Documentation

GLOBALPHONE: a Multilingual Text & Speech Database

Version 2.3

©XLingual, LLC
5529 Kentucky Avenue, 2nd floor, Pittsburgh, PA 15232

March 26, 2006

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 recognition as well as for language identification tasks. The entire GlobalPhone corpus enables the acquisition of acoustic-phonetic knowledge of the following 16 spoken languages: Arabic (AR), Chinese-Mandarin (CH), Chinese-Shanghai (WU), Croatian (CR), Czech (CZ), French (FR), German (GE), Japanese (JA), Korean (KO), Brazilian-Portuguese (PO), Russian (RU), Spanish (SP), Swedish (SW), Tamil (TA), Thai (TH), and Turkish (TU). In each language about 100 sentences were read from each of 100 speakers. The read texts were selected from national newspapers available via Internet to provide a large vocabulary (up to 65,000 words). The majority of read articles cover national and international political news as well as economic news from 1995-1998. The speech is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone (Sennheiser 440-6) and same recording equipment for all languages. The transcriptions are internally validated 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 complement the database. The entire GlobalPhone corpus contains over 300 hours speech spoken by more than 1500 native, adult 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), for language independent and language adaptive speech recognition, as well as for language identification tasks.

A. Language Selection

There are about 4500 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 (Chinese, Spanish, Russian, Arabic)
- Political relevance (Chinese, Arabic, Korean, Japanese, Spanish)
- Geographic coverage: European, Asian, Central Asian, Indian
- Phonetic coverage, e.g. tones (Chinese, Thai) and pharyngal sounds (Arabic)
- 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
- Morphologic variations:
    - Agglutinative languages (Turkish, Korean), compounding languages (German), not-inflecting languages (Chinese) and languages without segmentation on the word level (Chinese, Japanese, Thai)

Table 1 below shows the ranking of the most widespread languages of the world, the primary locations where the languages are spoken, and their pertaining speakers population. The languages covered by the GlobalPhone corpus are marked by "*".

Rank	*Language*	*Primary Locales*	*Speakers*	*Language Group*
1. *	Mandarin	China	907 Mio	Sino-Tibetan (Sinitic)

2. (*)	English	USA, UK, Can, Australia	456 Mio	Indo-European (Germanic)
3.	Hindi	India	383 Mio	Indo-European (Indo-Iran)
4. *	Spanish	Latin-America, Spain	362 Mio	Indo-European (Romance)
5. *	Russian	Russia, Indep. States	293 Mio	Indo-European (Slavic)
6. *	Arabic	N. Africa, Mid East	208 Mio	Afro-Asiatic (Semitic)
7.	Bengali	Bangladesh, India	189 Mio	Indo-European (Indo-Iran)
8. *	Portuguese	Brazil, Portugal, Angola	177 Mio	Indo-European (Romance)
9.	Malay-Indo.	Indonesia, Malay, Brunei	148 Mio	Austronesian (Polynesian)
10. *	Japanese	Japan	126 Mio	Isolate
11. *	French	F, Can, Africa, Switzerland	123 Mio	Indo-European (Romance)
12. *	German	G, Austria, Switzerland	119 Mio	Indo-European (Germanic)
15. *	Korean	Korea, China	73 Mio	Isolate
17. *	Tamil	India, SriLanka, Malaysia	67 Mio	Dravidian
20. *	Wu/Shanghai	China (Shanghai)	64 Mio	Sino-Tibetan (Sinitic)
25. *	Turkish	Turkey	57 Mio	Altaic (Turkik)
43. *	Serbo-Croatian	Balkan Europe	20 Mio	Indo-European (Slavic)
85. *	Swedish	Sweden, Finland	9 Mio	Indo-European (Germanic)

Table 1: Most widespread languages of the world

B. Task and Text Selection

In order to limit the effort of transcription which is the most time and cost consuming process of a database collection and to ease the collection of large amounts of similar text data, the data collection consists of read data from electronically available text sources. The texts were selected from national newspapers available via Internet. The texts were chosen from national and international political and economic topics to somewhat restrict the vocabulary. Most of the articles were selected between May 1996 and November 1997, which makes it possible to compare the usage of proper names (Politicians, companies, etc.) across languages.

