Wiktionary as a Source for Automatic Pronunciation Extraction

Tim Schlippe, Sebastian Ochs, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany
tim.schlippe@kit.edu, sebastian.ochs@student.kit.edu, tanja.schultz@kit.edu

Abstract

In this paper, we analyze whether dictionaries from the World Wide Web which contain phonetic notations, may support the rapid creation of pronunciation dictionaries within the speech recognition and speech synthesis system building process. As a representative dictionary, we selected *Wiktionary* [1] since it is at hand in multiple languages and, in addition to the definitions of the words, many phonetic notations in terms of the International Phonetic Alphabet (IPA) are available. Given word lists in four languages English, French, German, and Spanish, we calculated the percentage of words with phonetic notations in *Wiktionary*. Furthermore, two quality checks were performed: First, we compared pronunciations from *Wiktionary* to pronunciations from dictionaries based on the *GlobalPhone* project, which had been created in a rule-based fashion and were manually cross-checked [2]. Second, we analyzed the impact of *Wiktionary* pronunciations on automatic speech recognition (ASR) systems.

French *Wiktionary* achieved the best pronunciation coverage, containing 92.58% phonetic notations for the French *GlobalPhone* word list as well as 76.12% and 30.16% for country and international city names. In our ASR systems evaluation, the Spanish system gained the most improvement from *Wiktionary* pronunciations with 7.22% relative word error rate reduction.

Index Terms: pronunciation dictionary, rapid language adaptation, automatic speech recognition, crowdsourcing

1. Introduction

With some 6,900 languages in the world, data resources such as text files, transcribed speech or pronunciation dictionaries are only available in the most economically viable languages. Over the past years, the World Wide Web has been increasingly used as a text data source for rapid adaptation of ASR systems to new languages at low cost, e.g. websites are crawled to collect texts that are used to build language models. Moreover, prompts which might be read by native speakers to receive transcribed audio data, are extracted from the crawled text [3]. The creation of pronunciation dictionaries can be time consuming and expensive if they are manually produced by language experts. Thus our intention was to research if phonetic notations from the World Wide Web can be used to build pronunciation dictionaries from scratch or at least enhance existing ones.

If phonetic notations can be extracted together with the corresponding written words from the World Wide Web, the following scenario comes within reach: Text from the ASR system's target domain is crawled and a text normalization is performed automatically. The vocabulary of the normalized crawled text is extracted, and on the basis of phonetic notations from the World Wide Web, a pronunciation dictionary with that vocabulary is created. Subsequently, a language model is built on the collected text. Thereby, dictionary and language model in a new domain or language are generated without manual effort.

Our *Rapid Language Adaptation Toolkit (RLAT)* with its web-based interface is an ongoing effort towards that goal. It aims to reduce the human effort involved in building speech processing systems for new languages and domains. Innovative tools enable novice and expert users to develop speech processing models, such as acoustic models, pronunciation dictionaries, and languages models, to collect appropriate speech and text data for building these models, and to evaluate the results [4]. In this paper, we describe the extension of *RLAT* to extract pronunciations from *Wiktionary*.

Wiktionary is a wiki-based open content dictionary, available in many languages and checked by a big community frequently and carefully. It contains phonetic notations written in the International Phonetic Alphabet (IPA). The IPA, devised by the International Phonetic Association, is a standardized representation of the sounds of spoken language [5].

For our experiments, English, French, German, and Spanish were chosen as representative languages, selected from *GlobalPhone* [6]. *GlobalPhone* is a database collection that provides transcribed speech data for the development and evaluation of large speech processing systems in the most widespread languages of the world.

2. Previous Work

In the field of speech processing, the World Wide Web has been used as a data source for improving the language model probability estimation as well as for obtaining additional training material [7].

