Documentation

GLOBALPHONE: a Multilingual Text & Speech Database

Version 4.1

© XLingual, GmbH & Co. KG
Schwachhauser Heerstraße 251, 28211 Bremen, Germany

January 22 2016

TOC:

I. Summary

The GlobalPhone corpus was designed to provide read speech data for the development and evaluation of large continuous speech recognition systems in the most widespread languages of the world, and to provide a uniform, multilingual speech and text database for language independent and language adaptive speech processing as well as for language and speaker recognition tasks. The entire GlobalPhone corpus enables the acquisition of acoustic-phonetic knowledge of the following 22 spoken languages: Arabic (AR), Bulgarian (BG), Chinese-Mandarin (CH), Chinese-Shanghai (WU), Croatian (CR), Czech (CZ), French (FR), German (GE), Hausa (HA), Japanese (JA), Korean (KO), Brazilian-Portuguese (PO), Polish (PL), Russian (RU), Spanish (SP), Swahili (SWA), Swedish (SW), Tamil (TA), Thai (TH), Turkish (TU), Ukrainian (UA), and Vietnamese (VN). The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers were asked to read about 100 sentences each. The numbers of speakers and utterances varies between languages, see details below. The read texts were selected from national newspapers available via Internet to provide a large vocabulary. The majority of read articles cover national and international political news as well as economic news from some period between the years 1995 and 2012. The speech data is available in PCM encoding, 16 bit resolution, 16 kHz sampling rate, mono quality, recorded with a close-speaking microphone (e.g. Sennheiser 440-6). The transcriptions are validated post-recording and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects like laughing and hesitations. Speaker information such as age, gender, occupation, etc. as well as information about the recording setup complements the database. The entire GlobalPhone corpus contains about 450 hours of speech spoken by more than 2100 adult native speakers from 22 languages. The data are divided in speaker disjoint sets for training, development and evaluation (80:10:10) and are organized by languages and speakers.

II. Corpus Design

The goal of the GlobalPhone database collection was to provide read speech data suitable for multilingual large vocabulary continuous speech recognition (LVCSR) with a particular emphasis on language independent and language adaptive modeling. Due to the nature of the database, GlobalPhone supplies an excellent basis for research in the areas of (1) multilingual speech recognition, (2) rapid deployment of speech processing systems in under-resourced languages, (3) language and speaker recognition tasks, (4) multilingual speech synthesis and voice conversion, as well as (5) monolingual recognition in a large variety of languages.

A. Language Selection

There are about 7000 languages in the world. The majority of languages are spoken by less than 100,000 speakers; only about 150 languages (3%) have more than 1 Million speakers. To select a representative subset of languages, we considered the following characteristics:

- Size of speaker population (e.g. Chinese, Spanish, Russian, Arabic),
- Political relevance (e.g. Chinese, Arabic, Korean, Japanese, Spanish),
- Geographic coverage: European, African, Asian, Central Asian, Indian,
- Phonetic coverage, e.g. tones (Chinese, Hausa, Thai, Vietnamese), pharyngeal sounds (Arabic), palatalized and semi-palatalized sounds (Ukrainian), ejective (Hausa) and implosive consonants (Hausa, Swahili), to name only a few,
- Orthographic scripts:

    - Phonologic: alphabetic scripts (such as Latin, Cyrillic, Arabic); and syllable-based scripts (Japanese Kana, Korean Hangul),
    - ideographic scripts (pictographs): such as Chinese Hanzi and Japanese Kanji,
- Grapheme-to-phoneme relationship: very close relationship (e.g. Spanish, Slavic languages) versus lose relationship (e.g. English),
- Morphologic variations: Rich morphology like agglutinative languages (Turkish, Korean), compounding languages (German) versus poor morphology of not-inflecting languages (Chinese),
- Word segmentation: languages for which the written form provides segmentation in word units (e.g. German, French, Czech), does not provide any segmentation (e.g. Chinese, Japanese, Thai) or those, for which the segmentation is not appropriate for LVCSR (e.g. Korean, Vietnamese).

Table 1 below shows the ranking of the most widespread languages of the world by estimated number of native speakers as listed in the 2007 edition of National Encyklopedin (NE), the top eleven languages have updated figures from 2010. Since the consensus methods vary across countries, all numbers are based on estimates. Numbers above 95 million are rounded off to the nearest 5 million (see also Wikipedia, List of Languages). The languages covered by the GlobalPhone corpus are marked by "*".

Rank	*Language*	*Primary Locales*	*Speakers*	*Language Group*
1. *	Mandarin	China	955 Mio	Sino-Tibetan (Sinitic)
2. *	Spanish	Latin-America, Spain	405 Mio	Indo-European (Romance)
3. (*)	English	USA, UK, Can, Australia	360 Mio	Indo-European (Germanic)
4.	Hindi	India	310 Mio	Indo-European (Indo-Iran)
5. *	Arabic	N. Africa, Mid East	295 Mio	Afro-Asiatic (Semitic)
6. *	Portuguese	Brazil, Portugal, Angola	215 Mio	Indo-European (Romance)
7.	Bengali	Bangladesh, India	200 Mio	Indo-European (Indo-Iran)
8. *	Russian	Russia, Independent States	155 Mio	Indo-European (East Slavic)
9. *	Japanese	Japan	125 Mio	Isolate
10.	Punjabi	Punjab, India	100 Mio	Indo-European (Germanic)
11. *	German	G, Austria, Switzerland	89 Mio	Indo-European (Germanic)
13. *	Wu/Shanghai	China (Shanghai)	80 Mio	Sino-Tibetan (Sinitic)
14. *	French	F, Can, Africa, Switzerland	80 Mio	Indo-European (Romance)
16. *	Vietnamese	Vietnam	76 Mio	Austro-Asiatic (Khmer)
17. *	Korean	Korea, China	76 Mio	Isolate
20. *	Tamil	India, Sri Lanka, Malaysia	70 Mio	Dravidian
22. *	Turkish	Turkey	63 Mio	Altaic (Turkic)
25. *	Thai	Thailand	56 Mio	Sino-Tibetan (Thai)
30. *	Polish	Poland	40 Mio	Indo-European (West Slavic)
36. *	Hausa	Nigeria +12 other countries	34 Mio	Afro-Asiatic (West Chadic)
40. *	Ukrainian	Ukraine	30 Mio	Indo-European (East Slavic)
58. *	Serbo-Croatian	Balkan Europe	19 Mio	Indo-European (South W Slavic)
83. *	Czech	Czech Republic	10 Mio	Indo-European (West Slavic)
91. *	Swedish	Sweden, Finland	8.7 Mio	Indo-European (Germanic)
>100 *	Bulgarian	Bulgaria	8 Mio	Indo-European (South E Slavic)
*	Swahili	Tanzania, Kenya, Uganda, Democratic Republic of Congo	5 Mio	Bantu, Lingua franca in Southeast Africa / 60-150 Mio speakers

