1. Overview

Motivation
- Quality of pronunciation dictionary is important for Speech Recognition
- g2p models might be of different quality depending on training data

Goal of Work
- Creation of pronunciation dictionaries for new languages and domains rapidly and economically based on statistical grapheme-to-phoneme (g2p) models

Goals of this particular study
- Comparison of g2p models [Bisani and Ney, 2008] between:
  - Languages: English (en), German (de), Polish (pl), Spanish (es), Czech (cs), French (fr)
  - Different training data quality:
    1. GlobalPhone word-phoneme pairs (successfully applied to LVCSR): **GP**
    2. Wiktionary word-phoneme pairs (provided by Internet community): **wikt**

Evaluation criteria:
- **Consistency check** (with Phoneme Error Rate (PER))
  - Generalization ability of the g2p models
  - Consistency within each pronunciation dictionary
  - Comparison to validated GlobalPhone pronunciation dictionary
- **Complexity check**
  - g2p model sizes (number of non-pruned 6-grams plus their backoff scores)
  - Automatic Speech Recognition (ASR) performance

Word error rate using pronunciations generated with the g2p models

2. Evaluation of g2p Models: Consistency and Complexity

- 6-fold cross validation for consistency and complexity check, evaluation on 30% of respective dictionary
- Standard deviation in consistency less than 1% PER with only 1k phoneme tokens (with corresponding graphemes)
  (Trend to smaller deviations with more training material)
- **GP** consistency: Large range of PER (pl, bg, cs, es, ru < fr, hr, pt, de < en)
  - PER varies with amount of training data between 100 and 10k phoneme tokens (with corresponding graphemes)
  - 15k phoneme tokens necessary for reasonable results per language,
  - When automatically creating pronunciations based on Wiktionary (trained with only 5k phoneme tokens)
    - Czech (PER 3.7%): each 27th phoneme
    - French (PER 6.4%): each 16th phoneme
    - Spanish (PER 7.6%): each 13th phoneme
    to be changed to meet the validated quality of GlobalPhone

3. Evaluation of g2p Models: ASR Performance

<table>
<thead>
<tr>
<th>GlobalPhone (base form)</th>
<th>GlobalPhone g2p (1-best)</th>
<th>Wiktory g2p (1-best)</th>
<th>GlobalPhone (GP) Consistency (PER)</th>
<th>Wiktory (wiktOnGP + wikt2) Consistency (PER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>15.59</td>
<td>17.58</td>
<td>18.77</td>
<td>2.41</td>
</tr>
<tr>
<td>de</td>
<td>16.71</td>
<td>16.50</td>
<td>16.81</td>
<td>10.21</td>
</tr>
<tr>
<td>en</td>
<td>18.15</td>
<td>18.15</td>
<td>28.86</td>
<td>12.83</td>
</tr>
<tr>
<td>es</td>
<td>12.25</td>
<td>12.99</td>
<td>12.82</td>
<td>3.99</td>
</tr>
<tr>
<td>fr</td>
<td>20.91</td>
<td>22.68</td>
<td>23.79</td>
<td>5.28</td>
</tr>
<tr>
<td>pl</td>
<td>15.51</td>
<td>15.78</td>
<td>17.21</td>
<td>0.36</td>
</tr>
</tbody>
</table>

- Use **GP** and **wikt** g2p models trained with 30k phoneme tokens and corresponding graphemes to reflect saturated g2p model consistency (5k and 10k for cs and es)
- Replace pronunciations in dictionaries of GlobalPhone ASR systems with pronunciations generated with g2p models
- Reasonable performance degradations given the cost and time efficient generation process

**Wiki g2p model complexity**
- Model complexity keeps increasing for larger amounts of data but this has minor impact on quality