Furthermore, several approaches to automatic dictionary generation have been introduced in the past. Besling proposes heuristical and statistical methods [8]. Black et al. apply letter-to-sound rules for the dictionary production [9]. Often, these methods still require post-editing by a human expert or leverage off another manually generated pronunciation dictionary.

In [10], the *Lexicon Learner* presents words to the user. The user does not have to be a language expert to provide the pronunciations. Each word is accompanied by a suggested pronunciation, along with a synthesized wavefile. The prediction is based on letter-to-sound rules that the system infers from the user's answers, which are updated after each additional word. The rules are seeded during an initialization stage in which the *Lexicon Learner* asks the user for the phoneme most commonly associated with each letter. A similar dictionary creation process that combines machine learning with minimal human intervention was proposed by [11].

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

In [12], English phonetic notations in IPA and ad-hoc transcriptions are retrieved from the World Wide Web, and the pronunciations are compared to the *Pronlex* dictionary[1].

Our goal is to analyze a multilingual database in the World Wide Web such as *Wiktionary*, with a focus on quantity and quality of the pronunciations which includes the impact of the web-derived pronunciations on existing baseline ASR systems.

Our experiments to research the suitability of phonetic notations from the World Wide Web for pronunciation dictionary creation fall into the two categories:

1. *Quantity check*:
   (a) Given a word list, what is the percentage of words for which phonetic notations are found in a complete IPA representation?
2. *Quality check*
   (a) How many pronunciations derived from *Wiktionary* are identical to existing *GlobalPhone* pronunciations?
   (b) How does adding *Wiktionary* pronunciations impact the performance of ASR systems?

The word lists for the quantity check are taken from the *GlobalPhone* dictionaries plus a collection of international city and country names to investigate the coverage of proper names. Proper names can be of diverse etymological origin and can surface in another language without undergoing the process of assimilation to the phonetic system of the new language [13]. Therefore, pronunciations of proper names are of particular importance as they are difficult to generate with letter-to-sound rules.

3. Data

3.1. Wiktionary

Wiktionary is a collaborative project for creating a free lexical database in multiple languages, where meanings, etymologies and many phonetic notations in IPA standard are available. A big community verifies the entries frequently and carefully. The language diversity and size of *Wiktionary* is indicated in Table 1.

No.	Language	"Good" Entries	Admins	Active Users
1	French	1,786k	21	286
2	English	1,770k	100	1047
3	Lithuanian	542k	4	14
4	Turkish	268k	6	50
5	Chinese	257k	9	31
6	Russian	246k	6	139
7	Vietnamese	229k	5	31
8	Ido	171k	2	13
9	Polish	165k	25	79
10	Portuguese	156k	6	112

Table 1: *The ten largest Wiktionary language editions (July 2010)* [14].

3.2. GlobalPhone

In the *GlobalPhone* project, pronunciation dictionaries in 20 languages have been established. Widely read national newspapers available on the World Wide Web were selected as resources. Texts from national and international political and economic topics restrict the vocabulary.

	English	French	German	Spanish
Vocab size	58k	122k	38k	30k
Audio train	15.4 h	24.9 h	14.9 h	17.5 h
Audio test	0.5 h	2.0 h	1.5 h	1.6 h
OOV rate	0.72%	0.00%	0.01%	0.01%

Table 2: *Vocabulary size, length of audio data and OOV rates for our ASR systems.*

Table 2 shows the vocabulary sizes of the dictionaries we use to check the coverage of domain words and to build the baseline ASR systems, the size of the training and test audio data as well as the out-of-vocabulary (OOV) rates in the test sets. For French, we employed a French dictionary developed within the Quaero Programme which contains more domain words than the original *GlobalPhone* dictionary.

As *GlobalPhone* dictionaries contain phonetic notations based on the IPA scheme, a mapping between IPA units obtained from *Wiktionary* and *GlobalPhone* units is trivial [2].

