Overview

- A manually transcribed pronunciation dictionary for Icelandic (IPD) was examined through pattern analysis and alignment procedures.
- The aim was to create a more consistent version of the dictionary to improve grapheme-to-phoneme models needed in a TTS system under development.
- During cleaning the size of the dictionary was reduced from ~65,000 entries to ~40,000 entries.
- Significant improvement in PER for automatic transcriptions based on a g2p model.

Sources of variation:
- Dialects
- Homography
- Person and/or situation dependent
- Common alternative transcriptions processed, one alt. chosen
- Other entries with multiple transcriptions removed

1. Phonemes consistency

- Corrections with regard to a closed set of 58 transcription symbols
- IPA transcripts used
- 440 entries corrected, 159 entries removed (abbreviations, compounds)

2. Postaspiration

- A large inconsistency issue
- Partly a distinctive feature: /pʰahka/ vs. /pʰahka/ (bakka vs pakk)
- Automatic correction: /pʰahka/ vs. /pʰahka/
- 5,330 entries corrected

3. Diphthong consistency

- Seven different diphthong symbols: /ai, au, ou, ei, oey, vi, öi/
- Both regular grapheme(s)-to-diphthong correspondences and various context/dialect dependent combinations
- Entries not matching grapheme-diphthong rules deleted: ~4,500 entries

4. Variation

- Sources of variation:
  - Dialects
  - Homography
  - Person and/or situation dependent
  - Common alternative transcriptions processed, one alt. chosen
  - Other entries with multiple transcriptions removed

5. Compound analysis

- Decompounding and comparison of components across the dictionary
- Choose clear, standard pronunciation
- Over 200 errors identified, corresponding entries removed
- Over 13,400 compounds removed

6. Vowel length

- Partly a distinctive feature: /ama/ vs. /a ma/ (amma vs ama)
- Follows well defined rules for simple words
- More vague for derivations and compounds
- No changes made to the current (partly inconsistent) length

7. G2P Alignment

- Automatically generated g2p map
- Forced alignment performed using entries from the map as anchors
- Scarce mappings inspected for errors, 481 erroneous entries removed

G2P Transcription Experiments – Set up

- Two g2p models trained using the Sequitur toolkit, based on:
  - Cleaned IPD, ~40,000 entries (minus entries occurring in test/dev sets)
  - Raw IPD, ~65,000 entries (minus entries occurring in test/dev sets)
- Training parameters:
  - L=1 (max number of phonemes and graphemes in a grapheme)
  - M=5 (number of adjacent graphemes to consider)
  - Held-out set: 5%
- Development set, manually corrected: 700 entries
- Test set, manually corrected: 300 entries

G2P Transcription Experiments – Results WER/PER

<table>
<thead>
<tr>
<th></th>
<th>Raw IPD</th>
<th>Clean IPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>38%</td>
<td>5.65%</td>
</tr>
<tr>
<td>PER</td>
<td>25.33%</td>
<td>2.82%</td>
</tr>
</tbody>
</table>

G2P Transcription Experiments – Results Categories

<table>
<thead>
<tr>
<th>Error category</th>
<th>Raw IPD</th>
<th>Clean IPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel length</td>
<td>42</td>
<td>53</td>
</tr>
<tr>
<td>Postaspiration</td>
<td>46</td>
<td>19</td>
</tr>
<tr>
<td>Dialect</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Style/variation</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Consonant voiceness</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Other errors</td>
<td>38</td>
<td>9</td>
</tr>
<tr>
<td>SUM</td>
<td>144</td>
<td>87</td>
</tr>
</tbody>
</table>

Database example – Generate transcriptions by dialect

```
<id_12345>

dictionary
word | gramm. inf | pronunciation | variation | compounds
banka | nom,sg,gen | /paurkja/ | STANDARD | banka, bök

Generate North-Icelandic pronunciation:
input: bankalân transcr: /paurkalaun/
```

Future work, Resources

- Extend the database: compounds, dialect variations
- Implement transcription generation from database
- Post-processing steps for vowel length

Resources on GitHub:
https://github.com/Adamin/SLT2018

1 https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html