Cognitive Systems Lab

Karlsruhe Institute of Technology

Wiktionary as a source for Automatic Pronunciation Extraction

Tim Schlippe
Sebastian Ochs
Tanja Schultz

29 September 2010

Outline

1. Introduction

- World Wide Web (WWW) increasingly used as text data source for rapid adaptation of ASR systems to new languages and domains, e.g.
  - Crawl texts to build language models (LMs),
  - Extract prompts read by native speakers to receive transcribed audio data (Schultz et al. 2007)

- Creation of pronunciation dictionary
  - Usually produced manually or semi-automatically
  - Time consuming, expensive
  - Proper names difficult to generate with letter-to-sound rules

- Idea: Leverage off the internet technology and crowdsourcing

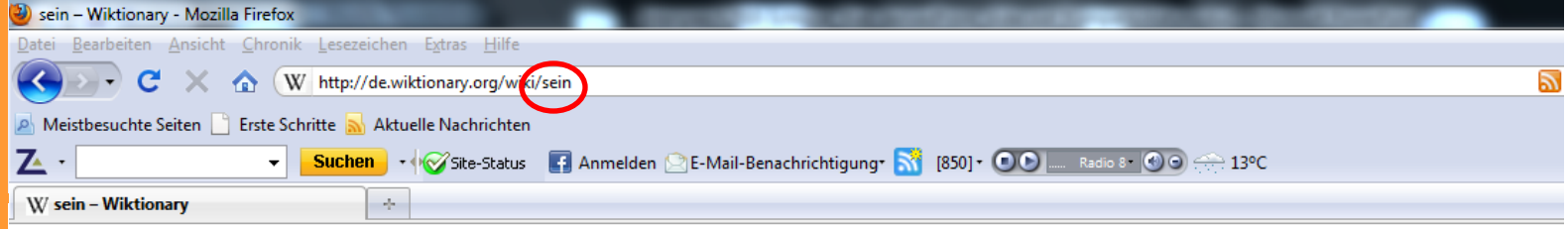
→ Is it possible to generate pronunciations based on phonetic notations found in the WWW?

2.1 Data – Wiktionary

- At hand in multiple languages
- In addition to definitions of words, many phonetic notations written in the International Phonetic Alphabet (IPA) are available
- Quality and quantity of entries dependent community and the underlying resources
- First *Wiktionary* edition: English in Dec. 2002, then: French and Polish in Mar. 2004

The ten largest *Wiktionary* language editions (July 2010) (http://meta.wikimedia.org/wiki/List of Wiktionaries)

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



Wikiwörterbuch

Wiktionary

['vɪkʃəˌnɛʁi], *n*

Das freie Wörterbuch

ein Wiki-basiertes freies Wörterbuch

Hauptseite
Themenportale
Zufällige Seite
Inhaltsverzeichnis

Mitarbeit
Eintrag erstellen
Autorenportal
Wunschliste
Literaturliste
Letzte Änderungen

Hilfe

Werkzeuge

In anderen Sprachen
العربية
Brezhoneg
Česky
Dansk
Ελληνικά
English

Eintrag Diskussion Lesen Bearbeiten Versic

Siehe auch: Sein

sein

sein (Deutsch) [Bearbeiten]

Hilfsverb [Bearbeiten]

Anmerkung:
Alle Verbindungen mit *sein* schreibt man nach neuer Rechtschreibung getrennt (da sein, weg sein, zusammen sein …).

Silbentrennung:
sein, Präteritum: war, Partizip II: ge·we·sen

Aussprache:
IPA: ['zaɪ̯n] (bin: [bɪn], bist: ['bɪst], ist: [ɪst], sind: ['zɪnt], seid: [zaɪ̯t]), Präteritum: ['vaːɐ̯], Partizip II: [gə'veːzən]

```
<dd><a href="/wiki/Hilfe:IPA" title="Hilfe:IPA">IPA</a>:
[<span class="ipa" style="padding: 0 1px;
text-decoration: none;">ˈzaɪ̯n</span>] (bin: [<span
class="ipa" style="padding: 0 1px; text-decoration:
none;">bɪn</span>], bist: [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">ˈbɪst</span>], ist: [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">ɪst</span>], sind: [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">ˈzɪnt</span>], seid: [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">zaɪ̯t</span>]), <span style="font-
size:95%;">Präteritum:</span> [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">ˈvaːɐ̯</span>], <span style="font-
size:95%;">Partizip II:</span> [<span class="ipa"
style="padding: 0 1px; text-decoration:
none;">gəˈveːzən</span>]</dd>
```