4. Experiments and Results

4.1. Automatic Dictionary Extraction from Wiktionary

Our Automatic Dictionary Extraction Tool takes a vocabulary list with one word per line. For each word, the matching *Wiktionary* page is looked up. If the page cannot be found, we iterate through all possible combinations of upper and lower case. However, for certain languages such as German, it may happen that different casing leads to slightly different pronunciations (e.g. "Weg" vs. "weg"). Each web page is saved and parsed for IPA notations. On a *Wiktionary* page, sometimes several IPA notations occur, either for different languages or for pronunciations variants. Certain keywords in the context of the IPA notations help us to decide which phonetic notation to extract. For simplicity, we only use the first phonetic notation, if multiple candidates exist. Our tool then outputs the detected IPA notations for the input vocabulary list and reports back those words for which no pronunciation could be found.

The Automatic Dictionary Extraction Tool is a new component of our *Rapid Language Adaptation Toolkit (RLAT)*, an extension of the *SPICE* system.

4.2. Quantity Check

With the vocabulary of our *GlobalPhone*-based dictionaries as input for our Automatic Dictionary Extraction Tool, we analyzed the coverage of pronunciations of words from our domain of political and economic topics. With a list of international city and country names, we additionally investigated the coverage of pronunciations for a subset of proper names. The city and country names were translated into the respective language for each check.

The results of our quantity checks are illustrated in Figure 1 and 2. French *Wiktionary* outperformed the other languages, with a 93.27% match for the French vocabulary list, 76.12% match for country names and 30.16% match for city names.

We were surprised that in our quantity check English *Wiktionary* performed much worse than the French one, despite having a comparable amount of pages. Therefore we analyzed how many pages include IPA notations. We can see in Table 3 that French contains almost six times as many pages with pronunciations as English. Even the coverage of pronunciations

[1] CALLHOME American English Lexicon, LDC97L20.

for the words from the *Pronlex* dictionary (22.67%) showed that significanty less English pronunciations can be found in the English *Wiktionary*. We also discovered that many pronunciations in a given *Wiktionary* are for foreign languages. In German *Wiktionary* for example, only 67% of the detected pronunciations are for German words, the rest being for Polish (10%), French (9%), English (3%), Czech (2%), Italian (2%), etc. In total, pronunciations for more than 100 languages were detected in German *Wiktionary*, most in very small quantities.

The results for our domain word lookups show that for French (114k found) we could build a nearly complete pronunciation dictionary for our domain using exclusively *Wiktionary* pronunciations. However, for the other three languages the out-of-vocabulary rate would be too high to get a reasonable performance. Therefore, we decided not to build new pronunciation dictionaries solely from *Wiktionary* but instead to enrich existing dictionaries with the retrieved pronunciations in the ASR evaluations, described in Section 4.3.2.

4.3. Quality Check

As we cannot ensure that the quality of pronunciations from the World Wide Web matches the quality of our *GlobalPhone*-based dictionaries, we analyzed how many phonetic notations derived from *Wiktionary* are identical to our existing pronunciations. Additionally, we built several baseline ASR systems based on the *GlobalPhone* data to investigate the impact of the pronunciations from *Wiktionary* on ASR performance.

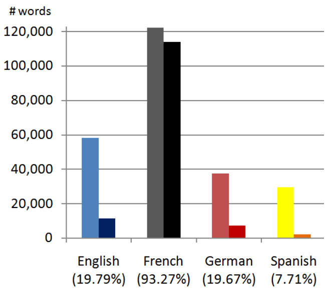

Figure 1: *Searched and found prons. for domain words.*

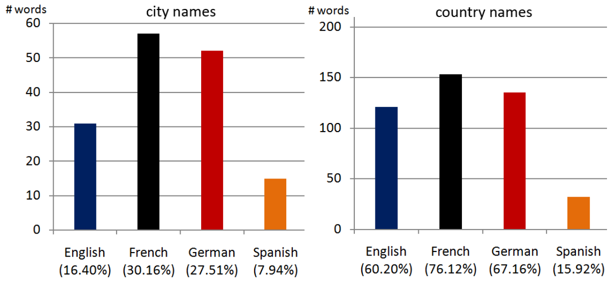

Figure 2: *Found pronunciations for 189 international city names and 201 country names.*

