AN ICELANDIC PRONUNCIATION DICTIONARY FOR TTS

*Anna Björk Nikulásdóttir, Jón Guðnason**

Reykjavik University
School of Science and Engineering
Reykjavik, Iceland

Eiríkur Rögnvaldsson

University of Iceland
Faculty of Icelandic and
Comparative Cultural Studies
Reykjavik, Iceland

ABSTRACT

This paper describes an Icelandic pronunciation dictionary for speech applications and its processing for use in a text-to-speech system for Icelandic. Cleaning and correction procedures were implemented to create a consistent training set for grapheme-to-phoneme conversion modeling, needed for the automatic extension of the dictionary. Experiments with the original version of the dictionary and the cleaned version described in this paper as training sets for a joint sequence g2p algorithm show a clear benefit of using clean data for training, both in terms of PER and in terms of categories of errors made by the g2p algorithm. The results of the dictionary processing where also used to create an initial version of an open source database for Icelandic speech applications.

Index Terms— Pronunciation dictionary, TTS, g2p, Icelandic

1. INTRODUCTION

Pronunciation dictionaries are one of the core components of automatic speech recognition and automatic speech synthesis systems. Their main purpose is to function as a look-up dictionary for the system, mapping between the grapheme and phoneme representations of an entry. A dictionary will, however, always be a limited resource in that a speech application system will constantly encounter new words. Letter-to-sound rules and/or grapheme-to-phoneme (g2p) models, learned from the dictionary, can be used to automatically transcribe unknown words.

For the ongoing development of an open text-to-speech (TTS) system for Icelandic, a high quality, consistent pronunciation dictionary is needed, as well as a reliable g2p model to automatically transcribe words not found in the lexicon.

An Icelandic pronunciation dictionary was created for the first Icelandic ASR project, *Hjal* [1]. In 2012 it was extended with about 9,000 words for use in a commercial TTS system for Icelandic and now contains about 60,000 words. The dictionary has been a valuable resource in later ASR projects in Iceland [2, 3] and at Google [4]. The words in the dictionary were collected using text corpora available at the time, and then manually transcribed, word by word, without any lexicographic tools. The Icelandic pronunciation dictionary, hereafter IPD, thus has a flat structure similar to the widely used open CMU pronunciation dictionary for English [5]. The dictionaries are stored as text files with one word and its transcription per line. Due to the rich morphology and productive compounding mechanism in Icelandic, the Icelandic lexicon would likely benefit very much from a deeper structure allowing for interconnection between entries. An example of a speech application lexicon of this kind is Combilex [6], where morphological analysis is used to increase consistency and to facilitate maintenance of the lexicon.

In this paper we describe the processing of the IPD for use in Icelandic TTS system. The main focus is on creating a consistent, unambiguous training set for g2p modeling. Furthermore, an initial lexicon database for speech applications is introduced, based on the dictionary processing and automatic g2p transcriptions.

2. RELATED WORK

During the work on an open ASR system for Icelandic and an ASR system for Icelandic parliamentary speeches, some consistency checks, error corrections and extensions using a g2p algorithm have already been performed, although not described in detail [2, 3]. The original pronunciation dictionary and the extended versions from these projects are available at the Icelandic language resources website, *Málföng*[1]. Some of the processing steps described in this paper were to some extent performed in these previous projects.

Jansche [4] also described a systematic approach to identify errors in the IPD by grapheme to phoneme alignment. Starting with an initial set of g2p mappings, valid mappings were added by iteratively aligning portions of the dictionary. As the parameter set was complete, words and transcriptions that failed to align were marked as erroneous. In this way a list of 480 words with errors was extracted and suggestions

*Thanks to the Icelandic Language Technology Fund for funding.

[1] http://malfong.is/

for corrections made. The list of words extracted by this procedure, together with the original dictionary sources is available at Google's language resources github repository. However, the list of erroneous words there with correction suggestions contains 659 words.[2] This list will be used as a reference at the end of our dictionary processing.

3. DICTIONARY PROCESSING

The aim of the dictionary processing described in this paper is to reduce errors and to increase consistency in the transcriptions. With a training set with consistent transcriptions, a more reliable grapheme-to-phoneme model can be trained.