Table 1: Most widespread languages of the world by estimated number of native speakers (NE 2007)

B. Task and Text Selection

To limit the effort of transcription which is the most time and cost consuming process of a data collection and to ease the collection of large amounts of similar text data, the collection consists of read speech from electronically available text sources. The texts were selected from national newspapers available from the web. Texts were chosen from international and national political and economic topics to somewhat restrict the vocabulary. Most of the articles were selected between May 1996 and November 1997, and between 2006 and 2012 which allows to compare of proper names (Politicians, companies, etc.) across languages.

For the GlobalPhone corpus we used the following newspapers: Assabah for Arabic, Banker, Cash, and Sega for Bulgarian, Peoples Daily for Mandarin and Shanghai Chinese, HRT and Obzor Nacional for Croatian, Ceskomoravsky Profit Journal and Lidove Noviny newspaper for Czech, Le Monde for French, Frankfurter Allgemeine und Süddeutsche Zeitung for German, CRI online and RFI for Hausa, Hankyoreh Daily News for Korean, Nikkei Shinbun for Japanese, Folha de Sao Paulo for Portuguese, Dziennik Polski for Polish, Ogonyok Gaseta and express-chronika for Russian, La Nacion for Spanish, Voice of America and various others for Swahili, Goeteborgs-Posten for Swedish, Thinaboomi Tamil Daily for Tamil, Bangkok Biz news and Daily News for Thai, Zaman for Turkish, nine newspaper resources for Ukrainian, and Tin Tuc among others for Vietnamese. Table 2 gives the web addresses for the newspapers.

Language	*Internet link*
Arabic	http://www.tunisie.com/Assabah
Bulgarian	http://www.banker.bg , http://www.cash.bg , http://www.segabg.com
Ch-Mandarin	http://www.snweb.com
Ch-Shanghai	http://www.snweb.com
Croatian	http://hrt.com.hr/vijesti/hrt , http://nacional.hr, http://www.tel.hr.hrvatski-obzor
Czech	http://press.medea.cz/press/pr/index.html
French	http://www.lemonde.fr
German	http://www.faz.de , http://www.sueddeutsche.de
Hausa	http://hausa.cri.cn , http://ha1.chinabroadcast.cn , http://www.bbc.co.uk/hausa http://www.dw-world.de/hausa , http://www.hausa.rfi.fr http://www.voanews.com/hausa/news
Japanese	http://www.nikkeihome.co.jp
Korean	http://news.hani.co.kr
Portuguese	http://www.uol.com.br/fsp
Polish	http://www.dziennik.krakow.pl/
Russian	http://www.ropnet.ru/ogonyok
Spanish	http://www.nacion.co.cr
Swahili	http://www.voaswahili.com
Swedish	http://www.gp.se
Tamil	http://www.thinaboomi.com
Thai	http://www.bangkokbiznews.com , http://www.dailynews.co.th http://www.manager.co.th , http://www.matichon.co.th , http://www.naewna.com , http://www.thairath.co.th
Turkish	http://www.zaman.com.tr
Ukrainian	http://umoloda.kiev.ua, http://day.kiev.ua, http://ukurier.com.ua, http://pravda.com.ua, http://chornomorka.com, http://tsn.ua, http://champion.com.ua, http://ukrslovo.org.ua, http://epravda.com.ua
Vietnamese	http://www.tintuconline.vn , http://www.nhandan.org.vn , http://www.tuoitre.org.vn

Table 2: Newspaper links

In 2012 we crawled a massive amount of text data based on the strategy presented in [Vu et al., Interspeech 2010] to quickly and efficiently build language models for 19 GlobalPhone languages. The text data were crawled over several days, and day-wise language models were linearly interpolated to create a final language model per language using the SRI Language Model Toolkit. The resulting language models were benchmarked and described in [Schultz et al, ICASSP 2013]. Pruned versions are available for free download from the following website [http://csl.uni-bremen.de → Projects → GlobalPhone].

III. Data Collection

A. Collection Site

To avoid artifacts, which might occur when collecting speech of native speakers living in a non-native environment, we collected the complete GlobalPhone database in the home countries of the native speakers. The data collection was done in two batches, the first batch between May 1996 and November 1997, and a second batch between 2003 and 2012. During the first batch we collected Arabic in Tunis, Sfax and Djerba, Tunisia; Mandarin in Beijing, Wuhan and Hekou, China; Wu (Shanghainese) in Shanghai, China; Croatian in Zagreb, Croatia, and parts of Bosnia; Czech in Prague, Czech Republic; French in Grenoble, France; German in Karlsruhe, Germany; Japanese in Tokyo, Japan; Korean in Seoul, Korea; Portuguese in Porto Velho and Sao Paulo, Brazil; Polish in Poland, Russian in Minsk, Belarus; Spanish in Heredia and San Jose, Costa Rica; Swedish in Stockholm and Vaernamo, Sweden; Tamil in India, and Turkish in Istanbul, Turkey. In the second batch between 2003 and 2012 we collected Bulgarian in Sofia in 2005, Hausa in Cameroon in 2011, Swahili in Nairobi, Kenya in 2012, Thai in Bangkok, Thailand in 2003, Ukrainian in Donezk, Ukraine in 2011, and Vietnamese in Hanoi and Ho Chi Minh City, Vietnam in 2009.