• • •

sein (Französisch) [Bearbeiten]

Substantiv, *m* [Bearbeiten]

Silbentrennung:
sein, Plural: seins

Aussprache:
IPA: [sɛ̃]
Hörbeispiele: sein, Plural: —

```
<dd><a href="/wiki/Hilfe:IPA" title="Hilfe:IPA">IPA</a>:
[<span class="ipa" style="padding: 0 1px;
text-decoration: none;">sɛ̃</span>]</dd>
```

2.2 Data – GlobalPhone

- For our experiments, we build ASR systems with *GlobalPhone* data for English, French, German, and Spanish
- In *GlobalPhone*, widely read national newspapers available on the WWW with texts from national and international political and economic topics were selected as resources
- Vocabulary size and length of audio data for our ASR systems:

	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

- *GlobalPhone* dictionaries

… had been created in rule-based fashion, manually cross-checked

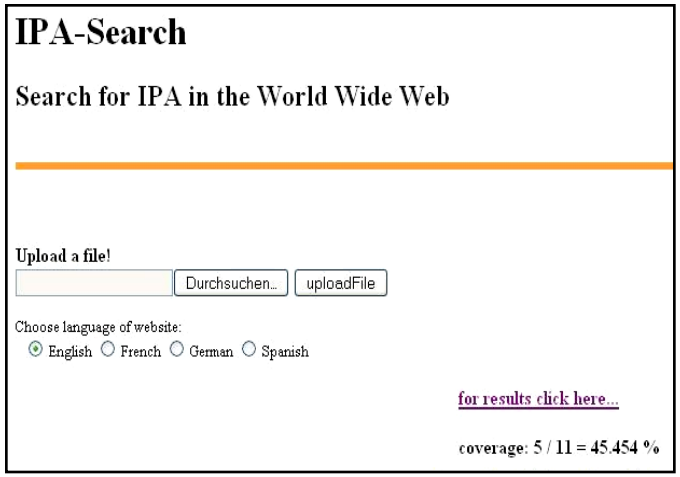
… contain phonetic notations based on IPA scheme

→ mapping between IPA units obtained from *Wiktionary* and *GlobalPhone* units is trivial (Schultz, 2002)

3. Experiments and Results

- Quantity Check:
  - Given a word list, what is the percentage of words for which phonetic notations are found in *Wiktionary*?
    - Quantity of pronunciations for *GlobalPhone* words
    - Quantity of pronunciations for proper names (e.g. New York)
- Quality Check:
  - How many pronunciations derived from *Wiktionary* are identical to existing *GlobalPhone* pronunciations?
  - How does adding *Wiktionary* pronunciations impact the performance of ASR systems?

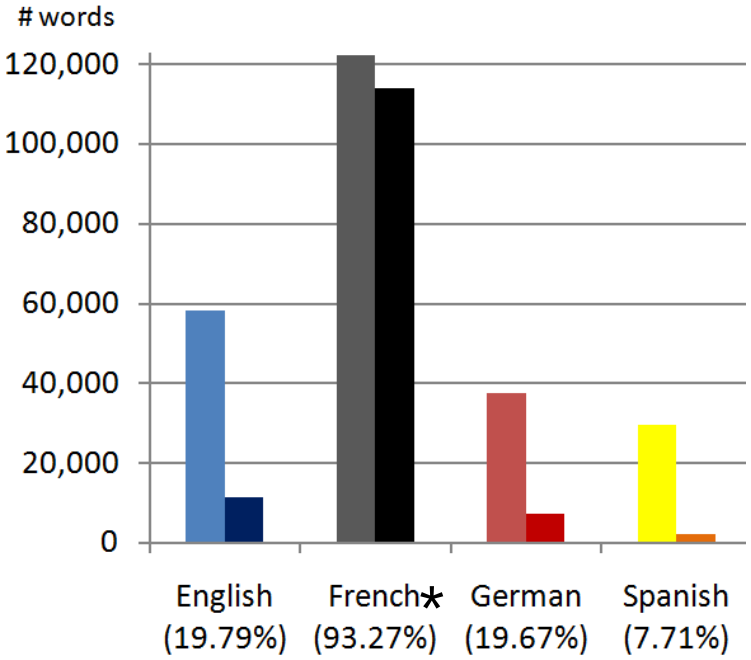
3.1 Experiments and Results – Extraction



