# GRAPHEME-TO-PHONEME MODEL GENERATION FOR INDO-EUROPEAN LANGUAGES

*Tim Schlippe, Sebastian Ochs, Tanja Schultz*

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany

## ABSTRACT

In this paper, we evaluate grapheme-to-phoneme (g2p) models among languages and of different quality. We created g2p models for Indo-European languages with word-pronunciation pairs from the *GlobalPhone* project and from *Wiktionary* [1]. Then we checked their quality in terms of consistency and complexity as well as their impact on Czech, English, French, Spanish, Polish, and German ASR. While the *GlobalPhone* dictionaries were manually cross-checked and have been used successfully in LVCSR, *Wiktionary* pronunciations have been provided by the Internet community and can be used to rapidly and economically create pronunciation dictionaries for new languages and domains.

***Index Terms—*** web-derived pronunciations, multilingual speech recognition, pronunciation modeling

## 1. INTRODUCTION

With more than 6,900 languages in the world, the biggest challenge today is to rapidly port speech processing systems to new languages with low human effort and at reasonable cost. Especially, the creation of pronunciation dictionaries for speech processing systems can be time-consuming and expensive if they are manually written by language experts. The World Wide Web has been increasingly used as a text data source for rapid adaptation of ASR (Automatic Speech Recognition) systems and initial investigations to leverage off available pronunciations have been described [2][3]. In [2], we automatically retrieved pronunciations in terms of the International Phonetic Alphabet (IPA) [4] from *Wiktionary* [1], a multilingual wiki-based open content dictionary. Based on these, we enriched existing pronunciation dictionaries and analyzed their impact as pronunciation variants on LVCSR. Additionally, the g2p correspondences from the web-derived word-pronunciation pairs can be used to build statistical g2p models. These models can be used to generate pronunciations for out-of-vocabulary (OOV) words or to produce pronunciation variants. However, bad pronunciations in the training dictionary may decrease the quality of the acoustic models. Bad pronunciations in the decoding dictionary can also result in higher word error rates. To achieve optimal ASR performances, we need to ensure to use dictionaries which have been produced with high-quality g2p models – especially, if we use word-pronunciation pairs from the World Wide Web without a cross-check of language experts to build data-driven g2p models. For our quality analysis of pronunciations provided by the Internet community (*Wiktionary*) and validated ones (*GlobalPhone*), we built g2p models for Indo-European languages from 6 *Wiktionary* editions and 10 *GlobalPhone* dictionaries. *GlobalPhone* dictionaries had been created in a rule-based fashion and were manually cross-checked to reach professional quality [5]. First we check the g2p model consistency. For that, we built g2p models with increasing amounts of word-pronunciation pairs from *GlobalPhone* and *Wiktionary* as training material. We applied them to test sets from the respective source and computed the phoneme error rate (PER) to the original pronunciations. Furthermore, we evaluate the *Wiktionary* g2p models on the *GlobalPhone* test sets to investigate if the web-derived data meets the quality of validated dictionaries. Then we select g2p models which had all been trained with a comparable number of training material. With these, we investigate their relations among g2p consistency, complexity and their usage for ASR. For the ASR experiments, we replaced the pronunciations in the dictionaries of six *GlobalPhone* speech recognizers (Czech, English, French, Spanish, Polish, and German) and investigated the change in performance by using exclusively pronunciations generated from *Wiktionary* and *GlobalPhone* g2p models for training and decoding.

## 2. RELATED WORK

[3] retrieve English pronunciations from the World Wide Web and compare those to the *Pronlex* dictionary[1]. [6] and [7] consider g2p accuracy as an indicator of dictionary consistency. [6] compare the consistency of dictionaries through a ratio between the entropy of graphones (joint units of graphemes and corresponding phonemes) and their mutual information. [7] and [8] apply the following technique: They analyze the consistency of dictionaries with an n-fold cross validation where a part of the dictionary is used as training data to extract g2p rules and another part as test data to verify the rules. For

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. The authors would like to thank Maximilian Bisani for useful comments.