B. Recording Equipment

The recording equipment for the first batch of data collection (1996-1997) consists of a portable Sony DAT-recorder TDC-8 and a close-talking Sennheiser microphone HD-440-6. The data was digitally recorded at a 48 kHz sampling rate at 16bit linear quantization. For further processing the data was down sampled to 16kHz sampling rate. The recording equipment is identical for all GlobalPhone languages of the first batch but the German language. For German data recordings the same microphone was used but speech was recorded on a SUN Ultra-Sparc2. For analog/digital conversion the on-board hardware was used at 16kHz sampling rate at 16bit resolution. For the second batch of data collection (2003-2012), i.e. for the languages Bulgarian, Hausa, Swahili, Thai, Ukrainian, and Vietnamese we used the same microphone but a modern laptop-based collection toolkit, with digital recording at 16kHz and 16bit resolution.

C. Recording Setup

Recordings were done in ordinary, but quiet rooms. The room sizes range from small to big rooms and vary between office and private rooms, in very few cases some recordings were done in public places. The surrounding noise level ranges from quiet to loud, however the majority of recordings were done in quiet environments, to not distract the subjects. The conditions in terms of noise level and recording room setup is reported for each session in the corresponding session information file (see section IV.B.3). The subjects were instructed about the equipment handling prior to the recording session. They were introduced to the projects goals and allowed to read the texts before recording.

D. Subject Recruitment

The aim of the collection was to recruit equal numbers of subjects of both genders, adult people of various ages and different education levels. To control the subjects' characteristics, a session sheet was filed for each recording session (see section IV.B.3). This sheet covers speaker characteristics, like native language, place where the speaker was raised, dialect, sex, age, and education level of the speaker, a question about the health conditions of the speaker, and whether the speaker is smoking or not. Beside the speaker's information, also the environmental setup was described, like the characteristic of the room, background noise and recording conditions (see above).

IV. Database Structure and File Formats

The final corpus consists of the information files (about speaker and session), the audio files, and the corresponding transcriptions. This section describes the naming conventions and formats of these files which are the same across all languages of the GlobalPhone corpus.

A. Naming Conventions

Language Identification Code (LID)

For each GlobalPhone language, a 2-character language ID code was given according to Table 3 below. The table also shows the official two-letter code ISO 639-1 established in 2002 and the three-letter code ISO 639-3 established in 2007. As can be seen from Table 3, most but not all GlobalPhone language ID codes follow the ISO 639 norm, which in part is due to historical reasons, the naming for GlobalPhone was started in 1995. The latest GlobalPhone collection (Swahili) now follows the three-letter code ISO 639-3.

No	*Language*	*LID*	*ISO 639-1 2-letter code*	*ISO 639-3 3-letter code*
1	**Arabic**	AR	ar	ara
2	**Bulgarian**	BG	bg	Bul
3	**Chinese Mandarin**	CH	Zh	Cmn
4	**Chinese Shanghai**	WU	-	Wuu
5	**Croatian**	CR	Hr	Hrv
6	**Czech**	CZ	Cs	Ces
7	**French**	FR	Fr	Fra
8	**German**	GE	de	Deu
9	**Hausa**	HA	Ha	Hau
10	**Japanese**	JA	Ja	Jpn
11	**Korean**	KO	Ko	Kor
12	**Polish**	PL	Pl	Pol
13	**Portuguese**	PO	pt	Por
14	**Russian**	RU	Ru	Rus
15	**Spanish**	SP	Es	Spa
16	**Swahili**	SWA	Sw	Swa
17	**Swedish**	SW	Sv	Swe
18	**Tamil**	TA	Ta	tam
19	**Thai**	TH	Th	tha
20	**Turkish**	TU	Tr	tur
21	**Ukrainian**	UA	Uk	ukr
22	**Vietnamese**	VN	Vi	vie

Table 3: Language code in GlobalPhone (LID)

Speaker Identification Code (SID)
The corpus is organized in speakers, where each speaker gets a unique 3-digit speaker identification code (SID). This SID starts at 001 for the first recorded speaker per language and is incremented for each new speaker throughout the recordings of that language.

Turn Identification (TID)
Recordings per speaker are organized and hard segmented into turns. One turn refers to the recording of one utterance, usually referring to one sentence. The turn identification (TID) is incremented for each utterance in the monologue starting at 1.

B. File Structure and Format

1. Audio File Format

The audio file format for the GlobalPhone speech data is PCM linear encoding in 16bit resolution, mono recordings (1 channel) based on 16kHz sampling rate, little endian byte-order (low-high), without any header. Files of this format are named "`.adc`". Alternatively, the data could be provided in the wav format, i.e. the audio container format based on the Microsoft defined Resource Interchange File Format (RIFF). These files are named "`.wav`" according to the common standards. They include a 44-byte header which contains the format information. Alternatively, it might be useful to deliver the data in a compressed form, for example to safe disk space and transfer time (when provided electronically via download or upload). For this purpose the audio data are available in compressed form based on the compression algorithm "Shorten" developed by Tony Robinson. The compression was performed lossless, i.e. the original adc format can be retrieved without any information loss. The files of this format are named "`.adc.shn`".

The standard way of uncompressing "`.shn`" files is to use shorten utility tools provided by Tony Robinson or otherwise on the internet. Please be aware that XLingual GmbH & Co. KG has no rights to the shorten tools. The shorten tool was originally developed by Tony Robinson and is included in many operation systems, such as Unix and derivatives. A Windows version can be freely downloaded for example from http://www.etree.org/shnutils/shorten/ for installation on a Windows machine (Windows XP, Windows 7). After installation, an interactive GUI will guide through the decompression process. The toolbox will propose settings for the decompression (PCM linear encoding, 16bit resolution, 16 kHz, mono, little-endian). We recommend to use these pre-settings for decompressing the "`adc.shn`" files.