- Manually select in which *Wiktionary* edition to search for pronunciations
- Our Automatic Dictionary Extraction Tool takes a vocab list with one word per line
- For each word, the matching *Wiktionary* page is looked up
  (e.g. http://fr.wiktionary.org/wiki/abandonner)
- If the page cannot be found, we iterate through all possible combinations of upper and lower case
- Each web page is saved and parsed for IPA notations:
  - Certain keywords in context of IPA notations help us to find the phonetic notation
    (e.g. `title="Annexe:Prononciation/français">` `<span class="API" title="prononciation API">/a.bɑ̃.dɔ.ne/</span>`)
  - For simplicity, we only use the first phonetic notation, if multiple candidates exist
  - Our tool outputs the detected IPA notations for the input vocab list and reports back those words for which no pronunciation could be found

3.2 Experiments and Results – Quantity Check

- Quantity of pronunciations for ***GlobalPhone* words**



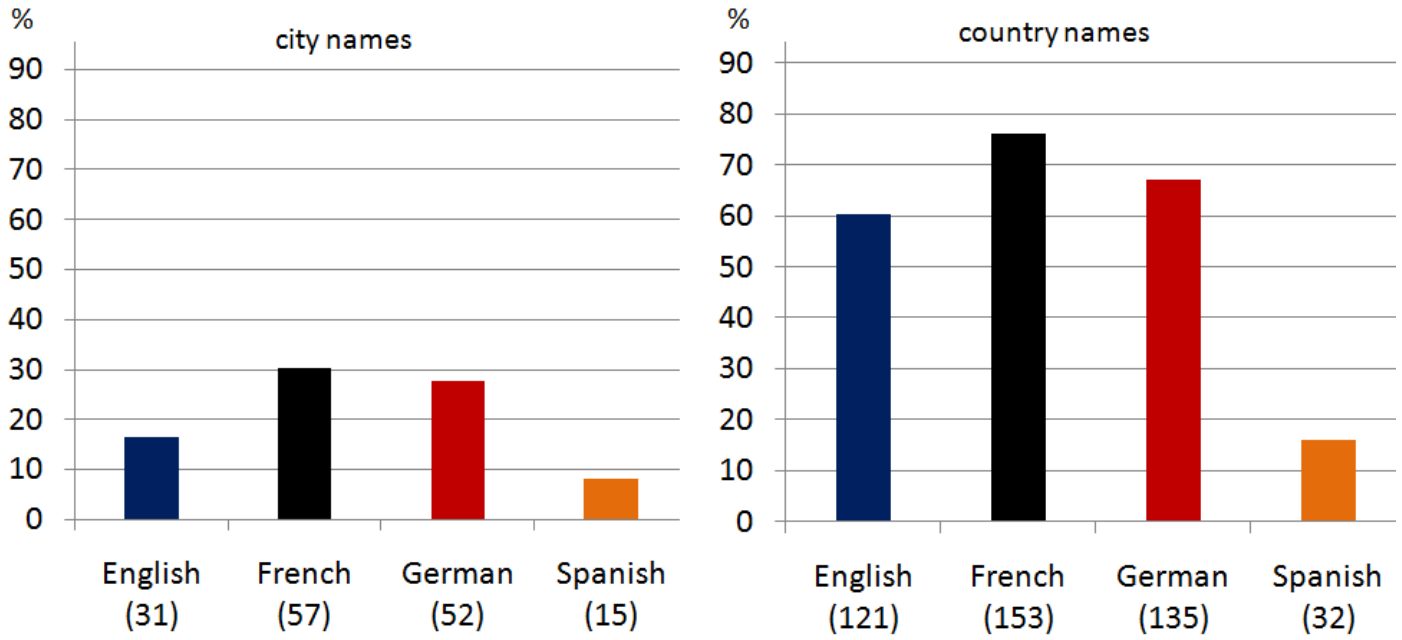
Searched and found pronunciations for words in the *GlobalPhone* corpora

*For French, we employed a word list developed within the Quaero Programme which contains more words than the original *GlobalPhone*