For the GlobalPhone corpus we used the following newspapers: Assabah for Arabic, 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 Sueddeutsche Zeitung for German, Hankyoreh Daily News for Korean, Nikkei Shinbun for Japanese, Folha de Sao Paulo for Portuguese, Ogonyok Gaseta and express-chronika for Russian, La Nacion for Spanish, Goeteborgs-Posten for Swedish, Thinaboomi Tamil Daily for Tamil, and Zaman for Turkish. Table 2 gives the web addresses for the newspapers.

Language	Internet link
Arabic	http://www.tunisie.com/Assabah
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
Japanese	http://www.nikkeihome.co.jp
Korean	http://news.hani.co.kr
Portuguese	http://www.uol.com.br/fsp
Russian	http://www.ropnet.ru/ogonyok
Spanish	http://www.nacion.co.cr
Swedish	http://www.gp.se
Tamil	http://www.thinaboomi.com
Turkish	http://www.zaman.com.tr

Table 2: Newspaper links

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 from May 1996 to November 1997. During this time we collected Arabic speech in Tunis, Sfax and Djerba, Tunisia; Mandarin in Beijing, Wuhan and Hekou, China; Shanghai dialect in Shanghai, China; Croatian in Zagreb, Croatia, and parts of Bosnia; Czech at the Charles University in Prague, Czech Republic in 1999; French in Grenoble, France; German in Karlsruhe, Germany; Japanese in Tokyo, Japan; Korean in Seoul, Korea; Portuguese in Porto Velho and Sao Paulo, Brazil; Russian in Minsk, Belarus; Spanish in Heredia and San Jose, Costa Rica; Swedish in Stockholm and Vaernamo, Sweden; Tamil in India, Thai in Bangkok, Thailand, and Turkish in Istanbul, Turkey.

B. Recording Equipment

The recording equipment 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 sampled down to 16kHz sampling rate. The recording equipment is identical across all GlobalPhone languages but German. For the German part we used the same microphone but speech was recorded on a SUN Ultra-Sparc2. For A/D we used the on-board hardware at a 16kHz sampling rate at 16bit resolution.

The audio file format for the GlobalPhone speech data is PCM waveform files 16kHz, 16bit mono quality, byte-order low-high, without header. All files are lossless compressed using the "Shorten" algorithm written by Tony Robinson, see http://www.softsound.com/Shorten.html.

C. Recording Setup

All recordings were done in ordinary, but quiet rooms. The rooms range from small to big and vary between office and private rooms, in very few cases the 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, so that the speakers were not distracted. The quality of noise level and recording room setup is reported for each session (see section IV.B.3). The speakers were given instructions about the equipment handling in advance. They were introduced to the projects goals and are allowed to read the texts before recording.

D. Subject Recruitment

The aim of the collection was to recruit equal numbers of subjects of both sexes, adult persons of various age categories and different education levels. In order to control the subjects' characteristics, a session sheet was filled about each recording session (see section IV.B.3). This sheet contains 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

The final corpus consists of speakers' information files, audio files, and the according transcriptions. This section describes the naming conventions and formats of these files.

A. Naming Conventions

Language Identification (LID)
A 2-character language ID code was given according to the following list:

Language	*LID*
Arabic	AR
Chinese Mandarin	CH
Chinese Shanghai	WU
Croatian	CR
Czech	CZ
French	FR
German	GE
Japanese	JA
Korean	KO
Russian	RU
Portuguese	PO
Spanish	SP
Swedish	SW
Tamil	TA
Thai	TH
Turkish	TU

Table 3: language ID

Speaker Identification (SID)
The 3-digit speaker ID 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)
The turn ID is incremented for each utterance in the monologue starting at 1.

