GlobalPhone: Pronunciation Dictionaries in 20 Languages

Tanja Schultz and Tim Schlippe
Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany
tanja.schultz@kit.edu

Abstract

This paper describes the advances in the multilingual text and speech database GLOBALPHONE a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. GLOBALPHONE was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers GLOBALPHONE supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages. Very recently the GLOBALPHONE pronunciation dictionaries have been made available for research and commercial purposes by the European Language Resources Association (ELRA).

Keywords: Speech, Text, and Dictionary Resources for Multilingual Speech Processing

1. Introduction

With more than 7100 languages in the world (Lewis et al., 2013) and the need to support multiple input and output languages, it is one of the most pressing challenge for the speech and language community to develop and deploy speech processing systems in yet unsupported languages rapidly and at reasonable costs (Schultz, 2004; Schultz and Kirchhoff, 2006). Major bottlenecks are the sparseness of speech and text data with corresponding pronunciation dictionaries, the lack of language conventions, and the gap between technology and language expertise. Data sparseness is a critical issue due to the fact that today's speech technologies heavily rely on statistically based modeling schemes, such as Hidden Markov Models and n-gram language modeling, as well as neural network based approaches such as Deep Neural Networks for acoustic modeling and Recurrent Neural Networks for language modeling. Although these machine learning approaches and algorithms are mostly language independent and proved to work well for a variety of languages, reliable parameter estimation requires vast amounts of training data.

Unfortunately, large-scale data resources for research are available for less than 100 languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. Furthermore, the lack of language conventions concerns a surprisingly large number of languages and dialects. The lack of a standardized writing system for example hinders web harvesting of large text corpora and the construction of pronunciation dictionaries and lexicons. Last but not least, despite the well-defined system building process, it is cost- and time consuming to handle language-specific peculiarities, and requires substantial language expertise. Also, it is challenging to find system developers who have both, the necessary technical background and the native expertise of a language in question. Thus, one of the pivotal issues to develop speech processing systems in multiple languages is the challenge of bridging the gap between language and technology expertise (Schultz, 2004).

More than ten years ago we released a multilingual text and speech corpus GLOBALPHONE to address the lack of databases which are consistent across languages (Schultz, 2002). By that time the database consisted of 15 languages but since then has been extended significantly to cover more languages, more speakers, more text resources, and more word types along with their pronunciations. In addition, GLOBALPHONE was adopted as a benchmark database for research and development of multilingual speech processing systems. Recently, we published the latest status of the GLOBALPHONE speech and text resources along a description of the Rapid Language Adaptation Toolkit which was used to crawl additional text resources for languages modeling, and presented the performances of the resulting speech recognition systems to provide a reference and benchmark numbers for researchers and developers who work with the GLOBALPHONE corpus (Schultz et al., 2013). This paper focuses on the formats, building process, and phone sets of the pronunciation dictionaries, as they have been recently released.

2. The GlobalPhone Corpus

GLOBALPHONE is a multilingual data corpus developed in collaboration with the Karlsruhe Institute of Technology (KIT). The complete data corpus comprises (1) audio/speech data, i.e. high-quality recordings of spoken utterances read by native speakers, (2) corresponding transcriptions, (3) pronunciation dictionaries covering the vocabulary of the transcripts, and (4) baseline n-gram language models. The first two are referred to as GLOBALPHONE Speech and Text Database (GP-ST), the third as GLOBALPHONE Dictionaries (GP-Dict), and the latter as GLOBALPHONE Language Models (GP-LM). GP-ST is distributed under a research and a commercial license by two authorized distributors, the European Language Resources Association (ELRA) (ELRA, 2012) and Appen Butler Hill Pty Ltd. (Appen Butler Hill Pty Ltd, 2012). GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website (LM-BM, 2012).

The entire GLOBALPHONE corpus provides a multilingual database of word-level transcribed high-quality speech for the development and evaluation of large vocabulary speech processing systems in the most widespread languages of the world. GLOBALPHONE is designed to be uniform across languages with respect to the amount of data per language, the audio quality (microphone, noise, channel), the collection scenario (task, setup, speaking style), as well as the transcription and phone set conventions (IPA-based naming of phones in all pronunciation dictionaries). Thus, GLOBALPHONE supplies an excellent basis for research in the areas of (1) multilingual speech recognition, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) multilingual speech synthesis, as well as (6) monolingual speech recognition in a large variety of languages.