3.1. Phoneset

The IPD contains transcriptions in the IPA[3] and SAMPA[4] transcription systems:

`banki` pauŋ̊cɪ `bauJ0J_I` ('bank')

To examine the consistency of both versions, the transcriptions were analyzed with regard to a closed set of phoneme symbols in the respective system. In line with [7], a closed set of 58 transcription symbols was defined for each system.

The original IPD contains 65,020 entries. When analyzing the IPA transcriptions, 599 entries with non valid symbols were found, and 4,027 erroneous entries in the SAMPA transcriptions. Most errors in the IPA transcriptions are due to invalid symbols at the end of the transcriptions ('\' or '/') and word separation symbols (space character or '#'). Other common errors include using a grapheme instead of a phoneme symbol (e for ɛ, o for ɔ), and misplacing of the length symbol or the voicelessness diacritic. For further processing the IPA version of the transcriptions was chosen, due to fewer phoneset dependent errors. The processed dictionary can be automatically converted to SAMPA or X-SAMPA as needed.

For this first processing step, the same procedure was performed as for all succeeding steps :

1. Identify error/inconsistency through pattern analysis
2. If simple automatic correction is possible: perform correction
3. If correction is not performed: remove entry from dictionary

Non-valid symbols were replaced by their valid alternative (ɛ for e, etc.) or removed (\, / etc.). Entries containing word separation symbols were removed from the dictionary. These entries are either abbreviations (`DNA`, `ADSL`), words including a hyphen (`Austur-Berlín`, `Suður-Afríka`) or compounds, where a straight forward transcription would cause an error, because vowels at the compound boundary would be interpreted as a diphthong (`dægurmálaútvarpi` /taiːɣʏrmaula#uːtvaɾ̥pɪ/). All these categories of entries are very productive and should be handled by a text processing module rather than within the core pronunciation dictionary. In total only 159 entries matched these categories and were removed from the dictionary, leaving 64,861 entries for further processing.

3.2. Postaspiration

In Icelandic the plosives `p, t, k` are pronounced with postaspiration word initially before vowels and liquids: `kaffi` - /kʰafɪ/ ('coffee') or `tromma` - /tʰrɔma/ ('drum'). If the postaspiration is not considered in the transcription a distinctive feature might be missing between two words, as this example from the IPD shows: `pakka` - /pahka/ ('package' acc/dat/gen., 'to pack', instead of the correct /pʰahka/) and `bakka` - /pahka/ ('tablet' acc/dat/gen.; 'to back'). Such examples do not only affect these two entries, but all word forms and compounds in the dictionary containing these words. In this case there are 12 entries containing `pakka` and 35 entries containing `bakka` in the dictionary.

Postaspiration seems to be one of the largest inconsistency issues in the IPD. It is sometimes included and sometimes not, and often the same word in different word forms or compound combinations, like `pakka`, is not transcribed in a consistent way throughout the dictionary. This is fairly easy to correct automatically for word initial position. To correct the transcriptions where an affected word is a head word of a compound, i.e. the plosive does not stand word initially in the entry, compound analysis is needed. When analyzing words that should have postaspiration on the initial plosive, 5,330 words were found to miss the postaspiration and were corrected, whereas 679 words already included the postaspriation symbol. These numbers do not take the occurrences of the corresponding words as head words in compounds into account.

3.3. Diphthong consistency

There are seven different diphthong symbols in the Icelandic phoneset, five of them, /ai, au, ou, ei/ and /œy/ can occur either short or long, and two of them, /ʏi/ and /ɔi/, can only be short. There are one-to-one grapheme correspondences to each diphthong phoneme, but also several other grapheme combinations produce a phoneme combination containing a diphthong in certain contexts. As an example, `bæi` and `bagi` are both transcribed /paiːjɪ/ but only æ is the direct grapheme correspondence to the diphthong /ai/.