For Unix operating systems, Unix derivatives and emulation programs, such as Cygwin under Windows, the most convenient solution is to rely on the integrated shorten tools -- cygwin for example comes with the shorten tools integrated (try > "shorten --help"). The extraction of the GlobalPhone "`.adc.shn`" files can be performed by running the command "shorten -x filename"; e.g. "`shorten -x GE001_1.adc.shn`" will result in the uncompressed file "GE001_1.adc" in the above described uncompressed adc-format.

2. Storage Media

The data are delivered on either DVD, external hard drive or transferred electronically via ftp or similar services. Each language data comes in separate directories named {`language`} and contains a Readme file (see section V.D.). Each language directory has up to four subdirectories, i.e. the directory "`spk`" which contains the speaker information and environmental conditions of a session, the directory "`trl`" which contains the transcriptions

in original script, an optional directory "rmn" which contains the corresponding transcriptions in Romanized version of this script, and the directory "adc" which contains the audio files. Alternatively, the latter directory might be "wav" which contains the wave-file container format or "adc.shn" for compressed audio files (see section above). For some languages, additional transcription formats might be provided, e.g. segmented in different word units (see Appendix for peculiarities of the GlobalPhone language packs).

The directory structure of the GlobalPhone Speech & Text database for each language looks like the following:

spk/	directory containing the speaker and recording information
trl/	directory containing the transcriptions in language specific encoding
rmn/	directory containing the Romanized transcriptions (optional)
adc/	directory containing the audio files in linear PCM encoding, mono (1 channel), 16bit resolution, 16kHz sampling rate
adc.shn/	shorten-compressed adc files (optional, see above)
wav/	adc files in wave format (optional, see above)

The directory "adc" (same for adc.shn and wav) consists of several sub-directories, one per speaker named by the 3-digit SID. The database is provided utterance wise, i.e. each spoken utterance comes with its pre-segmented audio file. This segmentation was performed automatically and the output was carefully post-processed and manually cross checked by native human listeners. Within each speaker directory are the individual audio files; the file names reflect the language, the speaker, and the turn index number as follows:

```
adc/{SID}/{LID}{SID}_{TID}.adc
```

where:

- adc/ = directory name (wav/ or adc.shn/)
- SID = speaker ID (e.g. "001")
- LID = language ID (one of "AR, CH, CR, .., VN")
- TID = turn ID (starting at "1")
- adc = adc-format audio file (wav or adc.shn)

Example: German/adc/001/GE001_10.adc
OR: German/wav/001/GE001_10.wav
OR: German/adc.shn/001/GE001_10.adc.shn

3. Speaker files spk/{LID}{SID}.spk

The speaker files are in plain ASCII containing the speaker data sheet (see section VI for the data sheet and an example for a Portuguese speaker below). The identity of the speaker has been anonymized (mapped to xxx) from the file for identity protection purposes. All information had been originally gathered and reported in German language and later been translated to English.

```
;timestamp: Tue May 26 12:06:11 MET DST 1998
;BEGIN
;LANGUAGE: Portuguese
;SUPERVISOR: Caleb EVERETT
```

;TAPELABEL: A
;RECORD DATE: June 10, 1996
;ARTICLE READORDER:1a,2a,3a,4a,5a,6a,7a
;TOPIC ARTICLE 1a:sports
;TOPIC ARTICLE 2a:economy
;TOPIC ARTICLE 3a:economy
;TOPIC ARTICLE 4a:nationalPolitics
;TOPIC ARTICLE 5a:internationalPolitics
;TOPIC ARTICLE 6a:internationalPolitics
;TOPIC ARTICLE 7a:other
;SPEAKERDATA ----------
;NAME OF SPEAKER: xxx
;SPEAKER ID:001
;NATIVE LANGUAGE: Portuguese
;RAISED IN: Rio de Janeiro
;DIALECT: Carioca
;SEX: female
;AGE:29
;OCCUPATION: nurse
;COLD OR ALLERGY: well
;SMOKER: nonsmoking
;RECORDING SETUP ----------
;RECORD PLACE: private big
;ENVIRONMENT NOISE: quiet
;RECORD CONDITIONS: Sits at a table opposite to the instructor
;COMMENTS: recording was interrupted by air plane noise
;END

4. Transcription files `trl/{LID}{SID}.trl`

The transcription files contain the spoken utterances transcribed in language specific encoding. We used the following encoding standards:

#	*Language*	*coding standard*
1	**Arabic**	ISO8859-1
2	**Bulgarian**	UTF-8
3, 4	**Chinese (MA,WU)**	Guobiao
5	**Croatian**	UTF-8 (before 2015: ISO8859-2)
6	**Czech**	UTF-8 (before 2015: ISO8859-2)
7	**French**	UTF-8 (before 2015: ISO8859-1)
8	**German**	ISO8859-1
9	**Hausa**	UTF-8
10	**Japanese**	JIS
11	**Korean**	Johabsh
12	**Russian**	KOI8
13	**Portuguese**	ISO8859-1
14	**Polish**	UTF-8
15	**Spanish**	ISO8859-1
16	**Swahili**	UTF-8
17	**Swedish**	ISO8859-1
18	**Tamil**	UTF-8 (only small parts are transcribed)
19	**Thai**	UTF-8
20	**Turkish**	ISO8859-9
21	**Ukrainian**	UTF-8
22	**Vietnamese**	UTF-8

Table 4: Encoding Standards for GlobalPhone transcriptions

All turns spoken by one speaker are listed in one transcription file. Both, the speaker's ID (SID) and the turnID (TID) are reported in the transcription file as comment lines started by ";". The file contains one turn per line preceded by the TID. The transcription files use the language specific script encoded in the coding standards described in Table 4.