B. File Structure

1. Storage Media

The data are delivered on either CD-ROM or DVD. They are named according the rule:

```
{language}1_n, {language}2_n, ..., {language}(n-1)_n, {language}n_n
```

where n is the total number of CD-ROMs/DVDs for this language. Each CD-ROM/DVD has four directories and a Readme file (see section V.D). The directory `spk/`contains the speaker and environmental conditions of a session, the directory `trl/` contains the transcriptions in original script, the directory `rmn/` the corresponding romanized version of this script. The directory `adc/` contains the compressed audio files.
Directory structure of CD-ROMs/DVDs:

spk/	directory containing the speaker and recording information
trl/	directory containing the transcriptions in language specific encoding
rmn/	directory containing the romanized transcriptions

adc/ directory containing the audio files

The directory `adc/` consists of several sub-directories, one per speaker named by the 3-digit SID. 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.shn
```

where:

adc/ = directory name
SID = speaker ID (e.g. "001")
LID = language ID (one of "AR, CH, CR, .., TU")
TID = turn ID (starting at "1")
adc.shn = compressed audio (shorten by T. Robinson)

Example: `Mandarin/adc/001/CH001_10.adc.shn`

2. Audio files

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

contains the audio of one spoken turn TID of speaker SID. The audio format is PCM 16bit 16kHz byte-order low-high lossless compressed with the program "shorten" written by Tony Robinson (see http://www.softsound.com/Shorten.html).

3. Speaker files

`spk/{LID}{SID}.dat`

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 removed (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, except for the comments.

```
;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:Sitzt an einem Tisch gegenueber des Testers
;COMMENTS:Sprecherin wurde durch ein vorbeifliegendes Flugzeug gestoert
;END
```

4. Transcription files

`trl/{LID}{SID}.trl`

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

Language	*coding standard*
Arabic	ISO8859-1
Chinese	Guobiao
Croatian	ISO8859-2
Czech	ISO8859-2
French	ISO8859-1
German	ISO8859-1
Japanese	JIS
Korean	Johabsh
Russian	KOI8
Portuguese	ISO8859-1
Spanish	ISO8859-1
Swedish	ISO8859-1
Tamil	No transcripts yet
Thai	UTF-8
Turkish	ISO8859-9

Table 4: Coding Standards

The transcription file contains all turns spoken by one speaker. 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. Non-verbal effects are marked by `<effect>`, and word fragments, which occur in false starts and stuttering are marked by `<effec->`. The following example shows an excerpt of a Portuguese transcription file in ISO8859-1 encoding.

```
;SprecherID 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`

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 simple one-to-one mapping functions (see Appendix). For other languages such as Chinese, Japanese, and Korean more sophisticated, automatic transformations had been introduced. 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 (see Arabic Appendix). These converted files are stored in the `rmn` directory.

```
;SprecherID 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 presidente 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
; 10:
Re+gis de Oliveira o pefelista que e+ candidato a vice-prefeito na chapa de
Pitta fez a iniciac:a~o na cri+tica ao governo PHC
; 11:
como se socorre bancos <fa-> falidos com bilho~es e se deixa outros setores
desamparados
```