In the IPD transcriptions, sometimes a combination of monophthongs is used instead of a diphthong, or the wrong

[2] https://github.com/googlei18n/language-resources/tree/master/third_party/
[3] http://www.internationalphoneticalphabet.org/
[4] https://www.phon.ucl.ac.uk/home/sampa/

symbols are used for the diphthongs, causing the analyzing algorithm to identify two monophthongs. Partly the differences are systematic, as the additional word list from 2012 (see Section 1) seems to follow other guidelines than the original dictionary. For example the diphthong /ou/ is in this list written as /ɔu/ as in the many surnames ending in `−dóttur` ('daughter-of' acc./dat./gen.): `Valsdóttur` - /valstɔuhtʏr/ vs. `Einarsdóttur` - /eiːnarstouhtʏr/ in the original dictionary. The transcripts are analyzed as /t ɔ u h t ʏ r/ and /t ou h t ʏ r/ respectively.

Furthermore, in the dialect of the Icelandic Westfjords, certain words are pronounced with a monophthong instead of a diphthong as in standard Icelandic. Both variants may be included in the IPD (see Section 3.4 for discussion on dialects and pronunciation variation).

Entries containing graphemes and grapheme combinations that should be transcribed with a diphthong according to [8] were extracted and the transcripts examined. About 14.5% of the 31,600 extracted entries do not match the diphthong transcription from the defined phoneset. No attempt was made to correct these entries, since in exceptional cases the grapheme combinations in question are at a compound boundary, e.g. `−agi−` in `músagildru: músa+gildru` ('mouse trap'). Instead, the entries were removed from the dictionary to be added later through compound analysis and g2p processing.

3.4. Transcription variants and dialects

The IPD was created for ASR where pronunciation variants need to be recognized, but where the need for the knowledge of the origin of the transcription variants is not as important as in a TTS system. The transcription variants in the IPD thus are not labeled in any way. For TTS as well as for consistent training of a g2p algorithm, there has to be a way to identify the source of different transcriptions for each entry. In that way different dialects and pronunciation styles can be produced, given the corresponding speech recordings. Variation in phonemic transcription can have several sources in Icelandic: a) dialect variation, b) homography, c) difference in situation dependent pronunciation style or individual differences and d) errors.

3.4.1. Dialects

Icelandic does not have strongly diverse dialects, rather one can speak of regional variants, mainly regarding pronunciation. The IPD contains four of the six regional pronunciation variants, at least to some extent: 1) Voiced pronunciation of /l m n ð/ before the plosives /p t k/. Standard pronunciation is voiceless. 2) Monophthongs before /ɪ ŋ/ (voiced and voiceless), whereas standard pronunciation uses diphthongs: `langa` ('want') /laŋka/ vs. /lauŋka/. 3) Monophthongs before `gi`: `agi` ('discipline') - /aːjɪ/ vs. /aiːjɪ/. and 4) The /xv/ vs. /kv/ pronunciation of the grapheme combination `hv`: `hvítur` ('white') - /xviːtʏr/ vs. /kviːtʏr/.

3.4.2. Homography

Homographs with different pronunciations are not very common in Icelandic. The main source of different pronunciation for the same grapheme combination is `<vowel>ll<vowel>`. The more common pronunciation is /tl/, but /l/ is also possible. Most homographs can be classified if grammatical information is available, like for `bolla` (/pɔtla/ 'cup' if accusative, /pɔla/ 'bun' if nominative) or `gellur` (/cɛlʏr/ 'kind of fish' if noun plural, /cɛtlʏr/ '(it) rings' if verb), but there are examples where semantic context is needed for disambiguation, as for `galli` and `galla` (/kalɪ/, /kala/ 'overall' or /katlɪ/, /katla/ 'flaw', nominative and accusative respectively).

3.4.3. Individual characteristics

Some variation cannot directly be assigned to dialect variation but is rather an individual pronunciation characteristic. They are partly systematic and thus are contained in the IPD. Examples are short diphthongs where standard pronunciation would be a long one, and at the same time lengthening of the following consonant, like `fleiri`('more') - /fleiːrɪ/ vs. /fleirɪ/ and pronunciation of /x/ rather than /k/ in certain contexts, e.g. `læknir` ('physician') - /laihknɪr/ vs. /laixnɪr/.