The transcriptions are supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects like laughing and hesitations for most languages. Non-verbal effects are marked by <effect>, and word fragments, which occur in false starts and stuttering are marked by <effect->. The following example shows an excerpt of a Portuguese transcription file in ISO8859-1 encoding.

```
;SpeakerID 001
; 1:
nenhum dirigente nacional do PFL compareceu a convenção que sacramentou a coligacao do partido com o malufismo em São Paulo
; 2:
a cúpula do PFL preferiu um acordo com o PSDB
; 3 :
Antônio Cabrera presidente do PFL paulista afirmou que sofrera pressões e ameaças para não selar a aliança com o PPB
; 4 :
Cabrera não citou a origem óbvia das pressões mas Maluf foi direto
; 5 : todo mundo na corte e no Bandeirantes se meteu para atrapalhar disse referindo-se ao governo PHC e ao governador Mario Covas que <demi-> que demitiu os pefelistas de seu secretariado
; 6 :
na linha de negar o óbvio Maluf tentou desmentir que o PFL nacional preferisse a candidatura Serra
; 7 :
disse que estivera em Brasília com as principais lideranças <pele-> pefelistas e que não haveria restrições a Pitta
; 8:
Incomodado com a pergunta sobre a posição do PFL nacional Maluf chamou Cláudio Lembo pefelista ligado ao Marco Maciel vice-presidente da República
; 9 :
Lembo para desconforto de Maluf afirmou que o PFL local deseja a vitória da de Pitta
; 10 :
Régis de Oliveira o pefelista que é candidato a vice-prefeito na chapa de Pitta fez a iniciação na crítica ao governo PHC
; 11 :
como se socorre bancos <fa-> falidos com bilhões e se deixa outros setores desamparados
```

5. Romanized Transcription files `rmn/{LID}{SID}.rmn`

In many languages, the transcribed files are additionally converted to ASCII-7 coding using language specific mappings. The example below corresponds to the Portuguese transcription file above. For most of the languages, these language specific mappings were relatively straightforward reversible one-to-one mapping functions (see Appendix). For other languages such as Chinese, Japanese, and Korean more sophisticated, automatic transformations had been introduced. For Vietnamese, Bulgarian, and Czech no Romanization is provided. For Vietnamese, Tamil, and Korean transcripts are provided in two formats, i.e. segmented into single syllables and in concatenated multi-syllables using machine learning algorithms. For Thai we performed no romanization but segmented the script using statistical methods. For the Arabic language the transcription was directly performed in the Romanized version in order to accomplish the needs of speech recognition (the Arabic script does not write short vowels). The trl directory of Arabic contains Romanized Modern Standard Arabic transcriptions. The transcription conventions were developed in this project and provide full vocalization. These Romanized transcriptions were then automatically converted into a

second Romanized transcription that was motivated by the CallHome standards. These converted files are stored in the rmn directory. See Appendix for language specific peculiarities of the provided transcriptions.

;SpeakerID 001
; 1:
nenhum dirigente nacional do PFL compareceu a convenc:a~o que sacramentou a coligacao do partido com o malufismo em Sa~o Paulo
; 2:
a cu+pula do PFL preferiu um acordo com o PSDB
; 3:
Anto^nio Cabrera 10resident do PFL paulista afirmou que sofrera presso~es e ameac:as para na~o selar a alianc:a com o PPB
; 4:
Cabrera na~o citou a origem o+bvia das presso~es mas Maluf foi direto
; 5:
todo mundo na corte e no Bandeirantes se meteu para atrapalhar disse referindo-se ao governo PHC e ao governador Mario Covas que <demi-> que demitiu os pefelistas de seu secretariado
; 6 :
na linha de negar o o+bvio Maluf tentou desmentir que o PFL nacional preferisse a candidatura Serra
; 7 :
disse que estivera em Brasi+lia com as principais lideranc :as <pele-> pefelistas e que na~o haveria restric :o~es a Pitta
; 8:
Incomodado com a pergunta sobre a posic:a~o do PFL nacional Maluf chamou Cla+udio Lembo pefelista ligado ao Marco Maciel vice-presidente da Repu+blica
; 9 :
Lembo para desconforto de Maluf afirmou que o PFL local deseja a vito+ria da de Pitta

V. The GlobalPhone Corpus

A. Speaker information and statistics

The following tables and two graphs describe the speakers' characteristics such as gender and age distribution for all languages in the GlobalPhone corpus. Table 5 summarizes gender, age category, and smoking (y=smoker, n=nonsmoker) as well has health status (y=feels healthy, n=feels sick or has allergies). Table 6 shows the setup and environmental noise conditions during the recording sessions. The "x" indicates that information is not available.

Lid	spk	Gender			Age Category						Smoking			Healthy		
		F	M	x	<19	20-29	30-39	40-49	>=50	x	Y	n	x	y	n	X
AR	78	43	35	0	20	35	13	6	4	0	14	55	9	70	7	1
BG	77	45	32	0	7	37	8	11	14	0	15	62	0	0	0	77
CR	94	56	38	0	21	30	14	15	13	1	43	51	0	88	6	0
CZ	102	45	57	0	16	70	2	9	5	0	0	0	102	0	0	102
FR	100	51	49	0	3	52	16	13	14	2	0	0	100	0	0	100