Language	# pages	# pages with prons	% pages with prons
French	1,786k	912k	51.1%
English	1,770k	159k	9.0%
German	110k	48k	43.6%
Spanish	48k	8k	16.7%

Table 3: *Quantity of Wiktionary pages containing pronunciations.*

4.3.1. Coverage of Identical Pronunciations

We did the following preparations for each language: We extracted those domain words out of the *GlobalPhone*-based dictionaries for which *Wiktionary* provides pronunciations. Then we extracted the phonetic notation of each word from *Wiktionary*. To have comparable phonetic notations, a mapping from *GlobalPhone* pronunciation units to IPA units was performed. Since *GlobalPhone* units are based on the IPA scheme such a mapping is straightforward. Finally, we compared the *Wiktionary* pronunciation for each word to all *GlobalPhone* pronunciations of the same word and counted the identical ones. As in our quantity checks, French *Wiktionary* showed most identical pronunciations (73.99%). The results are summarized in Table 4.

No.	Language	# prons.	% equal	# new
1	French	114k	74%	30k
2	Spanish	2k	50%	1k
3	German	7k	28%	5k
4	English	12k	26%	9k

Table 4: *Amount of compared pronunciations, percentage of identical ones and amount of new pronunciation variants.*

4.3.2. Impact on ASR Performance

Since the identical pronunciations from *Wiktionary* contribute nothing new to our existing dictionaries, we discarded them and added all remaining pronunciations as new variants (Table 4). With these new enriched dictionaries, a training and decoding was performed (*System1*). As shown in Table 5, the new pronunciations did not always have a good influence on the performance of the ASR systems.

To evaluate the impact of the new pronunciations on the Linear Discriminant Analysis (LDA) of our systems, we compared the class separability measure introduced by [15] for the 43-dimensional LDA matrices of our baseline systems to our systems after training with the additional *Wiktionary* pronunciations. The results which are illustrated in Table 5 correlate with the word error rate (WER).

	WER baseline	WER System1	rel. improv.	rel. improv. class sep.
French	23.43%	23.25%	0.79%	+1.24%
English	21.51%	22.46%	-4.44%	-0.95%
German	21.60%	21.67%	-0.31%	-0.50%
Spanish	14.68%	14.42%	1.76%	+0.50%

Table 5: *Impact of using all Wiktionary pronunciations for training and decoding (System1).*

For French and Spanish, we observe a decrease in WER for *System1*. The class separability demonstrates that training with the new pronunciations improved our LDA. However, we notice a degradation for the English and German systems. We explain this with faulty pronunciations from *Wiktionary* and with pronunciations not matching our training and test data in speaking style (accent etc.) which may cause the worse class separability. Also, we assume that the consistency of the dictionaries decreases by merging pronunciations of different sources.

Assuming that speakers in training and test set use similar speaking styles and vocabulary, and that our training process automatically selects the most likely pronunciations, we might see an improvement if we remove all *Wiktionary* pronunciations from the *System1* dictionaries that were not used in training. Table 6 illustrates the portion of *Wiktionary* pronunciations which are used in our training processes.

	# wikt prons	% wikt prons
French	3,000	10.11%
English	845	9.86%
German	1,439	27.02%
Spanish	259	22.90%

Table 6: *Amount and percentage of Wiktionary pronunciation selected in training.*

As described, we reduced the dictionaries in *System1* and performed another decoding (*System2*). With this approach, we were able to further improve the French and Spanish systems. The German system could be noticeably improved, even though it did not benefit from the training with added *Wiktionary* pronunciations in *System1* before and its LDA matrix developed worse class separability. The English system performed even worse than before. This shows that our previous assumption about the similarity of training and test set in speaking style and vocabulary was incorrect for this particular system. It seems that the test set makes use of pronunciations that do not occur in the training. The results of *System2* are listed in Table 7.