[1] CALLHOME American English Lexicon, LDC97L20.

g2p conversion, different methods are applied: Knowledge-based approaches with rule-based conversion systems were developed which can typically be expressed as finite-state automata [9] [10]. Often, these methods require specific linguistic skills and exception rules formulated by human experts. In contrast to knowledge-based approaches, data-driven approaches are based on the idea that, given enough examples, it should be possible to predict the pronunciation of unseen words purely by analogy. The benefit of the data-driven approach is that it trades the time- and cost-consuming task of designing rules, which requires linguistic knowledge, for the much simpler one of providing example pronunciations. [11] proposes a data-driven approach with heuristical and statistical methods. We use Sequitur G2P, a data-driven g2p converter developed at RWTH Aachen University which works with joint-sequence models [12]. As in [13], we evaluate the quality and complexity of the g2p models over increasing amount of data.

## 3. PRONUNCIATION EXTRACTION FROM WIKTIONARY

To accumulate training data for g2p models, we downloaded dumps of 6 *Wiktionary* editions (cs, de, en, es, fr, pl) for which we hold dictionaries from the *GlobalPhone* database and parsed them for IPA notations. We searched for strings which contain at least one character in the Unicode range between 0250 and 02AF surrounded by delimiters such as "/ /", "[ ]", etc. This procedure allows a website-independent collection of pronunciations. Sometimes several IPA notations occur on a *Wiktionary* page – either for different languages or for pronunciation variants. Usually the first pronunciation belongs to the target language. Therefore we used only the first pronunciation, if multiple candidates exist. In German *Wiktionary* for example, only 67% of the detected pronunciations are tagged as pronunciations for German words. The remainder is for Polish (10%), French (9%), English (3%), Czech (2%), etc. For some websites, there is no information to which language the pronunciations belong. Therefore it can happen that such inappropriate pronunciations are collected and corrupt the g2p model accuracy. To save time and cost, it is important to discover corrupted models early and not only through high word error rates after a speech recognizer has been built with the resulting dictionary.

## 4. EVALUATION OF G2P MODELS

### 4.1. Experimental Setup

For our g2p model generation and evaluation, we used pronunciations from 10 *GlobalPhone* dictionaries and from the 6 *Wiktionary* editions. The *GlobalPhone* dictionaries contain words of national and international political and economic topics from national online newspapers. For comparison, we mapped IPA pronunciations from *Wiktionary* to *GlobalPhone* phonemes. As *GlobalPhone* dictionaries contain phonemes based on the IPA scheme, a mapping between IPA units obtained from *Wiktionary* and *GlobalPhone* units is trivial [5]. For our experiments, Sequitur G2P models with a maximum M-gram size of $M$=6 and a maximum graphone size of $L$=1 (0 or 1 grapheme combined with 0 or 1 phoneme per graphone) worked out to be best for our amount of training data [12].

### 4.2. Quality Criteria

Our experiments to investigate the quality of the pronunciation dictionaries fall into the three categories:

- Consistency Check:
  Generalization ability of the g2p models
  - Consistency within each dictionary
  - Comparison to validated dictionary
- Complexity Check:
  g2p model sizes (number of non-pruned 6-grams plus their backoff scores)
- ASR Performance:
  Word error rate using pronunciations generated with the g2p models

### 4.3. Consistency Check

Table 1 shows how we analyzed the consistency within the *GlobalPhone* dictionaries (*GP*) and the *Wiktionary* editions (*wikt*) as well as between *Wiktionary* and the human cross-checked *GlobalPhone* dictionaries (*wiktOnGP*). For *GP* and *wikt*, we built g2p models with increasing amounts of word-pronunciation pairs in the dictionaries. Then we applied these to words from the same dictionary and computed the phoneme error rate (PER) between the new and the original pronunciations. For *wiktOnGP*, we computed the PERs of pronunciations generated with *Wiktionary* g2p models and evaluated on the original *GlobalPhone* pronunciations to analyze how close we can get to validated pronunciations with *Wiktionary* g2p models.