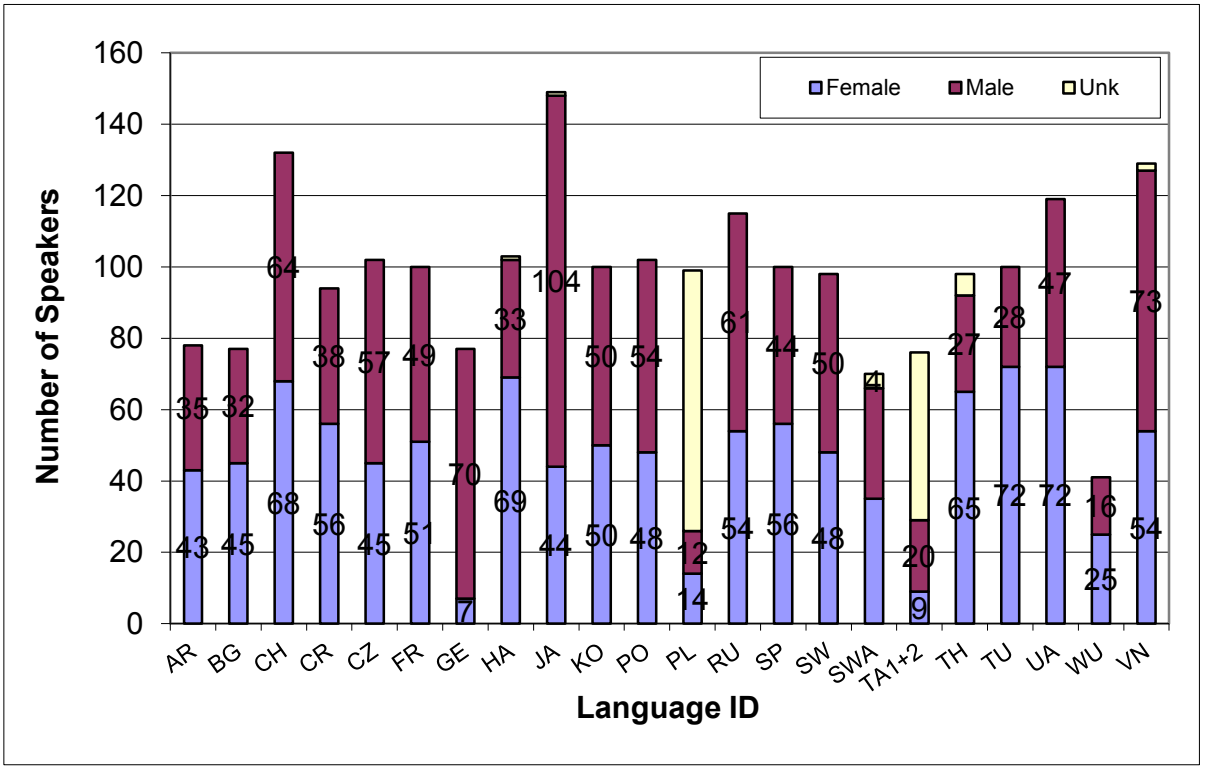
GE	77	7	70	0	0	0	0	0	0	77	0	0	77	0	0	77
HA	103	69	33	1	21	42	25	8	7	0	0	103	0	0	0	103
JA	149	44	104	1	22	90	5	2	28	2	22	95	32	98	18	33
KO	100	50	50	0	7	70	19	3	0	1	26	66	8	75	17	8
MA	132	68	64	0	16	96	16	3	0	1	10	122	0	132	0	0
PO	102	48	54	0	6	58	27	5	5	1	9	93	0	93	9	0
PL	99	14	12	73	0	12	2	9	3	73	4	29	66	28	0	71
RU	115	54	61	0	9	76	9	15	6	0	41	73	1	102	12	1
SP	100	56	44	0	20	54	13	5	8	0	14	85	1	86	13	1
SW	98	48	50	0	9	50	12	11	16	0	12	86	0	81	17	0
SWA	70	35	31	4	19	38	9	2	2	0	2	68	0	0	0	70
TA1	29	9	20	0	0	0	0	0	0	29	0	0	29	0	0	29
TA2	47	0	0	47	0	0	0	0	0	47	0	0	47	0	0	47
TH	98	65	27	6	31	67	0	0	0	0	0	98	0	0	0	98
TU	100	72	28	0	30	30	23	14	3	0	42	58	0	88	12	0
UA	119	72	47	0	9	50	19	23	18	0	4	27	88	0	0	119
WU	41	25	16	0	1	2	13	14	11	0	8	33	0	41	0	0
VN	129	54	73	2	13	71	19	15	11	0	23	88	18	0	0	129
∑	2159	1030	995	134	280	1030	264	183	168	234	289	1292	578	982	111	1066

