Karlsruhe Institute of Technology

Institute for Anthropomatics and Robotics
Cognitive Systems Lab (CSL)
http://csl.anthropomatik.kit.edu

GlobalPhone: Pronunciation Dictionaries in 20 Languages

Tanja Schultz and Tim Schlippe

The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May 2014

1. The GlobalPhone Corpus

- Multilingual database with high quality read speech with transcriptions and pronunciation dictionaries in 20 languages
- More than 400 hours transcribed audio data of more than 2,000 native speakers
- Excellent basis for multilingual speech processing (Speech Recognition, Speech Synthesis, Speaker ID, Language ID)

2. The GlobalPhone Pronunciation Dictionaries

Dictionary Statistics

Languages	Unit	#Phones	#Units	#Dict entries
Arabic	w	44	29669	31840
Bulgarian	w	44	20288	20465
Bulgarian$_{EXT}$	w	44	260k	260k
Croatian	w	30	22522	29602
Croatian$_{EXT}$	w	30	143k	143k
Czech	w	41	32942	33049
Czech$_{EXT}$	w	41	277k	277k
French	w	38	20710	36837
German	w	41	46035	48979
Hausa	w	33	42127	42711
Japanese	s	31	58829	58829
Korean	w	41	50220	50220
Korean	s	41	1276	1276
Mandarin	w	139	73444	73444
Mandarin	c	46	3113	3113
Polish	w	36	36484	36484
Polish$_{EXT}$	w	36	125k	125k
Portuguese	w	45	58787	58803
Russian	w	47	31719	32964
Spanish	w	40	36176	45467
Swedish	w	48	28069	28214
Tamil	w+s	41	219k	219k
Thai	s	44	22189	25326
Turkish	w	29	33514	33757
Ukrainian	w	51	7745	7934
Ukrainian$_{EXT}$	w	51	40k	40k
Vietnamese	s	59	30166	38696

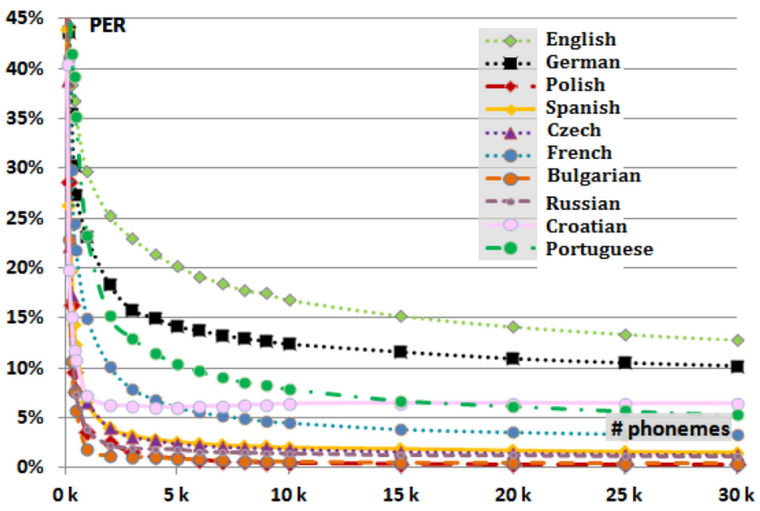
- Dictionaries in languages from
  - Europe
  - Africa
  - America
  - Asia
- Large variety of language peculiarities relevant for speech and language processing
  - Phonetic characteristics, phonotactics
  - Writing systems, word segmentation
  - Morphological variations

Dictionary Formats

```
{120} {{M_s WB} M_t M_o M_d M_v M_a M_d M_tS M_a {M_tj WB}}
{12го} {{M_d WB} M_v M_jE M_n M_a M_d M_tS M_a M_tj M_g {M_o WB}}
{12ти} {{M_d WB} M_v M_jE M_n M_a M_d M_tS M_a M_tj M_t {M_i WB}}
:
{Адольфа} {{M_a WB} M_d M_o M_lj M_f {M_a WB}}
{Азартная} {{M_a WB} M_z M_a M_r M_t M_n M_a {M_jA WB}}
{Азатакан} {{M_a WB} M_z M_a M_t M_a M_k M_a {M_n WB}}
{Азебаржан} {{M_a WB} M_z M_jE M_b M_a M_r M_Z M_a {M_n WB}}
{Азербайджан} {{M_a WB} M_z M_jE M_r M_b M_a M_j M_Z M_a {M_n WB}}
{Азербайджан(2)} {{M_a WB} M_z M_jE M_r M_b M_a M_d M_Z M_a {M_n WB}
{Азербайджан(3)} {{M_a WB} M_z M_jE M_r M_b M_a M_Z M_a {M_n WB}}
{Азербайджан(4)} {{M_a WB} M_z M_jE M_b M_a M_r M_Z M_a {M_n WB}}
:
```

- Phone naming conventions consistent across all languages, leveraging the International Phonetic Alphabet (IPA)
- Tags
  - Word boundary (WB)
  - Tone (in Hausa, Mandarin, Vietnamese)
  - Length (Long/Short) (in Hausa)
- Pronunciation variants

G2P Relationship



- G2P depends on the language (EN > DE > {PT, FR, HR} > {ES, RU, CZ, PL, BG})
- Close G2P Relationship: 5k phones ok, 15k saturation
- E.g. German needs 6x more examples than Portuguese for same prediction performance

3. Rapid Language Adaptation Toolkit (RLAT)

- Provides innovative methods and interactive web-based tools to:
  - Develop speech processing components for new languages at low costs
  - Continuously harvest, normalize, and process text data from web
  - Create prompts for recordings, create vocabulary lists
  - Select appropriate phone sets for new languages efficiently
  - Automatically generate pronunciation dictionaries
  - Iteratively build and evaluate speech processing components
- **http://csl.ira.uka.de/rlat-dev**

4. Language Models and Vocabulary Lists

Language/Unit		3-gram PPL LM_B	3-gram PPL LM	OOV [%]	#Vocab	#Token [Mio]
Arabic	w	no additional resources yet				
Bulgarian	w	454	351	1.0	274k	405
Croatian	w	721	647	3.6	362k	331
Czech	w	1421	1361	4.0	267k	508
French	w	324	284	2.4	65k	-
German	w	672	555	0.3	38k	20
Hausa	w	97	77	0.5	41k	15
Japanese	s	89	76	1.0	67k	1600
Korean	s	25	18	0	1.3k	500
Mandarin	c	262	163	0.8	13k	900
Portuguese	w	58	49	9.8	62k	11
Polish	w	951	904	0.8	243k	224
Russian	w	1310	1150	3.9	293k	334
Spanish	w	154	108	0.1	19k	12
Swedish	w	423	387	5.3	73k	211
Tamil	w+s	730	624	1.0	288k	91
Thai	s	70	65	0.1	22k	15
Turkish	w	-	45	13.2	29k	7
Ukrainian	w	594	373	0.5	40k	94
Vietnamese	s	218	176	0	30k	39

5. Availability

- Pronunciation dictionaries available from authorized distributor: European Language Resources Association (ELRA)
- Speech and text corpus available from two authorized distributors: ELRA and Appen Butler Hill Pty Ltd.
- Benchmark numbers and language models available for free download: http://csl.ira.uka.de/GlobalPhone