To verify the pronunciation quality, we performed a 6-fold cross validation as follows: For each *Wiktionary* edition and each *GlobalPhone* dictionary, we randomly selected 30% of the total number of word-pronunciation pairs for testing. From the remainder, we extracted increasing amounts

| | Train | Test |
|---|---|---|
| *GP* | GlobalPhone | GlobalPhone |
| *wikt* | Wiktionary | Wiktionary |
| *wiktOnGP* | Wiktionary | GlobalPhone |

**Table 1**. Consistency check setup.

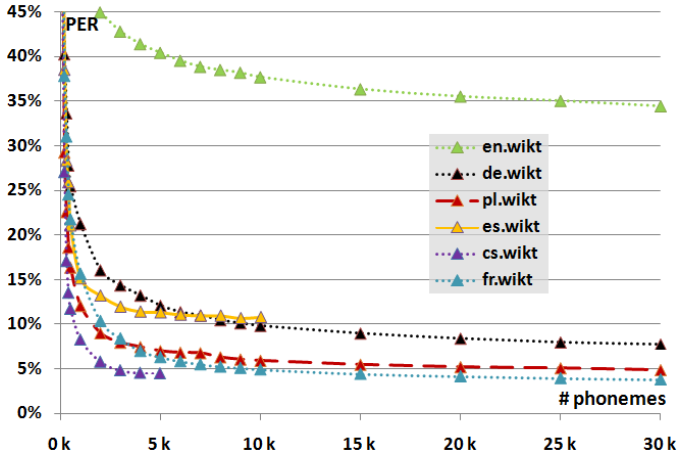


**Fig. 1**. Consistency of *wikt*.

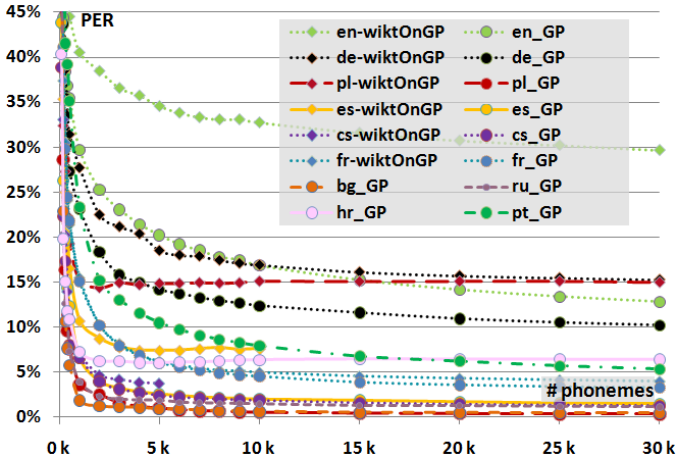


**Fig. 3**. Consistency of *GP* and *WiktOnGP*.

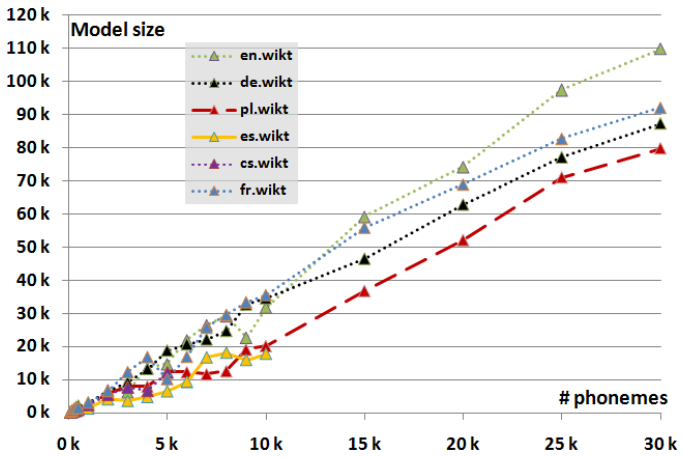


**Fig. 2**. Wiktionary g2p model complexity.

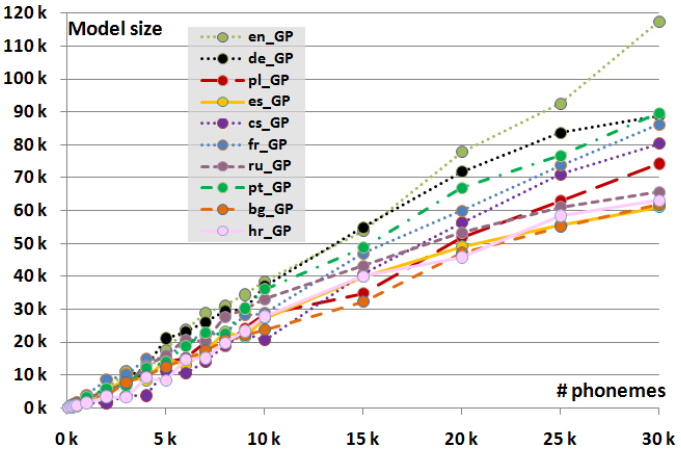


**Fig. 4**. GlobalPhone g2p model complexity.

of entries based on their accumulated phoneme count and used them for training the g2p models in each fold. Fig. 1 and 3 demonstrate differences in g2p consistency among the languages. Comparing both figures shows that for Czech, English, French, Polish, and Spanish, *GP* was more consistent internally than *wikt* except for German. *GP* came closer to the validated *GlobalPhone* pronunciations than *wiktOnGP* for all languages. Fig. 3 reveals noticeable differences between the PERs of *GP* and *wiktOnGP*. For Czech, English, and Spanish, the PERs of *wiktOnGP* are located between *wikt* and *GP* of the same language. However, for German, French, and Polish, the dictionaries were consistent internally but did not fit together in the cross-dictionary evaluation. Fig. 1 and 3 show variations in PERs for amounts of training data between 100 and 7k phonemes. For more than 7k phonemes, the PERs decrease with more training data. But we learn that for the 10 languages word-pronunciation pairs containing 15k phonemes were sufficient to have constant quality as the curves start to saturate at 15k phonemes for all 10 languages.

