Garnishing a phonetic dictionary for ASR intake

Iben Nyholm Debess
Sandra Saxov Lamhauge
Grunnurin Føroysk Teldutala
ibendebess@gmail.com
sandralamhauge@gmail.com

Peter Juel Henrichsen
Danish Language Council
pjh@dsn.dk

Abstract

We present a new method for preparing a lexical-phonetic database as a resource for acoustic model training. The research is an offshoot of the ongoing Project Ravnur (Speech Recognition for Faroese), but the method is language-independent. At NODALIDA 2019 we demonstrate the method (called SHARP) online, showing how a traditional lexical-phonetic dictionary (with a very rich phone inventory) is transformed into an ASR-friendly database (with reduced phonetics, preventing data sparseness). The mapping procedure is informed by a corpus of speech transcripts. We conclude with a discussion on the benefits of a well-thought-out BLARK design (Basic Language Resource Kit), making tools like SHARP possible.

1 Introduction

We introduce a new method for pre-processing phonetic databases for use in ASR development. Our research, to be presented at NODALIDA 2019, is an offshoot of the ongoing Faroese ASR project (automatic speech recognition) called Ravnur. After giving some background on the project proper, we turn to the main focus of the present paper: the algorithm SHARP.

We first introduce the Ravnur components and the principles behind them (section 2), and then go into details with SHARP (section 3). In conclusion we offer some remarks on the challenges and advantages of developing an 'eco-system' of inter-dependent language technology resources.

Project Ravnur was initiated in January 2019 with the purpose of creating all the necessary constituents for developing high quality ASR for Faroese. One of the challenges of ASR for small languages is the sparsity of language resources, making the development of such resources a vital part of the project (Nikulasdóttir et al., 2018). Existing speech and language materials for Faroese have been developed for other purposes (Helgason et al., 2005; Johannesen et al., 2009; Hansen, 2014; Bugge, 2018; Debess, 2019), but these alone are insufficient in size, quality and/or availability. Beginning almost from scratch allowed us the advantage of establishing rational and explicit principles for all aspects of data collection, annotation, and processing.

2 The Faroese BLARK

A BLARK (Basic Language Resource Kit) is defined as the minimal set of language resources necessary for developing language technology for a particular language (Krauwer, 2003; Maegaard et al., 2006). Although the BLARK is not the main theme of this paper, it is detailed below as a prerequisite to the following section on SHARP.

2.1 Inter-dependent language resources

Only non-proprietary file formats are used (txt, csv, html, rtf, textGrid, wav, flac).

- SAMPA: the phonetic inventory is inspired by the SAMPA initiative providing computer-readable phonetic alphabets[1]. Following the tradition within Faroese phonetic research and description, our SAMPA includes the most common, salient, and distinctive phones and diacritics (Rischel, 1964; Helgason, 2003; Árnason, 2011; Thráinsson et al., 2012; Knooihuizen, 2014; Petersen and Adams, 2014; Weyhe,

[1] According to John Wells, the founding father of the international SAMPA initiative, the project is long closed, the website no longer maintained. As recommended by Wells (p.c.), we hereby put our suggestion for a Faroese SAMPA definition forward, inviting future projects to use it as a reference. The phone table (and documentation) is available at https://lab.homunculus.dk/Ravnshornid

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 395–399
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

2014). Our work is closely coordinated with the (now completed) Faroese TTS project (Helgason and Gullbein, 2002; Helgason et al., 2005).

- PoS: the tagset for Faroese complies with the Pan-European PAROLE meta-tagset (Bilgram and Keson, 1998).
- Dictionary: the dictionary encompasses largely all function words and irregular content words, and a substantial part of highly frequent content words. Each entry includes pronunciation, PoS, and frequency information. The dictionary is versatile by design and can be used for many purposes including traditional lexicographic editions, teaching materials (e.g. CALL and CAPT), TTS development, interactive voice-response systems, and more. The dictionary currently holds about 3,000 entries, aiming at 25,000 by January 2021.
- Speaker sessions: transcripts of speech recordings documenting the phonetic and prosodic variation of modern Faroese. Reading materials comprise a word list, a closed vocabulary reading (numerals 1-100, calculator commands), a phrase list (eliciting prosodic variation, intonation patterns, etc.), and a few samples of connected text (2-5 minutes each). Each session produces roughly 20 min. of speech. The speech corpus currently holds 8 hours of speech (26 speakers), aiming at 200 hours by January 2021 (project end). All acknowledged contemporary dialects of Faroese (Thráinsson et al., 2012) are covered.
- Transcript Corpus: the recordings are transcribed manually by multiple transcribers (orthography and SAMPA) and time coded according to the Ravnur conventions (https://lab.homunculus.dk/Ravnshornid). Phonetic transcription of speech production is carried out by trained phoneticians.
- Background Text Corpus: at present, the background corpus holds 13M words (formal and informal styles). Some of the material is collected in collaboration with Sjúrður Gullbein from the TTS project and Hjalmar P. Petersen from the University of the Faroe Islands.
- Background Speech Corpus: the background speech corpus consists of audiobooks and material from UiO (Johannesen, 2009; Johannesen et al., 2009) and elsewhere.
- Tools: the text and speech tools developed in Project Ravnur can be accessed at (https://lab.homunculus.dk/Ravnshornid).