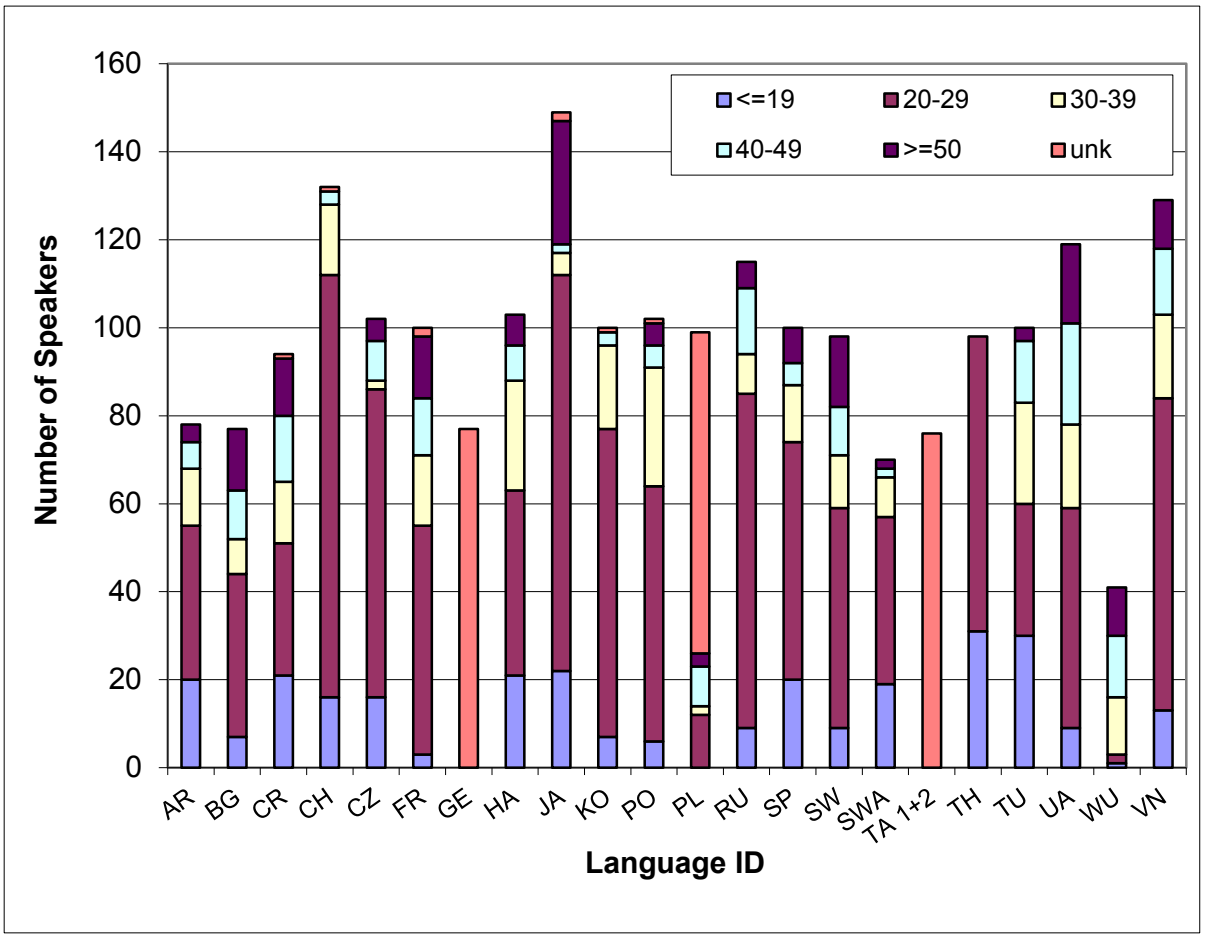
Table 5: Speaker characteristics

lid	spk	Recording Place					Environmental Noise			
		Small	Big	Public	Outdoor	x	Quiet	Middle	Loud	x
AR	78	44	17	2	7	8	60	11	6	1
BG	77	74	0	1	0	2	2	70	5	0
CR	94	69	22	3	0	0	74	16	4	0
CZ	102	0	0	0	0	102	102	0	0	0
FR	100	0	0	0	0	100	0	0	0	100
GE	77	77	0	0	0	0	77	0	0	0
HA	103	98	3	2	0	0	0	100	3	0
JA	149	114	24	0	1	10	31	102	0	16

KO	100	23	9	0	0	68	31	34	7	28
MA	132	50	12	0	0	70	13	108	11	0
PO	102	50	49	3	0	0	35	59	8	0
PL	99	0	0	0	0	99	0	0	0	99
RU	115	84	25	0	0	6	38	68	8	1
SP	100	88	3	0	9	0	85	15	0	0
SW	98	91	2	4	1	0	60	35	3	0
SWA	70	25	45	0	0	0	11	25	34	0
TA1	29	0	0	0	0	29	0	0	0	29
TA2	47	0	0	0	0	47	0	0	0	47
TH	98	0	0	0	0	98	97	1	0	0
TU	100	50	38	12	0	0	41	44	15	0
UA	119	1	3	1	0	114	2	2	0	115
WU	41	0	0	0	0	41	25	16	0	0
VN	129	111	0	0	0	18	111	0	0	18
∑	2159	1049	252	28	18	812	895	706	104	454

Table 6: Recording Conditions





B. Data Validation

For internal data validation, a three-pass approach was applied to process the data of the first collection batch: First, the speaker files were fed into the computer. Second the down-sampled DAT audio file of each speaker was split into turns by a silence detector. The speakers were instructed to pause at the end of every sentence during recording. The silence energy threshold and minimum duration of silence could be adjusted, so that each turn corresponds to one sentence. Third the sentences of the text file were assigned to the turns. The same native experts who collected the data listened to the utterances and checked if the text corresponds to the speech. Clearly audible spontaneous effects like false starts, obvious hesitations and stuttering were marked, minor differences between text and speech were corrected, incorrectly read utterances with major differences were deleted from the corpus. For the second batch of collection we used a laptop-based recording tool which recorded the utterances turn-wise in a push-to-talk scenario.

C. Training, Development, and Evaluation partition

Table 7 describes the data split into 3 sets, one training set for training the speech recognizer, one development set for testing during the systems development and one evaluation test set for reporting final numbers. The sets are disjoint, i.e. no speaker appears in more than one group and no article was read by two speakers from different groups.

Language	*Evaluation Spk*	*Development Spk*	*Training*
Arabic	027,039,108,137 + 6 TBA	005,036,107,164+ 6 TBA	All others

Bulgarian	040,059,063,068,095,109,110	051,055,058,084,090,100,106	All others
Croatian	037,038,039,040,041,042,043, 044,045,047	033,034,035,036,046,048,051,053,054, 057	All others
Czech	TBA		
French	091-098		All others
German	018,020,021,026,029,073	001,002,003,004,008,010	All others
Hausa	002, 014, 025, 028, 030, 052, 053, 062, 070, 088	018, 031, 034, 038, 046, 047, 050, 055, 058, 072	All others
Japanese	006,009,025,031,045,046,047,081,088,091,101		All others
Korean	019,029,032,042,051,064,069, 080,082,088	006,012,025,040,045,061,084,086,091, 098	All others
Mandarin	080-089	028-032, 039-044	All others
Portuguese	135,136,137,138,139,142,143, 312	064,065,072,073,102,103,104,132,133, 134	All others
Polish	TBA		
Russian	002,005,027,033,036,042,063,065,069,078,092,097,102,103,104,106,109,110,112,122		All others
Shanghai	TBA		
Spanish	011-018	001-010	All others
Swahili	TBA		
Swedish	040-044,060-064	045,046,047,048,049,066,067,068,069	All others
Thai	101-108	023,025,028,037,045,061,073,085	All others
Tamil	TBA		
Turkish	025,030,031,032,037,039,041, 046,056,063	001,002,003,005,006,008,013,014,015, 016,019	All others
Ukrainian	TBA		
Vietnamese	TBA		

Table 7: Partition in Training set, Development set, and Evaluation set

D. Miscellaneous

In some languages more than 100 speakers had been collected in total, but not all of them had been post-processed and validated so far. In order to make these files available (e.g. for unsupervised adaptation experiments, or those tasks which do not require validated transcripts), the deliverable contains an extra directory /RAW that includes the not-yet processed files. Those languages, in which these a RAW directory is available, are Arabic, Croatian, Russian, and Portuguese. For the French language a directory /PLUS is added to the deliverable, that includes phonetically balanced sentences spoken by each speaker of the database. The Czech database additionally provides 40 common “enrollment” sentences spoken by 49 speakers (speaker 001-049), which can be used for adaptation and normalization experiments.