Pronunciation can also vary according to the speech situation. When reading aloud, for example, pronunciation tends to be more clear than in casual conversation. This kind of variation is present in the IPD transcripts, although normally not as multiple entries for one word but rather across the dictionary in compounds and derivations of the affected words. An example is the transcription of the word `barna` ('children') which can be correctly transcribed as /bartna/ or /batna/. In the IPD there are 105 entries containing `barna`, like `barnabók` ('childrens' book'), transcribed /pa**tn**apouːk/ and `barnabókasöfn` ('childrens' libraries'), transcribed /p**artn**apɔːukasœpn/. This kind of inconsistency is rather common in IPD. Through compound analysis these kind of variants could be entered for each base word and propagated to the compounds like the other variants and dialects.

3.4.4. Processing of variants

There are 3,535 entries that have multiple transcriptions in the IPD. By far the most common alternative transcriptions are /k/ vs. /x/, as in `hver` ('who', 'geysir') /kvɛːr/ vs. /xvɛːr/ or `vaxa` ('grow') /vaksa/ vs. /vaxsa/. Other common alternatives are due to the difference in voiced vs. voiceless nasals and the lateral /l/ that distinguish the North-Icelandic pronunciation variants from standard Icelandic: /n̥/ vs. /n/, /l̥/ vs. /l/ etc. Some words have two transcript variations where one is correct and one erroneous, i.e. does not describe a pronunciation variant. An example are transcriptions containing /c/

vs. /k/ between high, front vowels /ɪ/ and /i/, where the pronunciation of /k/ would be impossible by Icelandic pronunciation rules. From the most common alternative transcriptions the ones corresponding to standard pronunciation were chosen, other entries with multiple transcriptions were removed from the dictionary. The alternatives will not be permanently discarded, but stored and labeled e.g. by dialect. Alternative transcriptions were kept for all identified homographs. The size of the dictionary after processing multiple entries is 54,360 entries.

3.5. Compound Analysis

Icelandic is highly productive in producing compounds. This means that the same words can be found as parts of numerous compounds in the dictionary. To identify inconsistencies in transcriptions of compound parts, decompounding and comparison of transcripts was performed. Another purpose of the compound analysis was to remove compounds from the g2p training set and identifying compound parts that can be linked across the lexicon. An initial database of compounds was created by decomposing all longer words and search for their parts in the lexicon. Then some manual extensions were made and the database used for decompounding contains 4,700 entries for modifiers and 5,840 head words. The entries are not lemmatized, i.e. there can be multiple entries for the same head lemma depending on its inflected forms.

This is a preliminary database, a more reliable analysis will be possible with the next version of the Database of Modern Icelandic Inflections [9], which will be connected to a large database of compounds, not yet published. A compound analysis algorithm utilizing this data was published in [10].

When comparing transcriptions of compound parts, over 200 errors were identified, along with other inconsistencies. The inconsistencies mostly have to do with the difference between clear and less clear pronunciation as described in Section 3.4. The different versions are collected and labeled, and for the current lexicon and g2p training set the more clear pronunciation variants were selected. The transcription variants are both valid but should be labeled and kept consistent within each version of a TTS lexicon. After the compound analysis over 13,400 compounds were removed from the dictionary, along with the about 200 identified errors.

3.6. Vowel length

In Icelandic all vowels can be short or long. This is also true for many consonants: [p], [t], [k], [c], [s], [f], [l], [m], [n] and [r] all have a long version as well as a short one. Compare e.g. `amma` ('grandmother') [amːa] and `ama` ('bother') [aːma] or `bolla` ('bun') [pɔlːa] and `bola` ('to budge', 'bull' acc./dat./gen.) [pɔlːa]. However, in the phonemic transcription, only the long vowels are marked, and thus vowel length becomes the only distinctive feature between words like `amma` and `ama`.