2.1. Language Coverage

To date, the GLOBALPHONE corpus covers 21 languages, i.e. Arabic (Modern Standard Arabic), Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. Since no pronunciation dictionary and language model data is made available in the Shanghai dialect at this point, the paper describes the resources for 20 languages only, hence the title of the paper. This selection of 20 languages covers a broad variety of language peculiarities relevant for Speech and Language research and development. It comprises wide-spread languages (e.g. Arabic, Chinese, Spanish, Russian), contains economically and politically important languages, and spans wide geographical areas (Europe, Africa, America, Asia).

The spoken speech covers a broad selection of phonetic characteristics, including pharyngeal sounds (Arabic), consonantal clusters (German), nasals (French, Portuguese), and palatalized sounds (Russian) as well as various lexical properties such as tonal languages (Mandarin, Thai, Vietnamese). The written language contains all types of writing systems, i.e. logographic scripts (Chinese Hanzi and Japanese Kanji), phonographic segmental scripts (Roman, Cyrillic), phonographic consonantal scripts (Arabic), phonographic syllabic scripts (Japanese Kana, Thai), and phonographic featural scripts (Korean Hangul). The languages cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), compounding languages (German), and also include scripts that completely lack word segmentation (Chinese, Thai).

For details on the Data Acquisition and corpus statistics we refer to (Schultz et al., 2013). This paper focus on the description of the GLOBALPHONE pronunciation dictionaries.

3. GlobalPhone Pronunciation Dictionaries

Phone-based pronunciation dictionaries are available for each of the 20 GLOBALPHONE language. The dictionaries cover all word units which appear in the transcription data of the GLOBALPHONE speech recordings.

The word forms are given in UTF-8 or Unicode format for 16-bit encodings to preserve the original script but for some languages are also be provided in Romanized version (Arabic, Korean, Japanese, Korean, Mandarin) using ASCII encoding to fit the Romanized script as appearing in the directory `/rmn` in the speech & text corpus. The conversion between the Roman and the language specific script is provided in the documentation.

If the grapheme-to-phoneme relationship in a language permits, the dictionaries were constructed in a rule-based manner using language specific phone sets plus noise and silence models. The number of phones was generally optimized for the purpose of automatic speech recognition. The phone set and its distribution in the dictionaries are described in large detail in the documentation. After this automatic creation process the dictionaries were manually post-processed word-by-word by native speakers, correcting potential errors of the automatic pronunciation generation process, and adding pronunciations for acronyms and numbers.

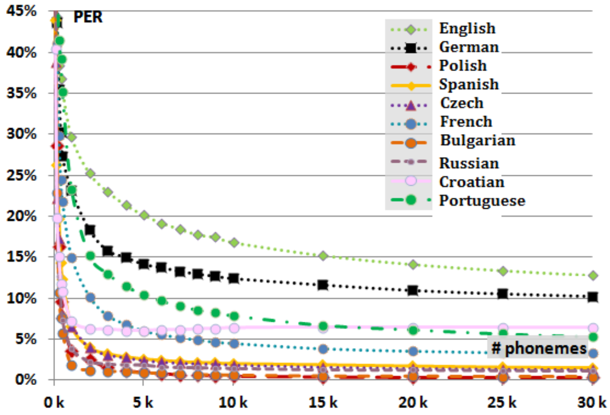

Figure 1: Phone Error Rate (PER) over Training Data for 10 languages (see also (Schlippe et al., 2014))

Figure 1 describes the quality of grapheme-to-phoneme converters as a function of languages (closeness of grapheme-to-phoneme relation) and of the amount of training data for these converters. In this particular experiment (see (Schlippe et al., 2014) for more details), we used the GLOBALPHONE dictionaries as ground truth, randomly picked a certain amount of dictionary entries (dictionary entry = unit plus corresponding pronunciation), trained a grapheme-to-phoneme converter (we applied the Sequitur converter from (Bisani and Ney, 2008)), applied the converter to unseen units in order to generate a new dictionary entry, and finally evaluated the results in terms of Phone Error Rates (PER) by comparing the converter output to the correct dictionary entry from the GLOBALPHONE dictionaries. The amount of data on the x-axis of the graph in figure 1 is expressed in terms of the total number of phones in the dictionary entries. As dictionary units are built of roughly five phones on average, 5k phones training material corresponds to about 1000 dictionary entries.