### 4.4. Complexity Check

For the second category, we investigated the complexity of the g2p models over training data and among languages and compared the complexity change to the consistency change. Fig. 2 and 4 show the increase in complexity of the g2p models with the increase of training material between 100 and 30k phonemes with corresponding graphemes. A comparison of Fig. 1 and 3 with Fig. 2 and 4 indicates that although the consistency saturates at 15k phonemes, the model complexity keeps increasing for larger amounts of training data. However, this has minor impact on quality in terms of consistency.

For the ASR performance checks, we decided to select g2p models which were trained with 30k phonemes and their corresponding graphemes to reflect a saturated g2p model consistency. 30k phonemes are contained in all *GlobalPhone* dictionaries and in most of the 6 *Wiktionary* editions. For the Czech and Spanish *Wiktionary* and *GlobalPhone* g2p models, we used the maximum number of phonemes (5k and 10k) which we could find in *Wiktionary*.

| | GlobalPhone (base form) | GlobalPhone g2p (1-best) | Wiktionary g2p (1-best) | GlobalPhone (with variants) | GlobalPhone g2p (n-best) | Wiktionary g2p (n-best) | GlobalPhone (*GP*) Consistency (PER) | Wikt. (*wiktOnGP* / (*wikt*)) Consistency (PER) | |
|---|---|---|---|---|---|---|---|---|---|
| cs | 15.59 | 17.58 | 18.72 | 15.62 | 18.06 | 19.32 | 2.41 | 3.75 | (4.47) |
| de | 16.71 | 16.50 | 16.81 | 17.11 | 17.06 | 17.40 | 10.21 | 15.27 | (7.74) |
| en | 14.92 | 18.15 | 28.86 | 11.52 | 18.66 | 37.82 | 12.83 | 29.65 | (34.44) |
| es | 12.25 | 12.59 | 12.82 | 11.97 | 12.32 | 12.81 | 1.99 | 7.63 | (10.78) |
| fr | 20.91 | 22.68 | 25.79 | 20.41 | 22.68 | 25.17 | 3.28 | 4.02 | (3.77) |
| pl | 15.51 | 15.78 | 17.21 | 14.98 | 15.68 | 17.34 | 0.36 | 15.02 | (4.86) |