The general rule for vowel length in Icelandic is that vowels are long in final position in a monosyllabic word and in a stressed syllable before one short consonant, otherwise vowels are short. Stress in Icelandic follows an alternating pattern: the first syllable of a word is always stressed, followed by a non-stressed syllable. For longer words alternating patterns of primary stress, no stress and secondary stress are the rule. In compounds a second primary stress can be located at the first syllable of the head word if the modifier has more than one syllable, otherwise the alternating pattern would be broken. Stress/length patterns for longer words, e.g. compounds of compounds, can be difficult to define explicitly and can simply be a question of phonetic perception of the transcriber.

Vowel length in the IPD generally follows the rule that if a compound modifier has more than one syllable, the following head word is treated as the word in isolation regarding the length marking. The word `son` ('son', acc. 'son-of' in compounds) has a long, stressed vowel in isolation /sɔːn/, but because of the alternating stress pattern it looses stress (and length) when connected to a single syllable modifier: `Pálsson (Páls+son)` /pʰaulsɔn/ vs. `Kristinsson (Kristins+son)` /kʰrɪstɪnsɔːn/. In the wordlist added in 2012 the vowel in /sɔn/ is always short.

In very long compounds the decision for a long vowel becomes still less explicit. It is e.g. questionable if the last vowels with length symbols in the following examples really are stressed:

`hagfræðiráðgjafarfyrirtækis`
/haxfraiðɪrauðcavarfɪːrɪrtʰaiːcɪs/

`iðnaðarráðuneytisins`
/ɪðnaðarauːðʏneiːtɪsɪns/

It will have to be examined if and how inconsistencies in length marking influence the quality of ASR and TTS, but it is impossible for a statistical g2p algorithm to learn length marking in longer words without knowledge of morphology. The stress patterns together with length labeling could be added in a post processing step, leaving the g2p training set without length labels. For now the length symbols are kept as is.

3.7. Alignment

Similar to [4] we seek to identify errors by aligning graphemes and phonemes. We start without a manual list of grapheme to phoneme mappings. Instead we identify the most common mappings by aligning words and transcriptions with equal symbol length, e.g. adam - /aː t a m/ or samruna - /s a m r ʏ n a/. Mappings occurring at least 1,000 times in this analysis are considered correct, resulting in an initial map of 56 grapheme-phoneme pairs. Then a list of eleven graphemes and grapheme combinations that are not represented by a

one-to-one relation to a phoneme is added to the map, e.g. 'hr' - /r̥/ or 'sl' - /s t l/. Additionally it is ensured, that each vowel mapping contains both the long and the short version of the vowel. Using the pairs from this map as anchors, all entries of the dictionary are aligned to create a final alignment map, additionally containing new mappings occurring more than 100 times in this second pass. The final map contains 119 mappings and is used to perform forced alignment of the whole dictionary. We choose forced alignment instead of marking failed alignments, since we want to inspect the discovered mappings. For the final filtering, g2p mappings occurring less than 20 times in the alignment results were inspected for errors. Of the 285 pairs 38 were valid, entries containing any of the other mappings were removed from the dictionary, all in all 481 entries.

These entries can contain spelling errors, like `tfra-flautunni` (for `töfraflautunni`, 'the magic flute'), transcription errors, `öfluga` - /œ k l ʏ ɣ a/ instead of /œ p l ʏ ɣ a/, often also a correct transcription of another related word form, like `þögnuðu` - /θ œ k n ɪ n ɪ/ as for `þögninni`. Not all entries containing a rare grapheme to phoneme mapping are erroneous. Many of them are foreign words containing non-native Icelandic grapheme and/or phoneme combinations, and should therefore be kept separately from the core g2p training set: `lasagne` - /l a s a ɲ j a/ or `Netscape` - /n ɛ h t s k ei p/. Further, transcriptions of single consonants, like `n` - /ɛ n/ are on this list, and occasional correct transcriptions, or false positives, as well.

As mentioned in Section 2, [4] described error detection from the IPD using an alignment algorithm. During the cleaning process described here, we encountered numerous errors not contained in the error list detected in that work. When comparing our cleaned dictionary to this list, we found that 29 erroneous entries from the total error list of 659 entries were still uncorrected in the processed IPD. We removed these remaining errors as well.