V. The GlobalPhone Corpus

A. Speaker information and statistics (Updated for all languages)

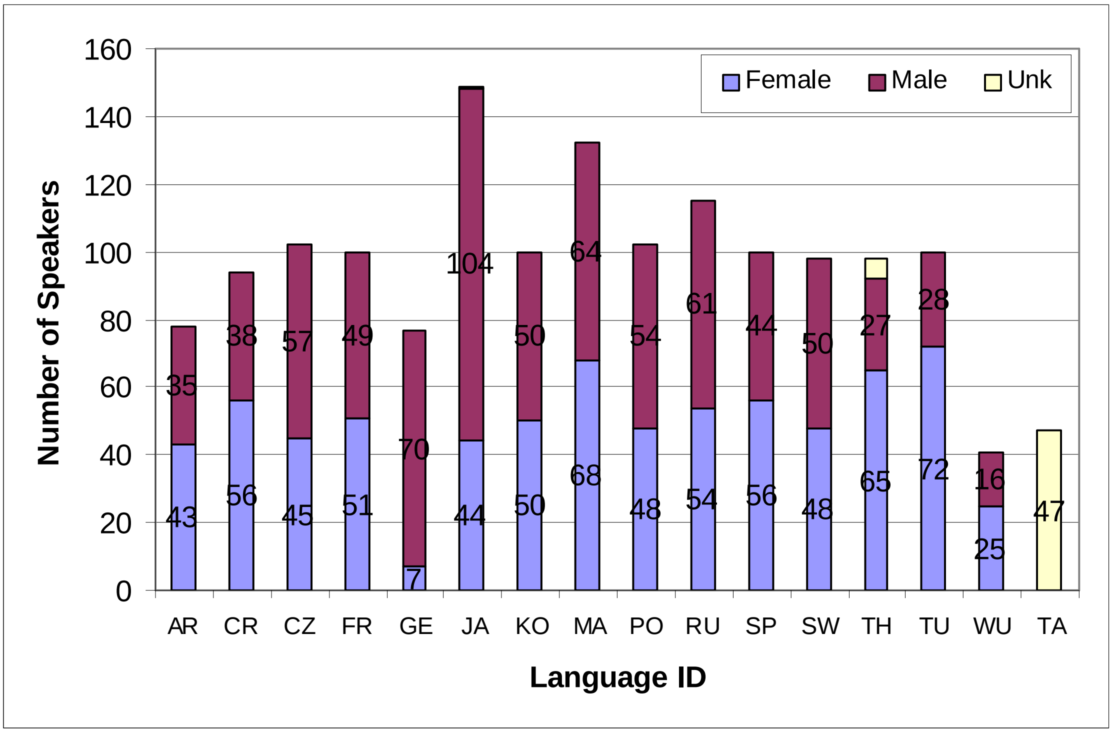
The following table and two graphs describe the speakers' characteristics such as the gender and age distribution for all languages in the GlobalPhone corpus. The first table shows gender, age category, and smoking (y=smoker, n=nonsmoker) as well has health status (y=feels healthy, n=feels sick or has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

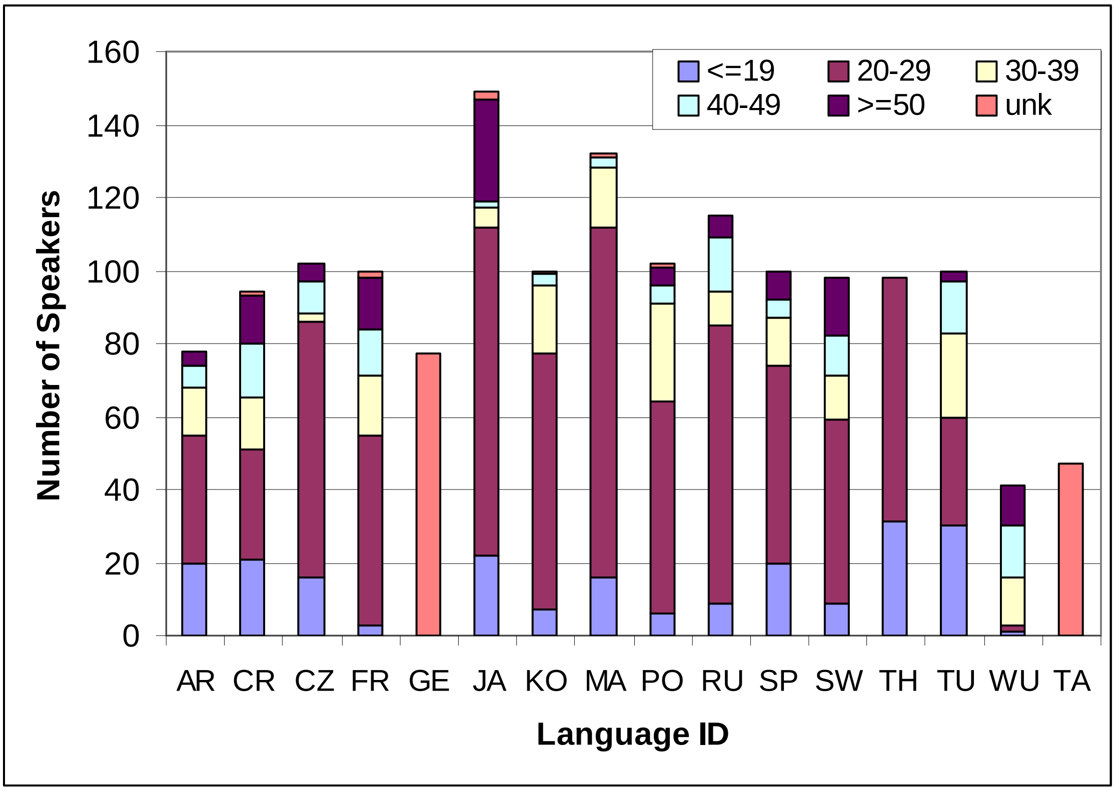
lid	spk	Gender			Age Category						Smoking			Healthy		
		F	M	x	<19	20-29	30-39	40-49	>=50	x	y	n	x	y	n	x
AR	78	43	35	0	20	35	13	6	4	0	14	55	9	70	7	1
CR	94	56	38	0	21	30	14	15	13	1	43	51	0	88	6	0
CZ	102	45	57	0	16	70	2	9	5	0	0	0	102	0	0	102
FR	100	51	49	0	3	52	16	13	14	2	0	0	100	0	0	100
GE	77	7	70	0	0	0	0	0	0	77	0	0	77	0	0	77
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
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
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
WU	41	25	16	0	1	2	13	14	11	0	8	33	0	41	0	0
TA	47	0	0	47	0	0	0	0	0	47	0	0	47	0	0	47
∑	1533	732	747	54	211	780	182	115	113	132	241	915	377	954	111	468