  - Morphological variants in the word lists could also be find in Wiktionary
  - French *Wiktionary* has highest match, possible explanations:
    - Strong French internet community (e.g. “Loi relative à l’emploi de la langue française “)
    - Several imports of entries from freely licensed dictionaries in *French Wiktionary (http://en.wikipedia.org/wiki/French_Wiktionary)*

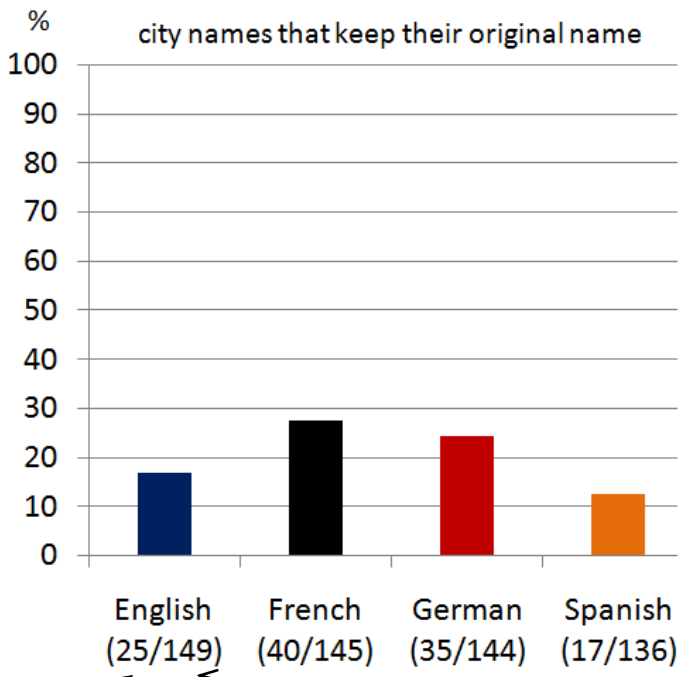
3.2 Experiments and Results – Quantity Check

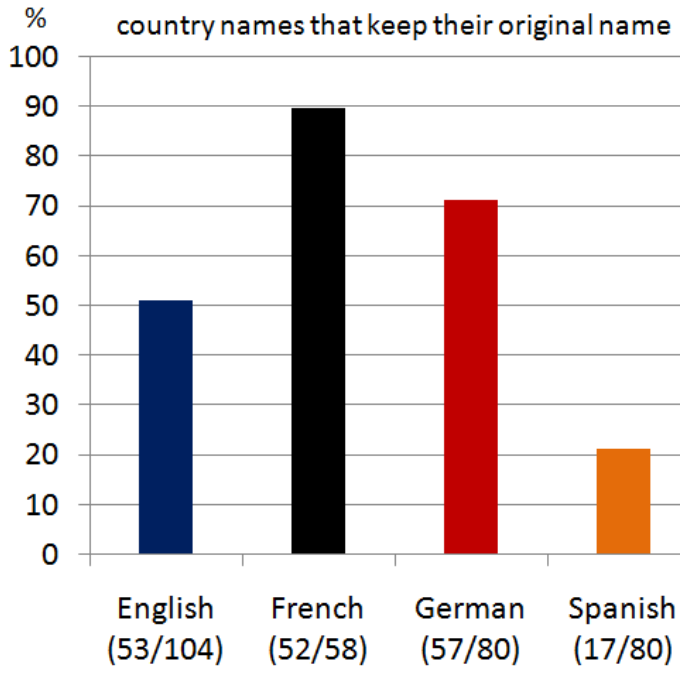
- Quantity of pronunciations for **proper names**
  - Proper names can be of diverse etymological origin and can surface in another language without undergoing the process of assimilation to the phonetic system of the new language (Llitjós and Black, 2002)
    → important as difficult to generate with letter-to-sound rules
  - Search pronunciations of 189 international city names and 201 country names to investigate the coverage of proper names:



3.2 Experiments and Results – Quantity Check

- Quantity of pronunciations for **proper names**
  - Results of only those words that keep their original name in the target language:





found prons. for country names that keep their original name

names which keep the original name in target language

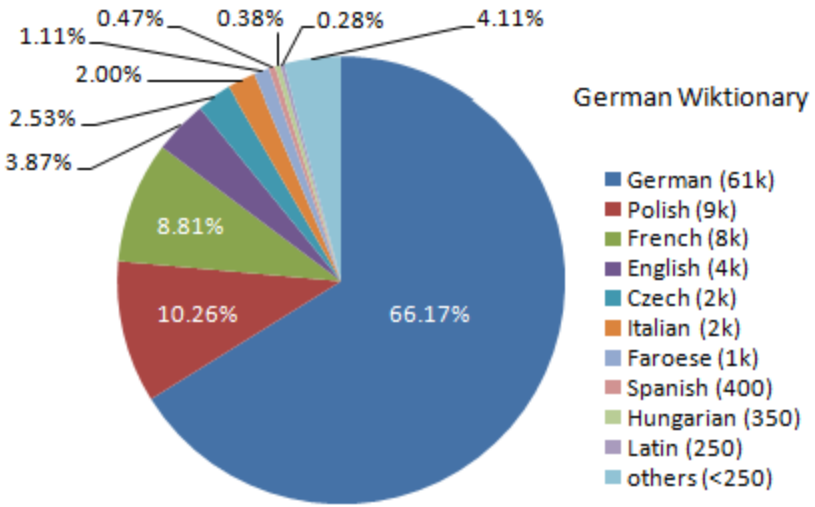
3.2 Experiments and Results – Quantity Check

- A look at the amounts of ***Wiktionary*** **pages containing prons.** also shows the differences in quantity betw. the *Wiktionary* editions:

Language	# pages	# pages with prons	% pages with prons	# prons	% prons in target language due to tag information
French	1,786k	912k	51.1%	954k	86.16%
English	1,770k	159k	9.0%	181k	14.92%
German	110k	48k	43.6%	91k	66.17%
Spanish	48k	8k	16.7%	19k	47.53%

- Many pronunciations in a given *Wiktionary* are for foreign languages

Example:



(According to the language information from the tags)

3.3 Experiments and Results – Quality Check

- **Impact of new pronunciation variants on ASR Performance**

  Approach **I**: Add all new *Wiktionary* pronunciations to *GlobalPhone* dictionaries and use them for training and decoding (*System1*)

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

Amount of *GlobalPhone* pronunciations, percentage of identical *Wiktionary* pronunciations and amount of new *Wiktionary* pronunciation variants

	WER baseline	WER System1	rel. improv.*
French	23.43%	23.25%	0.79%
English	21.51%	22.46%	-4.44%
German	21.60%	21.67%	-0.31%
Spanish	14.68%	14.42%	1.76%

Impact of using all *Wiktionary* pronunciations for training and decoding

*Improvements are significant at a significant level of 5%

→ How to ensure that new pronunciations fit to training and test data?

3.3 Experiments and Results – Quality Check

- **Impact of new pronunciation variants on ASR Performance**

  Approach **II**: Use only those *Wiktionary* pronunciations in decoding that were chosen in training (*System2*)

  - *Wiktionary* pronunciations chosen in training during forced alignment are of good quality for training data
  - Assumption:
    Similarity of training and test data in speaking style and vocabulary

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

Amount and percentage of *Wiktionary* pronunciations selected in training

*Improvements are significant at a significant level of 5%

	WER baseline	WER System1	rel. improv.*	WER System2	relative improvement*
French	23.43%	23.25%	0.79%	23.16%	1.17%
English	21.51%	22.46%	-4.44%	23.39%	-8.76%
German	21.60%	21.67%	-0.31%	21.07%	2.44%
Spanish	14.68%	14.42%	1.76%	13.62%	7.22%

4. Conclusion

- We proposed an efficient data source from the WWW that supports the rapid pronunciation dictionary creation
- We developed an Automatic Dictionary Extraction Tool that automatically extracts phonetic notations in IPA from *Wiktionary*
- Best quantity check results: French *Wiktionary* (92.58% for *GlobalPhone* word list, 76.12% for country names, 30.16% for city names)
- Best quality check results*:* Spanish *Wiktionary (*7.22% relative word error rate reduction)
- Particular helpful for pronunciations of proper names
- Results depend on community and language support
- *Wiktionary* pronunciations improved all system but the English one

Thanks for your interest!

ありがとうございます。

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, "Polyphone decision tree specialization for language adaptation," in *Proceedings of the ICASSP*, Instanbul, 2000.

[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.