The final version of the clean IPD contains 40,322 entries, meaning that we have removed about 24,000 entries from the original IPD. The list of removed entries contains errors, pronunciation variation, foreign words, abbreviations, inconsistencies and compounds. The number of false positives, i.e. correctly transcribed words that were not subject to the cleaning and correction procedures, can not be estimated (see also [4]).

4. DICTIONARY DATABASE

As a result of the dictionary processing an initial dictionary database was created. Each word is stored with its transcription and labeled transcription variants if applicable (dialect, style). Compounds are stored with references to their modifier and head, allowing for consistent transcription of a word throughout the dictionary. When choosing a dialect or pronunciation style, all compounds will automatically follow the defined dialect/style without any manual effort in ensuring that all entries affected by a certain variation are changed. Special postprocessing rules will handle possible assimilation between modifer and head, e.g. prevent repetition of consonants as in `Guðmunds+son` - /kvʏðmʏnts+sɔːn/ - /kvʏðmʏntsɔːn/ and change /nt/ to /m/ before a head word starting with a bilabial consonant: `land+búnaður` - /lant+punaðʏr/ - /lampunaðʏr/. Such postprocessing rules also increase consistency, since manual transcriptions might occasionally fail to follow such rules.

Compound analysis can be used as a first step when transcribing unknown words. Instead of immediately applying a g2p model, the correct transcription of one or both parts of an unknown compound can be extracted from the dictionary.

Further labels, like part-of-speech and foreign word labeling are also included, but more detailed morphological information on inflection and derivation will be needed to fully exploit the idea of consistent transcriptions of morphological units.

The final design of the database will aim at being applicable for other languages with similar word building mechanisms. Thus methods for the extraction of application specific dictionaries can be largely language independent.

5. G2P TRANSCRIPTION EXPERIMENTS

We trained g2p models with the Sequitur g2p toolkit [11, 12], on the one hand based on the 40k entries of the cleaned IPD and on the other hand based on the raw version of IPD. Since the test and development sets needed to be manually inspected and corrected in line with the consistency rules set for the TTS dictionary, the size of the test and development sets were limited to 700 and 300 entries respectively. The entries were randomly chosen from the IPD and from an extended version of the IPD, used in the previously mentioned open ASR system (see Section 2). The selection was purely random, i.e. no attempt was made to include a predefined portion of words having certain features, like different correct pronunciations and dialect variants. The words in the test and development sets were removed from both training sets.

The core adjustable parameters for the Sequitur toolkit are a) L, the maximum number of phonemes and graphemes in each *graphone*, a unit of zero to L graphemes mapped to zero to L phonemes (whereas mappings from zero graphemes to zero phonemes are ignored), and b) M, the number of adjacent graphones to consider, i.e. the gram length. The best results in our experiments were achieved with L=1 and M=5. Held-out set in all experiments was 5% of the size of the training set.

The results of the automatic g2p transcriptions are shown in Table 1. The phone error rate (PER) is 3.4% for the test set with a model trained on the processed version of the IPD, and 6.1% when using the raw, original version. Word error rate (WER) is about 12% higher absolute and over 50% higher

Training set	Test set	WER	PER
Raw IPD	Dev	38%	5.65%
Clean IPD	Dev	25.33%	2.82%
Raw IPD	Test	37.29%	6.1%
Clean IPD	Test	24.43%	3.4%

Table 1. Results of g2p testing using the original Icelandic pronunciation dictionary vs. a cleaned version.

Error category	Raw IPD		Clean IPD	
Vowel length	42	28.7%	53	61%
Postaspiration	50	34%	19	21.8%
Dialect	6	4%	0	0%
Style/variation	7	4.9%	2	2.3%
Consonant voiceness	3	2%	4	4.6%
Other errors	39	26.5%	9	10.3%
SUM	147	100.0%	87	100.0%

Table 2. Error categories found in the g2p results of the development set. Error counts and percentages are given for each category.

relative when training on the raw IPD compared to the cleaned version.