Table 6: 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
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
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
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
TH	98	0	0	0	0	98	97	1	0	0
TU	100	50	38	12	0	0	41	44	15	0
WU	41	0	0	0	0	41	25	16	0	0
TA	47	0	0	0	0	47	0	0	0	47
∑	1533	740	201	24	18	550	769	509	62	193

Table 7: Recording Conditions





B. Data Validation

For internal data validation, a three-pass approach was applied to process the data: 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, incorrect read utterances with major differences were deleted from the corpus.

C. Training, Development, and Evaluation partition

Table 5 describes how the data was split up into 3 sets, one training set for training the speech recognizer, one development set for testing during the systems development and on evaluation test set for reporting final numbers. The sets split up in a way that 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 Spk
Arabic	027,039,108,137 + 6 TBA	005,036,107,164+ 6 TBA	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
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
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
Swedish	040-044,060-064	045,046,047,048,049,066, 067,068,069	All others
Thai	015, 022, 023, 042, 046, 068, 081,088	013, 016, 031, 038, 051, 059, 070, 080	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

Table 8: 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 6 summarizes the characteristics of the GlobalPhone corpus with respect to total size in Gbyte, number of speakers, and spoken utterances. Numbers are given in total, and broken down into languages. The size in Gbyte is calculated based on the compressed audio files (see section B, Shorten). In total, more than 270 hours spoken speech had been recorded in more than 112.000 utterances of average length of 9 seconds, summing up to about 2 Million words spoken by more than 1300 native speakers.

Language	*Size*	*#speaker*	*#utterances*
Arabic	1.5GB	78(84)	4908
Croatian	1.1GB	94(3)	4499
Czech	1.9GB	102	12425
French	1.9GB	98	10273
German	1.3GB	77	10085
Japanese	2.1GB	144	13067
Korean	1.3GB	100	8107
Mandarin	2.2GB	132	10225
Portuguese	2.2GB	102(14)	10417
Russian	2.6GB	115(46)	12205
Spanish	1.5GB	100	6898
Swedish	1.4GB	98	11816
Tamil	1.1GB	- (47)	TBA
Thai	1.7GB	98	14039
Turkish	1.1GB	100	6950
Wu	0.6GB	41	2644
Total	28GB	1479(194)	124109