As figure 1 shows, creating pronunciations for languages with a close grapheme-to-phoneme (g-2-p) relationship like Bulgarian, Czech, Polish, and Spanish, 1000 to 2000 dictionary entries are sufficient for training a well performing converter. For example, the phone error rate on Polish is lower than 4% and drops to 3.2% with 30k phones (6k dictionary entries) for training. However, as the relationship gets weaker (Portuguese, French, German), significantly more training examples are required to achieve similar performances. For example, the German g-2-p converter requires 6 times more training material (30k phones) than the Portuguese one (5k phones) to achieve the same quality (10% PER) on the generated pronunciations. Furthermore, from the plots in figure 1 we - not surprisingly - conclude that g-2-p conversion for languages like Portuguese, French, German, and in particular English, will not reach the performance level of those with close g-2-p relationship.

3.1. Dictionary Formats

The format of the GlobalPhone dictionaries is very straight forward. Each line consists of one word form and its pronunciation. Words forms and their pronunciations are separated by blank. The pronunciation consists of a concatenation of phone symbols separated by blanks. Both, words and their pronunciations are given in tcl-script list format, i.e. enclosed in {}, since phones can carry various types of tags: a word boundary tag `WB`, indicating the boundary of a dictionary unit, tone tags `T`n where n indicates tonal variations, and length tags `L/S` which indicate long or short length of a pronunciation of the corresponding phone. All dictionaries contain the `WB` tags, while tone tags are represented in Hausa, Mandarin and Vietnamese, length tags in Hausa. The tags can be included as standard questions in the decision tree, for example in the case of `WB` for capturing crossword models in context-dependent modeling. Furthermore, most of the dictionaries contain pronunciation variants, which are indicated by (n) with $n = 2, 3, 4, \ldots$ showing the number of variants per word. The order in which variants occur is not necessarily related to their frequency in the corpus. Figure 2 provides a short excerpt from the Russian dictionary.

```
{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}}
:
{ярмо} {{M_jA WB} M_r M_m {M_o WB}}
{ярость} {{M_jA WB} M_r M_o M_s {M_tj WB}}
{ярче} {{M_jA WB} M_r M_tS {M_jE WB}}
{ярый} {{M_jA WB} M_r M_i2 {M_j WB}}
{ясно} {{M_jA WB} M_s M_n {M_o WB}}
{ясности} {{M_jA WB} M_s M_n M_o M_s M_t {M_i WB}}
{ясность} {{M_jA WB} M_s M_n M_o M_s {M_tj WB}}
{ясны} {{M_jA WB} M_s M_n {M_i2 WB}}
{ястребов} {{M_jA WB} M_s M_t M_r M_jE M_b M_o {M_v WB}}
{яти} {{M_jA WB} M_t {M_i WB}}
{ях} {{M_jA WB} {M_x WB}}
{ячейку} {{M_jA WB} M_tS M_jE M_j M_k {M_u WB}}
{ящиками} {{M_jA WB} M_StS M_i M_k M_a M_m {M_i WB}}
```

Figure 2: Short excerpt from the Russian Dictionary

3.2. Multilingual Phone Naming Conventions

To support the development of multilingual speech processing, the GlobalPhone phone naming conventions are consistent across all languages, leveraging the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999). This allows to merge the phone sets and acoustic models of different GlobalPhone languages in order to create multilingual phone inventories and acoustic models. It also supports the merging of dictionaries, for example for the purpose of multilingual speech synthesis or for building recognizers for code-switching. Furthermore, the merging of phone sets supports the creation of multilingual grapheme-to-phoneme rules and with this the option to potentially generate pronunciations across languages (see for example (Schlippe et al., 2013)). A mapping between language specific and GlobalPhone phone naming conventions is provided for each language in the documentation. Figure 3 shows the IPA consonant chart and the phone names of the corresponding phones for the GlobalPhone languages. A similar chart exists for vowels.

