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 Sprogøvnsen, Denmark

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

<table>
<thead>
<tr>
<th>THE COMPONENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SAMPA</strong></td>
</tr>
<tr>
<td><strong>Sound</strong></td>
</tr>
<tr>
<td><strong>Transcript Corpus</strong></td>
</tr>
<tr>
<td>Current size: 790,000 running words</td>
</tr>
<tr>
<td><strong>Reading Material</strong></td>
</tr>
<tr>
<td><strong>Dictionary</strong></td>
</tr>
<tr>
<td>Current size: 23,000 complete paradigms.</td>
</tr>
<tr>
<td><strong>PoS</strong></td>
</tr>
<tr>
<td><strong>Background corpus</strong></td>
</tr>
<tr>
<td>Current size: 25M words.</td>
</tr>
<tr>
<td><strong>Tools</strong></td>
</tr>
</tbody>
</table>

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**
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 inconsistencies annotated for location, frequency, kind and degree of inconsistency, and level of significance.

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

CONSISTENCY PRINCIPLE

**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

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.