Table 9: Corpus Statistics

F. *Delivery Version 2*

This delivery Version 2 consists of CD-ROMs or DVDs of compressed audio files, the corresponding transcribed utterances, and the original as well as the romanized script version organized in the four directories adc, dat, rmn, trl as described above. Furthermore the CD-ROMs or DVDs contain this documentation, selected publications on the database, and the compression tool “shorten” from Tony Robinson (this tool was downloaded from the web site http://www.softsound.com/Shorten.html and is provided as a courtesy to costumers, no rights whatsoever belong to XLingual, LLC). The following Readme file goes with each distribution.

```
###################################################################
#                                                                 #
#         Multilingual Text & Speech Database GLOBALPHONE         #
#                         Version 2.2                             #
#                                                                 #
#                   X L I N G U A L,  L L C                       #
#                         March 2006                              #
#                                                                 #
###################################################################

   This CD-ROM contains the multilingual text and speech database
   GLOBALPHONE, produced by XLingual, LLC.

   GLOBALPHONE  provides  multilingual  speech and text data in 16
   languages Arabic, Chinese-Mandarin, Chinese-Shanghai, Croatian,
   Czech, French, German,  Japanese, Korean, Portuguese,  Russian,
   Spanish,  Swedish, Tamil, Thai, and Turkish.  In each  language
   news article sentences were read by  about 100 native speakers.
   The articles cover national and international political news as
   well as economic news from  1995-2005.  The speech is available
   in  16bit,  16kHz mono quality,  recorded with a close-speaking
```

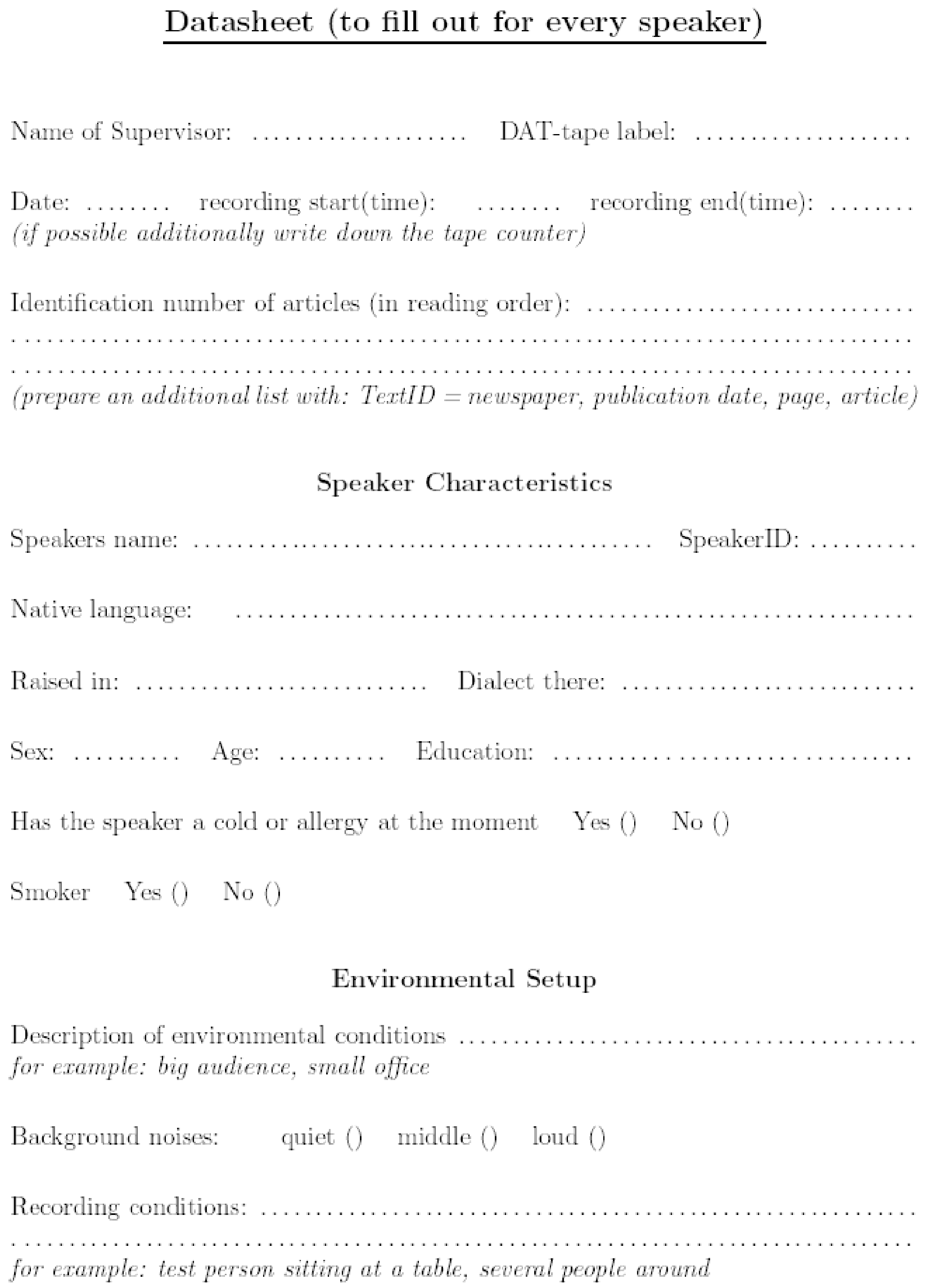
microphone and same recording equipment for all languages.