**Table 2**. WERs (%) of systems with dictionaries built completely with g2p generated pronunciations.

### 4.5. ASR Performance

Finally, we analyzed if we can use the pronunciations generated with our *Wiktionary* and *GlobalPhone* g2p models in ASR. Furthermore we were interested if our information about the pronunciation quality correlates with their impact on ASR performance. For it, we replaced the pronunciations in the dictionaries of six *GlobalPhone* ASR systems with pronunciations generated with *Wiktionary* and *GlobalPhone* g2p models. Then we trained and decoded the systems completely with those pronunciation dictionaries. First, we built and decoded ASR systems with dictionaries where only the most likely (1-best) pronunciation for each *GlobalPhone* word was produced with our g2p models. We compared these to *GlobalPhone* systems which were also limited to the first pronunciation (base form). Furthermore, we established systems with dictionaries where pronunciation variants (n-best) were also produced. For each word, we generated exactly the number of pronunciations with our models that occurs in the *GlobalPhone* dictionaries. The results of the ASR experiments together with the consistency results of the used g2p models are listed in Table 2. For all languages except for Spanish and French, the systems built with the 1-best g2p models performed better than those with the pronunciation variants. With the *Wiktionary* g2p models, we come close to the word error rates of the *GlobalPhone* systems for all languages but English. However, the *GlobalPhone* g2p systems performed slightly better which correlates with the *GP* and *wiktOnGP* consistency. We explain the high word error rates in English with a difficult g2p correspondance and corrupted training material from *Wiktionary*.

## 5. CONCLUSION AND FUTURE WORK

We have investigated the g2p model generation for Indo-European languages with pronunciations from 6 *Wiktionary* editions and 10 *GlobalPhone* dictionaries. We analyzed and compared their quality with regard to consistency and complexity and detected a saturation at 15k phonemes with corresponding graphemes as training material. Using exclusively pronunciations generated from *Wiktionary* and *GlobalPhone* g2p models for ASR training and decoding resulted in reasonable performance degradations given the cost and time efficient generation process. The severeness of degredation correlates with the g2p consistency. However, obtaining pronunciations generated with *Wiktionary* g2p models will lead to less manual editing effort than starting to write pronunciation dictionaries from scratch. A linguist or native speaker merely has to change in average each 27th phoneme for Czech (PER 3.8%), each 25th for French (PER 4.0%), and each 13th for Spanish (PER 7.6%) to meet validated *GlobalPhone* quality after applying the *Wiktionary* models from our ASR experiments. The worst effort reduction appears for English, where each third phoneme (PER 29.7%) has to be changed. In the future, optimization of our pronunciation extraction and filtering methods should improve the g2p models. Furthermore, we may integrate a speech synthesis component into a dictionary building process for accelerated and interactive editing of improper phonemes.

## 6. REFERENCES


[1] "Wiktionary - a wiki-based open content dictionary," Website, http://www.wiktionary.org.

[2] T. Schlippe, S. Ochs, and T. Schultz, "Wiktionary as a Source for Automatic Pronunciation Extraction," in *Interspeech*, 2010.

[3] A. Ghoshal, M. Jansche, S. Khudanpurv, M. Riley, and M. Ulinski, "Web-derived Pronunciations," in *ICASSP*, 2009.

[4] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[5] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *ICSLP*, 2002.

[6] M. Wolff, M. Eichner, and R. Hoffmann, "Measuring the Quality of Pronunciation Dictionaries," in *PMLA*, 2002.

[7] M. Davel and E. Barnard, "Developing Consistent Pronunciation Models for Phonemic Variants," in *Interspeech*, 2006.

[8] M. Davel and O. Martirosian, "Pronunciation Dictionary Development in Resource-Scarce Environments," in *Interspeech*, 2009.

[9] R. M. Kaplan and M. Kay, "Regular Models of Phonological Rule Systems," in *Computational Linguistics*, 1994.

[10] A. W. Black, K. Lenzo, and V. Pagel, "Issues in Building General Letter to Sound Rules," in *ESCA Workshop on Speech Synthesis*, 1998.

[11] S. Besling, "Heuristical and Statistical Methods for Grapheme-to-Phoneme Conversion," in *Konvens*, 1994.

[12] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, 2008.

[13] J. Kominek, "TTS From Zero - Building Synthetic Voices for New Languages," *Doctoral Thesis*, 2009.