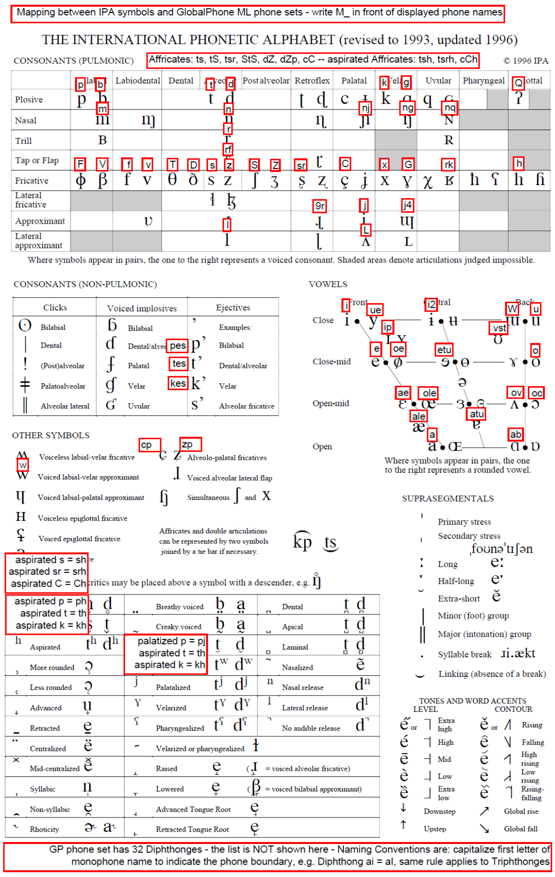
Mapping between IPA symbols and GlobalPhone ML phone sets - write M_ in front of displayed phone names
THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, updated 1996)
CONSONANTS (PULMONIC)
Affricates: ts, tS, tsr, StS, dZ, dZp, cC -- aspirated Affricates: tsh, tsrh, cCh
© 1996 IPA
p b t d k g Q m n nj ng nq r rf F V f v T D s z S Z sr C x G rk h 9r j j4 l L
Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.
CONSONANTS (NON-PULMONIC)
pes tes kes
VOWELS
i ue i2 W u ip vst e oe etu o ae ole ale atu a ab
Where symbols appear in pairs, the one to the right represents a rounded vowel.
OTHER SYMBOLS
w cp zp
aspirated s = sh
aspirated sr = srh
aspirated C = Ch
aspirated p = ph
aspirated t = th
aspirated k = kh
palatized p = pj
aspirated t = th
aspirated k = kh
SUPRASEGMENTALS
TONES AND WORD ACCENTS
GP phone set has 32 Diphthonges - the list is NOT shown here - Naming Conventions are: capitalize first letter of monophone name to indicate the phone boundary, e.g. Diphthong ai = aI, same rule applies to Triphthonges

Figure 3: Naming conventions for GlobalPhone

3.3. Dictionary Statistics

Table 1 gives an overview of the size of the phone sets, amount of vocabulary units covered, and amount of pronunciation variants in the GlobalPhone pronunciation dictionaries. In addition, the table lists the dictionary entry unit in which the pronunciations are provided, i.e. on the word level (w) if the script provides word-level segmentation, on the syllable (s) or character level (c). A description about how these units were derived from the text resources for the various languages is beyond the scope of this paper. We refer to our work published on GLOBALPHONE systems in various languages. For the Korean and Mandarin language, dictionaries are provided in two different units as listed in table 1. In addition, for some languages (Bulgarian, Croatian, Czech, Polish, Ukrainian) extended versions of the pronunciation dictionaries are available which cover a significantly larger vocabulary (indicated by $_{EXT}$). However, for these extented versions, the dictionary entries were automatically generated from grapheme-to-phoneme converters without extensive manual cross-checks.

Table 1: GLOBALPHONE Pronunciation Dictionaries

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

4. Language Models and Vocabulary Lists

To further increase vocabulary coverage, reduce Out-Of-Vocabulary (OOV) rates and improve language models, we implemented several tools which are integrated into our Rapid Language Adaptation Toolkit (RLAT).

RLAT is a web-based interface which aims to reduce the human effort involved in building speech processing systems for new languages. Innovative tools enable novice and expert users to develop speech processing models, such as acoustic models, pronunciation dictionaries, and language models, to collect appropriate speech and text data for building these models, and to iteratively evaluate the results. It allows to indicate starting web pages and then continues to crawl down to a given link depth. Histories ensure that the same pages are not crawled twice. RLAT includes features like the snap-shot function which gives automated feedback about the quality of text data crawled from the web. For this, the user specifies a time interval when new language models are automatically built based on the harvested data. The quality of the language models are then evaluated based on criteria, such as perplexity, OOV rate, n-gram coverage, vocabulary size, and WER given a test corpus and a speech recognition system. The outcome is presented in easy-to-digest graphs and reports. Other useful features are automatic cleaning and normalization processes, filtering methods which are modularized into language independent parts and language dependent aspects which can be modified easily by the user. Also, automatic language identification on the texts is applied to make the crawling process more efficient (see (Vu et al., 2010) for more details).