2.2 Consistency Principle

All BLARK components relate to and depend on each other: each word appearing in a transcript must correspond to a lexical entry. Each manuscript (for recording sessions) must represent all SAMPA phones, and so forth. The Consistency Principle allows the BLARK to develop like an eco-system where the individual components feed off and grow from each other in an iterative process.

3 Garnishing the dictionary

We are now in a position to discuss the SHARP algorithm for optimizing lexical-phonetic information prior to the training of ASR acoustic models.

3.1 Phone inventories

When phoneticians need to represent pronunciation phenomena in symbolic form, they largely follow one of two strategies, either abstracting over speakers and contexts (the lexical approach) or sampling actual speech productions (the descriptive approach). ASR projects typically apply the lexical strategy only, shying away from the burden of phonetic transcription. Since classical phonetic dictionaries (complying with structuralist minimal-pair tests) are usually considered too rich for ASR purposes, lexical-phonetic forms are reduced prior to acoustic training, deleting certain phone types and collapsing others. To the best of our knowledge, the concrete reduction procedure is most often based on technological considerations or gut feeling rather than linguistic principle.

By way of an example, most popular commercial ASR applications for Danish allow users to supply phonetics for new lexical insertions, but in impoverished form without symbols for stød, accent, prolongations, assimilations, and only a subset (not a very rational one) of the Danish vowel inventory. Such linguistically unwarranted restrictions limit the general usability of the users' accumulated lexical contribution, in effect tying it to a particular ASR product.

Thus, in keeping with the Consistency Principle, we needed to devise a principle-based procedure allowing us to maintain the versatility of the dictionary and yet provide the reduced phonetic forms required for acoustic training. Our solution, called SHARP, utilizes the transcript corpus for deriving a reduced SAMPA in a non-destructive way.

Lex	*Trsc*	
X	→ *X*	opposed phones MATCH
X Y	→ *Y X*	adjacent phones 'swapped'
X Y	→ *Y*	1 phone skipped
X Y Z	→ *Z*	2 phones skipped
X Y Z	→ *W Z*	2+1 phones skipped
X Y V Z	→ *Z*	3 phones skipped
X Y V Z	→ *W Z*	3+1 phones skipped
X	→ *Y*	two opposed phones skipped
X	→ _	fallback rule: IGNORE

Table 1. Transduction rules. Lex = lexical-phonetic tier, Trsc = transcript tier. Rules also apply in mirrored versions (e.g. *Y* → *X Y*). Transduction of identical strings uses the MATCH rule only. The IGNORE rule ensures completion (_ is the empty string). The term 'skipped' is used for symbols only occurring in one tier.

3.2 The phonetic mapping

As mentioned above, each word appearing in the Transcript Corpus is also represented in the Dictionary. We can therefore align the phonetic representation of any phrase appearing in a transcript with its corresponding lexical projection. For alignment of phone strings, we employ a finite state transducer (FST) with limited look-ahead. Pairs of phone strings are traversed left-to-right applying the transduction rules in table 1.

Consider the alignment of two phonetic renderings of "vónandi er hann ikki koyrdur útav" *hopefully he hasn't driven off (the road)*, one lexical and one descriptive.

```
Lex:    [vOWnandIerhanIHdZIkOrdur0WdEAv]
Trsc:   [vOWnandIer anIS   kORDIRU dEAv]
```

Observe that this alignment corresponds to the FST transitions (h→_), (H→S) , (dZI→_), (r→R), (d→D), (u→I), (r→R), and (0W→U).