	WER baseline	WER System2	relative improvement
French	23.43%	23.16%	1.17%
English	21.51%	23.39%	-8.76%
German	21.60%	21.07%	2.44%
Spanish	14.68%	13.62%	7.22%

Table 7: *Impact of using only those Wiktionary prons. in decoding that were chosen in training (System2).*

5. Conclusion and Future Work

In this paper, an economical data source from the World Wide Web that may support the rapid pronunciation dictionary creation has been proposed. With our Automatic Dictionary Extraction Tool, we developed a system which automatically extracts phonetic notations in IPA from *Wiktionary*.

We reported various results for four languages concerning quantity and quality checks. The quantity checks with lists of international cities and countries demonstrated that even proper names whose pronunciations might not be found in the phonetic system of a language are detectable together with their phonetic notations in *Wiktionary*. However, we did not gain similar results for all languages due to the different quantity and quality of the data sources.

Designers of ASR systems are supposed to ensure that they use exclusively pronunciations of good quality for dictionary creation. Using only the pronunciations for decoding which were used in the training process did not help to improve all systems. Therefore further investigations should be done here.

In the future, we plan to explore other language editions of *Wiktionary*. We intend to improve our detection and extraction to handle more than the first IPA notation, if multiple candidates exist on one page. Since we discovered that these additional notations can be pronunciation variants as well as pronunciations from other languages, we intend to improve our methods so that we can retrieve and categorize all available pronunciations including those not in the target language. Even if not enough pronunciations for a complete dictionary are found, one of our next tasks is to analyze whether existing IPA notations can still be useful for training grapheme-to-phoneme models of good quality.

We expect an increase of *Wiktionary*'s vocabulary and language coverage due to its growing community. Finally, we are going to investigate other websites for pronunciations, such as *Wikipedia*.

6. Acknowledgements

The authors would like to thank Qingyue He for her support and Ngoc Thang Vu for useful discussions.

7. References

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

[2] T. Schultz, "GlobalPhone: A multilingual speech and text database developed at Karlsruhe University," in *Proceedings of the ICSLP*, 2002, pp. 345–348.

[3] T. Schultz, A. W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[4] A. W. Black and T. Schultz, "Rapid language adaptation tools and technologies for multilingual speech processing," in *Proceedings of the ICASSP*, Las Vegas, USA, 2008.

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

[6] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1-2, pp. 31–51, 2001.

[7] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[8] S. Besling, "Heuristical and statistical methods for grapheme-to-phoneme conversion," in *Konvens*, Vienna, Austria, 1994.

[9] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proceedings of ESCA Workshop on Speech Synthesis*, Australia, 1998, pp. 77–80.

[10] J. Kominek and A. W. Black, "Learning pronunciation dictionaries: Language complexity and word selection strategies," in *Proceedings of the HLT Conference of the NAACL*, 2006, pp. 232–239.

[11] M. Davel and E. Barnard, "The efficient generation of pronunciation dictionaries: Human factors during bootstrapping," in *Proceedings of the 8th ICSLP*, Korea, 2004.

[12] A. Ghoshal, M. Jansche, S. Khudanpur, M. Riley, and M. Ulinski, "Web-derived pronunciations," in *Proceedings of the 2009 ICASSP*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4289–4292.

[13] A. F. Llitjós and A. W. Black, "Evaluation and collection of proper name pronunciations online," in *Proceedings of LREC2002*, Las Palmas, Canary Islands, 2002.

[14] "List of wiktionary editions, ranked by article count." [Online]. Available: http://meta.wikimedia.org/wiki/List_of_Wiktionaries

[15] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Proceedings of Interspeech*, 2007.