructure. *Nordic Atlas of Language Structures (NALS) Journal*, 4(1), pages 1–33.

Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. D. (2006). The BLARK concept and BLARK for Arabic. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 773- 778, Genova, Italy.

Mitchinson, J. (2012). *Danish in the Faroe Islands. A Post-Colonial Perspective*. Ph.d.-dissertation. University College London.

Nikulásdóttir, A. B., Helgadóttir, I. R., Pétursson, M., and Guðnason, J. (2018). Open ASR for Icelandic: Resources and a Baseline System. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3137-3142.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3414–3422.

Petersen, H. P. (2010). *The Dynamics of Faroese-Danish Language Contact*. Heidelberg: Universitätsverlag Winter.

Petersen, H. P. (2021). *Føroysk míllæra 3: Ljóðlæra*. Nám, Tórshavn.

Rischel, J. (1961). Om retskrivningen og udtalen i moderne færøsk. In M. A. Jacobsen & C. Matras (Eds.), *Føroysk-donsk orðabók*. Tórshavn: Føroya Fróðskaparfelag, pp. xiii-xxxvi.

Thráinsson, H., Petersen, H. P., Jacobsen, J. í L., and Hansen, Z. S. (2012). *Faroese, an overview and reference grammar*. Tórshavn: Fróðskapur, Reykjavík: Linguistic Institute, University of Iceland, 2nd edition.

8. Language Resource References

BLARK 1.0 for Faroese LT https://maltokni.fo

Debess, I. N. (2019). *FADAC Hamburg 1.0. Guide to the Faroese Danish Corpus Hamburg. Kieler Arbeiten zur skandinavistischen Linguistik 6. Institut für Skandinavistik, Frisistik und Allgemeine Sprachwissenschaft (ISFAS)*, FID Northern Europe https://macau.uni-kiel.de/receive/macau_publ_00002318

FTS Färöisk textsamling https://spraakbanken.gu.se/resurser/fts

Føroyskur talumálsbanki https://clarino.uib.no/korpuskel/corpus-list

GiellaLT korpus for færøysk. UiT Norges arktiske universitets tekstsamling, Version 01.12.2021. http://gtweb.uit.no/f_korp/?mode=fao

Johannesen, J. B., Priestly, P., Hagen, K., Åfarli, T. A., and Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus - an Advanced Research Tool. In K. Jokinen and E. Bick (Eds.), *NEALT Proceedings Series*, 4:73-80 http://tekstlab.uio.no/nota/scandiasyn/

Sigurðsson, E. F., Ingason, A. K., Rögnvaldsson, E., and Wallenberg, J. C. (2012). Faroese Parsed Historical Corpus (FarPaHC). Version 0.1. http://www.linguist.is/farpahc

Trosterud, T. and Jákupsson, H. (2012). A Finite State Transducer for Faroese. https://github.com/giellalt/lang- fao (2022).

Trosterud, T., Jákupsson, H. and Moshagen, S. N. (2021). Faroese proofing tool v1.1.1. https://github.com/giellalt/lang-fao

Tyers, F. M., Sheyanoa, M., Martynova, A., Stepachev, P., and Vinogradovsky, K. (2018). Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW)*, pages 144-150.

https://github.com/UniversalDependencies/UD_Faroese-OFT