Repeating the alignment procedure for all phrases in the Transcription Corpus, a list of rule instances develops. A sample from the rule list (excluding instances of the MATCH rule) is shown below, with the number of instances.

```
128     (j → _)
96      (I → _)
80      (U → I)   *
68      (r → _)
58      (I → 3)
58      (r → R)
43      (U → _)   *
36      (U → 3)   *
32      (d → _)
32      (i → I)
25      (d → D)
22      (E A → a)
21      (E A d → a)
```

Consider the three starred rules, all concerning the lexical phone [U], in 80 cases pronounced as [I], in 36 cases as [3], and in 43 cases not pronounced at all. There are several (less frequent) (U→*X*) rules for (*X*≠U). In comparison, the MATCH rule (U→U) has only 63 occurrences, contributing to the general impression that [U] is an unstable phone exposed to pronunciation variation.[2]

Several other phones are shown to be unstable in this sense, evident in rules such as (`→_), (5→_), (j→_), (4→E), (w→_), (u→o). Such rules we shall call *skewed*. Formally, skewed rules are determined by

$$count(X \longrightarrow X) < \sum count(X \longrightarrow Y) \textit{ for all } (Y \neq X).$$

3.3 Generations

Skewed rules are interpreted in SHARP as transformation rules and are applied everywhere in the Dictionary and Transcription Corpus (in size-order), creating new tiers of phonetic forms. In some cases, phone symbols are cut out of the SAMPA renderings (like (5→_)), in other cases two phones are collapsed into one (e.g. (u→o)), effectively reducing the cardinality of the phone inventory. We call this new lexical tier of transformed phonetic forms the Generation-1 tier (or simply G1).

The transduction procedure is repeated using G1 as lexical forms, producing a G2 tier, and so forth. With each new generation, the cardinality of the phone table decreases (often by 1-3 items) while the average inhabitation (number of exemplars) in the remaining types increases.

[2] Observe that rule types (*X*→_) outnumber the mirror form (_→*X*), as a sign of a general fact: phonetic dictionaries aim at a high degree of articulatory explicitness while speech production show exactly the opposite tendency.

At G12 the iteration stops naturally as no more skewed transduction rules can be found. At this point, 24 out of the original 45 SAMPA symbols are still present. It is an important observation, though, that the meaning of each remaining symbol at this point has changed. The symbols can therefore no longer be expected to represent the usual phonetic flavours.

3.4 Turning to ASR

Acoustic models for ASR are trained on sound samples, phonetically labeled. Two complementary factors affect the training efficiency, *parsimony* (a smaller set of labels provides more robust training) and *discrimination* (a larger set preserves more phonetic distinctions).

Since we now have a procedure for gradually reducing the phonetic richness as controlled by (the transcription of) actually occurring speech production, the next step is to evaluate the G0, G1, … G12 phonetic forms for training acoustic models. We use the sphinxtrain engine (ver. 5prealpha, cf. https://github.com/cmusphinx/), employing a standard ten-fold cross validation regime to yield statistically valid test figures.[3]

3.5 Preliminary results

Our initial results are encouraging if preliminary. Using our current smallish dictionary of 2990 entries and 1366 spoken phrases only, our acoustic models do not reach impressive results in terms of absolute WER figures (word-error rate). However, as our performance measures are reasonably consistent (cf. the narrow error bars in fig.1) it still makes sense to compare learning sessions across SHARP-generations.

From an initial WER at 50.2%, error rates improve rapidly: WER_{G1}=41.2%, WER_{G2}=31.6%, WER_{G3}=25.6%, …, WER_{G7}=13.8%. Of course, this impressive recovery is owed to the very poor outset, and also to an atypical ASR setup based on small linguistic databases. We do not know yet to what extent the SHARP algorithm will remain relevant in more realistic scenarios. However, it seems safe to conclude that SHARP may offer a relief to very small ASR projects in distress.

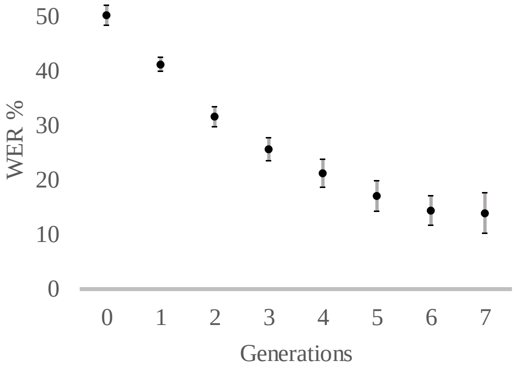

Figure 1. ASR results trained on SHARPened lexical-phonetic forms. The graph shows Word-error rates (WER) for each SHARP generation. Error bars: standard deviation for data sets after ten-fold cross validation. Average WER keeps improving somewhat in generations >7, however less significantly so as error margins increase.