All use of this corpus is subject to a license agreement.
Copyright holder of the data is
XLingual, LLC
5529 Kentucky Avenue, 2nd floor, Pittsburgh, PA 15232
Tanja Schultz (tanja@xlingual.com)

No parts of this text & speech database may be distributed or reproduced in any form or by any means without permission from the copyright holder. A copy of this README file should be kept with the data.

```
###################################################################
#  README for the Multilingual Text & Speech Database GLOBALPHONE #
###################################################################
```

VI. Datasheet



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

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.

Annex: Arabic

Collection Site

For the Arabic language an additional data collection was performed. The original, first data collection took place in July and August 1996 in Tunis, Sfax, and Djerba in Tunisia. In this collection 94 Tunisian speakers (SpeakerID 001-100) speaking Modern Standard Arabic were recorded. The second collection took place in Palestine between August and October 1999, where another 65 speakers (SpeakerID 101-165) from Palestine and Jordan were recorded.
Recording equipment and setup are identical among those collections.

Transcription conventions – trl files

The transcription of the Modern Standard Arabic recordings was not done in the traditional Arabic script but in a romanized form. This decision was motivated by the fact that most of the Arabic text resources (such as newspaper text) are written in Arabic script without diacritics, which means that this form does not provide the full vocalization of words (short vowels are missing). In order to better serve speech recognition purposes, a roman transcription convention was developed in this project, which does provide the full vocalization. As a consequence, the pronunciation of words can be easily derived from the transcriptions.

The conventions were inspired by Qalam and CAT (Classical Arabic Transliteration), two transliteration conventions originally developed to allow an ASCII-based exchange of Arabic text in computers. However, these conventions had to be changed (1) in order to better reflect the pronunciation, and (2) to allow unambiguous mapping between scripts. The table below shows the correspondence of Arabic and Roman characters used to transcribe the `trl` files.

Character	Qalam	CAT	GlobalPhone
aleef	aa	aa	A
baa	b	b	B
taa	t	t	T
thaa	th	th	Tt
jym	j	j	J
haa	H	h	H
khaa	kh	k	Kk
daal	d	d	D
dhaal	dh	z^|zh	Dd
raa	r	r	R
zayn	z	z	Z
syn	s	s	S
shyn	sh	s^|sh	Sc
saad	S	s|S	Ss
daad	D	d|D	Sd
taa	T	t|T	Td
zaa	Z	z|Z	Dt
ayn	'	@	Ar
ghayn	gh	g|gh	G
faa	f	f	F
qaaf	q	q	Q
kaaf	k	k	K
laam	l	l	L
mym	m	m	M
nuwn	n	n	N
haa	h	h	h
waaw	w	w|uu|oo	W
yaa	y	y	Y