The differences in the results are not only due to PER and WER. The errors made are also of different categories. While almost all errors made by the model trained on the clean dictionary concern vowel length or postaspiration, the raw model shows more non-matching transcripts because of dialect or style/variation and 26.5% of the errors are non categorized transcription errors, often due to wrong transcription of diphthongs. Table 2 shows the difference in error categories between the models, analyzed from the development set results. The model based on the raw IPD transcribes several words according to another dialect than defined in the test set. A word like `banka` is affected by two different dialects: the Northern Icelandic dialect, which has a voiced nasal before a plosive, and the dialect of the Westfjords which speaks a monophthong before `-nk` instead of a diphthong as in standard Icelandic. In the transcription of the development set both dialects are present: `bankaráðið` - /paŋkarauðɪð/ (Westfjords dialect), and `bankarnir` - /pauŋkatn̥ɪr/ (North Icelandic). No standard transcription of this grapheme combination is present in the results of the raw IPD g2p model.

The largest error groups in the results of the clean IPD training set, vowel length and postaspiration, can be dealt with using postprocessing rules. Compound analysis can ensure initial postaspiration in head words and, followed by syllabification and stress pattern analysis, the vowel length can be set according to the rules described in Section 3.6.

6. CONCLUSION AND FUTURE WORK

In this paper we described the processing of a manually crafted pronunciation dictionary with transcription variants and a flat structure. Through several automatic correction and cleaning procedures we succeeded in increasing consistency and removing errors, resulting in a more reliable grapheme-to-phoneme model for automatic transcriptions. Furthermore, an initial version of a relational dictionary database for speech applications was created that will allow for the extraction of coherent versions of a pronunciation dictionary with regard to dialects and pronunciation variants. The database will have to be developed further and a connection to the Database of Modern Icelandic Inflection should be considered. The database will be used to create pronunciation dictionaries and g2p models for an open TTS system for Icelandic currently under development. Post processing rules, syllabification and stress pattern extraction will be added to the system's frontend to create a fully-fledged TTS dictionary for Icelandic.

All design decisions will take language independence into account, and thus the database schema and dictionary construction methods will be applicable for other languages with word building mechanisms similar to Icelandic.

7. REFERENCES

[1] Eiríkur Rögnvaldsson, "The Icelandic speech recognition project Hjal," in *Nordisk Sprogteknologi. Årbog 2003*, Henrik Holmboe, Ed., pp. 239–242. 2004.

[2] Anna Björk Nikulásdóttir, Inga Rún Helgadóttir, Matthías Pétursson, and Jón Guðnason, "Open ASR for Icelandic: Resources and a Baseline System," in *Proceedings of LREC*, Miyazaki, Japan, May 7-12 2018.

[3] Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason, "Building an ASR corpus using Althingi's parliamentary speeches," in *Proceedings of Interspeech*, Stockholm, Sweden, August 20-24 2017, pp. 2163–2167.

[4] Martin Jansche, "Computer-aided quality assurance of an Icelandic pronunciation dictionary," in *Proceedings of LREC*, Reykjavik, Iceland, May 26-31 2014.

[5] Carnegie Mellon University, "CMU pronunciation dictionary 0.7b," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2015.

[6] Korin Richmond, Robert Clark, and Sue Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proceedings of Interspeech*, Makuhari, Chiba, Japan, September 26-30 2010, pp. 1974–1977.

[7] Eiríkur Rögnvaldsson, "Phonetic transcription guideline: Icelandic," Tech. Rep., 2003.

[8] Eiríkur Rögnvaldsson, *Íslensk hljóðfræði. Kennslukver handa nemendum á háskólastigi*, Málvísindastofnun Háskóla Íslands, 1989.

[9] Kristín Bjarnadóttir, "The Database of Modern Icelandic Inflection," in *Proceedings of the Workshop on Language Technology for Normalization of Less-Resourced Languages - SaLTMiL 8 - AfLaT2012*, 2012, pp. 13–18.

[10] Jón Friðrik Daðason og Kristín Bjarnadóttir, "Kvistur: Vélræn stofnhlutagreining samsettra orða," *Orð og tunga*, vol. 17, pp. 115–132, 2015.

[11] Maximilian Bisani and Hermann Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[12] "Sequitur g2p," https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html, 2016.