Our work is clearly in progress, and the specifics of the SHARP implementation are bound to change as our BLARK matures. Among many new features we would like to test context sensitive transformation rules ($X\ A\ Y \rightarrow X\ B\ Y$) as used by phonologists. However, this step (and many others) make sense only for much larger pools of phonetic samples.

4 Concluding remarks

Much R&D in speech technology has been hampered by implicit or explicit obligations to recycle existing, often inadequate, databases. One example is the government-supported Danish ASR project in the mid-2000s leaning on mediocre speech data from NST, lexical data from the Danish TTS project and various sources unrelated to the project objectives (cf. Kirchmeier et al 2019). Recognition rates never met international standards, much labour was wasted on smartening up poor data, and yet the delivered modules could not, for legal reasons, be shared publicly.

In contrast, the Faroese ASR project, starting afresh, could adopt strict consistency principles to be followed by all, from lexicographers to field workers. Carefully synchronized lexical and descriptive procedures paved the way for the SHARP tool presented in this paper, exploiting the complementarity of theory-driven and data-driven phonetics and getting the most out of our smallish, but undefiled databases.

"More data will solve any problem", "Principles are for sissies", "Fire your

[3] For the sake of reproducibility, we use a flat language model with minimum likelihood (0%) for all *n*-grams (*n*>1) and equal likelihood for individual words.

linguists!". Such fresh attitudes are currently shared by many developers. We invite the serious ASR manufacturer to rediscover the power of linguistic precision.

Acknowledgments

Project Ravnur wishes 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.

References

Anna Björk Nikulasdóttir, Inga Rún Helgadóttir, Matthías Pétursson and Jón Guðnason. 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.

Bente Maegaard, Steven Krauwer, Khalid Choukri and Lise Damsgaard Jørgensen. 2006. The BLARK concept and BLARK for Arabic. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 773-778.

Edit Bugge. 2018. Attitudes to variation in spoken Faroese. *Journal of Sociolinguistics* 22(3):312-330.

Eivind Weyhe. 2014. Variatión av i og u í herðingarveikari støðu í føroyskum. *Fróðskaparrit*, 61:116-136. Fróðskapur, Tórshavn, Faroe Islands.

Höskuldur Thráinsson, Hjalmar P. Petersen, Jógvan í Lon Jacobsen and Zakaris Hansen. 2012. *Faroese, an overview and reference grammar*, 2nd edition. Fróðskapur, Tórshavn, Faroe Islands, and Linguistic Institute, University of Iceland, Reykjavík, Iceland.

Iben Nyholm Debess. 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

Janne Bondi Johannesen. 2009. A corpus of spoken Faroese. *Nordlyd*, 36(2):25-35.

Janne Bondi Johannesen, Joel Priestly, Kristin Hagen, Tor Anders Åfarli and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Kristiina Jokinen and Eckhard Bick (eds.). 2009. *NEALT Proceedings Series*, 4:73-80.

Jonathan Adams and Hjalmar P. Petersen. 2014. *A Language Course for Beginners*, 3rd edition. Stiðin, Tórshavn, Faroe Islands.

Jørgen Rischel. 1964. Toward the Phonetic description of Faroese vowels. *Fróðskaparrit*, 13:99-113.

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

Kristján Árnason. 2011. *The Phonology of Icelandic and Faroese*. Oxford University Press, Oxford, UK.

Pétur Helgason. 2003. Faroese Preaspiration. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2517-2520. Universidad Autònoma de Barcelona, Barcelona, Spain.

Pétur Helgason and Sjúrður Gullbein. 2002. Phonological norms in Faroese speesch synthesis. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2269-2272, Denver, Colorado.

Pétur Helgason, Sjúrður Gullbein and Karin Kass. 2005. Færøsk talesyntese: Rapport marts 2005. In Henrik Holmboe (ed.). 2005. *Nordisk sprogteknologi 2005 - Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*, pages 51-58.

Remco Knooihuizen. 2014. Variation in Faroese and the development of a spoken standard: In search of corpus evidence. *Nordic Journal of Linguistics*, 37(1):87-105.

Sabine Kirchmeier, Peter Juel Henrichsen, Philip Diderichsen and Nanna Bøgebjerg Hansen. 2019. *Dansk Sprogteknologi i Verdensklasse*.

Steven Krauwer. 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,* Moscow, Russia.

Thomas Bilgram and Britt Keson. 1998. The Construction of a Tagged Danish Corpus. *Proceedings of the 11th Nordic Conference of Computational Linguistics, NODALIDA 1998*, pages 129-139.