The example below shows the transcription of the utterances spoken by speaker 001 as can be found in file `AR001.trl`.

```
;SprecherID 001
; 1:
ALTKWYN ALMHNY
; 2:
NiD~aAM JaDiYD LiL-TTaSdaRRuF FiY MaRaAKiZi AL-TTaKWiYN
; 3:
AeaBRaMaT BiLaADuNaA FiY AeaWaAIiLi HaDhaA AL-ShShaHR AeiTTiFaAQa QaRDd
Ma3a AL-BaNKi AL-DDuWaLiY LiL-AeiNShaAE Wa AL-TTa3MiYR QaSdDa Ta7WiYLi
JuZEiN HaAMMiN MiNa AL-MaShRuW3 AeaL-ThThaANiY LiL-TTaKWiYN Wa AL-
TTaShGhiYL AeaLLaDhiY YaHDiFu AeiLaU Ta3SdiYR JiHaAZ AeaL-TTaKWiYNi AL-
MiHaNiY Wa AL-TTaShGhiYL BuGhYaTea AL-TTaRFiY3i FiY QuDRaTei AL-
AeiQTiSdaADi AL-WaTtaNiY 3aLaU MuJaABaHaTei AL-MuNaAFaSa(Te) Wa DhaLiKa <-
Bi-AeiDdFaA_Ei> MaZiYDiN MiNa AL-MuLaAEaMa(Te) BaYNa HaYaAKiLi AL-TTaKWiYN
Wa AL-TTaShGhiYL Ma3a 7aAJiYaAT AeaL-MuNShaA~T Wa Bi-Ta7SiYNi
AeiNTaAJiYYaTei AL-YaD AL-3aAMiLa(Te) Wa 7aRaKiYYaTiHaA
; 4:
HaDhaA Wa YaTaDdaKMMaNu AL-MaShRuW3 3iDDaTe MuKaWWiNaAT AeaSaASiYYa(Te)
MiNHaA
```

Transcription conventions – rmn files

In addition to the transcriptions described above, the rmn files provide another representation of the spoken utterances. The main purpose of this representation is to make it easier to merge the GlobalPhone data with other corpora and with other dialects of spoken Arabic. These conventions are based on the LDC CallHome conventions. Some enhancements were developed primarily for two reasons: first, to make it easier for transcribers to encode the data consistently, and second, to capture distinctions necessary for speech synthesis. The latter is important to realize seamless integration of recognition, translation, and synthesis in a speech-to-speech translation framework. The conventions are described in large detail in the attached document “EGA-conventions.pdf”.

The mapping between the two conventions was done automatically. The example below shows the correspondent transcription of the utterances spoken by speaker 001 as can be found in file `AR001.rmn`.

```
;SprecherID 001
; 1:
Altkwyn Almhny
; 2:
niZAm jadEd li+il+ttaSarruf fE marAkizi il+ttakwEn
; 3:
abramat bilAdunA fE awACili haVA il+$$ahr ittifAQa QarD maca il+banki
il+ddUlE li+il+Cin$AC wa il+ttacmEr QaSda taHwEli juzCin hAmmin mina
il+ma$rUc al+FFAnE li+il+ttakwEn wa il+tta$GEl allaVE yahdifu ila tacSEr
jihAz al+ttakwEni il+mihanE wa il+tta$GEl buGyaB(t)a il+ttarfEci fE
QudraB(t)i il+CiQtiSAdi il+waTanE cala mujAbahaB(t)i il+munAfasaB wa Valika
<bi+CiDfACi> mazEdin mina il+mulACamaB bayna hayAkili il+ttakwEn wa
il+tta$GEl maca HAjiyAt al+mun$CAt wa bi+taHsEni intAjEyaB(t)i il+yad
il+cAmilaB wa HarakEyatihA
; 4:
haVA wa yataDakmmanu il+ma$rUc ciddaB(t) mukawwinAt asAsEyaB minhA
```

Annex: Portuguese

Portuguese Romanization

The following list describes the original characters/diacritics used in Brazilian Portuguese script, the corresponding ISO8859-1 code and romanized form as used in the romanized transcription files and the pronunciation dictionary.

Original Character	ISO-code	Romanized character
á	\341	a+
Á	\301	A+
é	\351	e+
É	\311	E+
í	\355	i+
Í	\315	I+
ó	\363	o+
Ó	\323	O+
ú	\372	u+
Ú	\332	U+
ã	\343	a~
Ã	\303	A~
õ	\365	o~
Õ	\325	O~
ü	\374	u^
Ü	\334	U^
à	\340	a:
À	\300	A:
ç	\347	c:
Ç	\307	C:
â	\342	a^
Â	\302	A^
ê	\352	e^
Ê	\312	E^
ô	\364	o^
Ô	\324	O^