Based on RLAT we crawled text data for several days, and each day one language model was built based on the daily crawled text data. The final language model was then created by a linear interpolation of all daily language models. The interpolation weights were computed using the SRI Language Model Toolkit (Stolcke, 2002), optimized on the GLOBALPHONE development sets. The experimental results in (Vu et al., 2010) indicated that the text data from the first few days are most helpful and therefore receive the highest interpolation weights in the final language model. Since the outcome of the crawling process depends on the input websites, the starting pages have to be chosen carefully. In some cases (Croatian, Japanese, Korean, Thai) the crawling process finished prematurely. In those cases we selected additional websites to harvest more diverse text data.

Table 2: GLOBALPHONE Text Data & Language Models

Language/Unit		3-gram PPL		OOV	#Vocab	#Token
		LM$_B$	LM	[%]		[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

The final best language model for each language was then built based on the interpolation of the language models from a variety of websites. Since some scripts lack a segmentation into words or do not provide a suitable definition of 'word units' (Chinese, Korean, Japanese, Tamil, Thai, and Vietnamese) we defined syllables or characters as token units for the purpose of speech recognition. Table 2 gives an overview of the amount of crawled text data, the trigram perplexities (PPL), out-of-vocabulary (OOV) rates, and the vocabulary sizes of the GLOBALPHONE language models, for both the full (LM) and the pruned benchmark language models (LM$_B$), which are available for download from our website (LM-BM, 2012). The symbols in the second column after the language name indicate the token units used, i.e. (w) for word-based, (s) for syllable-based, and (c) for character-based token units. Since the main focus of this paper is on the pronunciation dictionary, we refer the interested reader to (Vu et al., 2010) for more details on the language modeling part.

5. Summary

In this paper we presented the latest status of the GLOBALPHONE speech and language resources in 20 different languages with a particular emphasis on the pronunciation dictionaries. We described the dictionary formats and phone naming conventions, summarized the number of entries covered in the pronunciation dictionaries and the amount of text data along with the characteristics of the language models. These resources are available to the community for research and development of multilingual speech processing systems. The GLOBALPHONE speech and text resources are distributed under a research and a commercial license by two authorized distributors, the European Language Resources Association (ELRA) (ELRA, 2012) and Appen Butler Hill Pty Ltd. (Appen Butler Hill Pty Ltd, 2012). The pronunciation dictionaries are distributed by ELRA, while the GP-LMs are freely available for download from our website (LM-BM, 2012).

6. References

Appen Butler Hill Pty Ltd. (2012). Speech and Language Resources 2012. Appen Butler Hill Speech and Language Resources 2012 - Product Catalogue.

Bisani, M. and Ney, H. (2008). Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, 50:434–451.

ELRA. (2012). European language resources association (ELRA). ELRA catalogue. Retrieved November 30, 2012, from http://catalog.elra.info.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Lewis, M., Simons, G., and Fennis, C., editors. (2013). *Ethnologue: Languages of the World (17th ed.)*. SIL International, Dallas, TX.

LM-BM. (2012). Benchmark GlobalPhone Language Models. Retrieved November 30, 2012, from http://csl.ira.uka.de/GlobalPhone.

Schlippe, T., Volovyk, M., Yurchenko, K., and Schultz, T. (2013). Rapid bootstrapping of a ukrainian large vocabulary continuous speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 26-31.

Schlippe, T., Ochs, S., and Schultz, T. (2014). Web-based tools and methods for rapid pronunciation dictionary creation. *Speech Communication*, 56:101–118.

Schultz, T. and Kirchhoff, K. (2006). *Multilingual Speech Processing*. Elsevier Academic Press.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A Multilingual Text and Speech Database in 20 Languages. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8126 – 8130, Vancouver, BC, Canada, May 26-31.

Schultz, T. (2002). Globalphone: A Multilingual Speech and Text Database Developed at Karlsruhe University. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 345–348, Denver, CO.

Schultz, T. (2004). Towards Rapid Language Portability of Speech Processing Systems. In *Conference on Speech and Language Systems for Human Communication (SPLASH)*, volume 1, Delhi, India, November.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.

Vu, N. T., Schlippe, T., Kraus, F., and Schultz, T. (2010). Rapid bootstrapping of five eastern european languages using the rapid language adaptation toolkit. In *Proceedings of InterSpeech*, pages 865–868, Makuhari, Japan.