E. Corpus Statistics

Table 8 summarizes the characteristics of the GlobalPhone corpus with respect to total size in Gigabyte, number of speakers, and spoken utterances. Numbers are given in total, and broken down into languages. The size in Gigabyte is calculated based on the uncompressed audio file. In total, more than 450 hours spoken speech had been recorded in more than 200.000 utterances of average length of 9 seconds, summing up to about 3 Million words spoken by more than 2100 native speakers.

Language	*Size*	*#speaker*	*#utterances*
Arabic	2.1 GB	78(84)	4908
Bulgarian	2.4 GB	77	8682
Croatian	1.8 GB	94(3)	4499
Czech	3.5 GB	102	12425
French	3.0 GB	98	10478
German	2.0 GB	77	10084
Hausa	975 MB	103	7895
Japanese	3.7 GB	144	13067
Korean	2.3 GB	100	8107
Mandarin	3.4 GB	132	10225
Polish	2.7 GB	99	10130
Portuguese	2.9 GB	102(14)	10344
Russian	2.9 GB	115(46)	12203
Spanish	2.4 GB	100	6898
Swahili	1.3 GB	70	7728
Swedish	2.4 GB	98	11816
Tamil	2.6 GB	- (76)	TBA
Thai	3.1 GB	98	14039
Turkish	1.9 GB	100	6950
Ukrainian	1.6 GB	119	12814
Wu	1.1 GB	41	2644
Vietnamese	2.2 GB	129	18842
Total	52,3 GB	2076 (223)	204778

Table 8: Corpus Statistics

F. Delivery Version 4

This delivery Version 4 consists of CD-ROMs, DVDs or external hard drive of uncompressed audio files, the corresponding transcribed utterances, and the original as well as the Romanized script version (where available) organized in the four directories adc, spk, rmn, trl as described above. Furthermore, the delivery contains this documentation, and selected publications about the database.

VI. Datasheet (Information on speaker and recording session)

Datasheet (to fill out for every speaker)

Name of Supervisor: ..... DAT-tape label: .....

Date: ..... recording start(time): ..... recording end(time): .....
(if possible additionally write down the tape counter)

Identification number of articles (in reading order): .....
.....
.....
(prepare an additional list with: TextID = newspaper, publication date, page, article)

Speaker Characteristics

Speakers name: ..... SpeakerID: .....

Native language: .....

Raised in: ..... Dialect there: .....

Sex: ..... Age: ..... Education: .....

Has the speaker a cold or allergy at the moment Yes () No ()

Smoker Yes () No ()

Environmental Setup

Description of environmental conditions .....
for example: big audience, small office

Background noises: quiet () middle () loud ()

Recording conditions: .....
.....
for example: test person sitting at a table, several people around

VII. Bibliography

This section contains a short list of the most relevant work related or performed based on the GlobalPhone data collection. For a most recent list of publications, please check the directory „Docs+Tools/Papers“. This directory contains all relevant and most recent work as pdf. You may also download all the papers from http://csl.uni-bremen.de Publications.

A. Corpus design and other aspects of the GlobalPhone collection

(1) Tanja Schultz: G*lobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University.* In: Proceedings of the International Conference of Spoken Language Processing (ICSLP-2002), Denver, CO, October 2002.

(2) Tanja Schultz, Martin Westphal, and Alex Waibel : *The GlobalPhone Project: Multilingual LVCSR with JANUS-3.* In : Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, pp 20--27, Plzen, Czech Republic, April 1997.

(3) Tanja Schultz and Alex Waibel : *Das Projekt GlobalPhone: Multilinguale Sprach-erkennung*. In : Computers, Linguistics, and Phonetics between Language and Speech. Proceedings of the 4th Conference on NLP (Konvens-1998), pp 179-189, Bonn, Germany, October 1998. (In German)

B. Multilingual Speech Recognition using the GlobalPhone database

(4) Tanja Schultz and Alex Waibel: *Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition.* In: Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001.

(5) Alex Waibel, Petra Geutner, Laura Mayfield-Tomokiyo, Tanja Schultz, and Monika Woszczyna: *Multilinguality in Speech and Spoken Language Systems*. In: Proceedings of the IEEE, Special Issue on Spoken Language Processing, Volume 88(8), pp 1297-1313, August 2000.

(6) Tanja Schultz, *Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen.* In: Berichte aus der Informatik, Aachen: Shaker Verlag, 2001. (Phd Thesis, in German).

(7) Tanja Schultz and Alex Waibel: *Polyphone Decision Tree Specialization for Language Adaptation*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2000), Istanbul, Turkey, June 2000.

(8) Tanja Schultz and Alex Waibel: Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP-1998), Vol. 5 pp 1819--1822, Sydney, Australia, November 1998.

C. Language Specific Speech Recognition using the GlobalPhone database

(9) Sinaporn Suebvisai, Paisarn Charoenpornsawat, Alan Black, Monika Woszczyna, and Tanja Schultz: *Thai Automatic Speech Recognition.* In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2005), Philadelphia, PA, March 2005.

(10) Sebastian Stüker and Tanja Schultz*: Grapheme-based Russian Speech Recognition.* In: Proceedings of the International Conference on Speech and Computer (SPECOM-2004), St. Petersburg, Russia, September 2004.

(11) Kenan Çarki, Petra Geutner, and Tanja Schultz*: Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages.* In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2000), Istanbul, Turkey, June 2000.

(12) Jürgen Reichert, Tanja Schultz, and Alex Waibel: *Mandarin Large Vocabulary Speech Recognition using the GlobalPhone Database.* In: Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-1999), pp 815-818, Budapest, Hungary, September 1999.

(13) Daniel Kiecza, Tanja Schultz, and Alex Waibel: *Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR.* In: 1999 Proceedings of the International Conference on Speech Processing (ICSP-1999), pp 323-327, Seoul, Korea, August 1999.