Collection of SLR in the Asian-Pacific area

Asunción Moreno[1], Khalid Choukri[2], Phill Hall[3], Henk van den Heuvel[4], Eric Sanders[4], Francesco Senia[5], Herbert Tropf[6]

[1] UPC, Spain; [2] ELDA, France; [3] Appen, Australia; [4] SPEX, The Netherlands; [5] Loquendo SpA, Italy; [6] Siemens AG, Germany

asuncion@gps.tsc.upc.es

Abstract

The goal of this project (LILA) is the collection of a large number of spoken databases for training Automatic Speech Recognition Systems for telephone applications in the Asian Pacific area. Specifications follow those of *SpeechDat*-like databases. Utterances will be recorded directly from calls made either from fixed or cellular telephones and are composed by read text and answers to specific questions. The project is driven by a consortium composed by a large number of industrial companies. Each company is in charge of the production of two databases. The consortium shares the databases produced in the project. The goal of the project should be reached within the year 2005.

1. Introduction

According to ELRA recent surveys, we do expect that, in the coming years, Human Language Technologies (HLT) developers will need resources to develop basic technologies related to speech recognition (areas such as telephony, consumers, cars, news transcription and information retrieval, etc.) , speech synthesis (including multiple voices, voice conversion), speech to speech translations, for areas with dialogue capabilities. This will require the development of multi speaker speech databases, Onomastica/pronunciation lexica, (morpho-syntactically tagged) text corpora, preferably of transcribed texts, prosodically tagged lexica, aligned textes, etc.

To do so we need to address such issues for all the languages, probably not the 6000 existing languages but at least the most important ones that could be around 300-500 according to the current trend of globalization. Such approach should consider market sectors like Office, Fixed and Mobile Telephony, In-Car applications, "Other" Consumer applications. After Europe (including East), America and Oriental regions, some partners are ready to move to LILA and then to other places.

The catalogue of ELRA [1] includes telephone SpeechDat-like [2] SLR for Shangai Mandarin, Mandarin and Cantonese. Soon microphone speech databases in Japanese, Cantonese, Korean, Mandarin, Taiwan Mandarin and Thai as collected in the SpeeCon project will be made publicly available. However, many countries and applications are not covered in the ELRA catalogue. The goal of this project [3] is the collection of a large number of spoken databases for training Automatic Speech Recognition Systems in the Asian Pacific area.

The project is funded by an Industrial consortium. The consortium is open to industrial or public members. All members have access to the databases produced within the consortium.

The consortium is composed by Loquendo SpA (Italy), Microsoft (US), Motorola (US), NSC (Israel), Phonetic Systems (Israel), Scansoft (Belgium), Siemens AG (Germany) and Telisma (France). The University Polytechnic of Catalonia (UPC) (Spain) coordinates the project. Validation of the databases is performed by the Dutch institute SPEX.

Each company is in charge of the production of one or more databases. The consortium shares the databases produced in the project. The goal of the project should be reached within the year 2005.

A production model is being studied to shorten the production time and save cost The production model is performed by ELDA (France), Appen (Australia), ATLAS (Spain), and Human Voices (Israel). Databases will be available at ELRA catalogue at the end of the project.

The paper describes the project. Section 2 describes the recording areas, languages to record and specifications. Section 3 describes in detail one of the countries to record, its economic factors, population, languages and dialects distribution, and population. Section 4 describes the validation issues to assure quality in the recorded databases and finally, section 5 describes the main issues of the production model.

2. Specifications

The project at its first stage has started a study on the different languages spoken in the various countries of the Pacific Asia area. Each country has been analysed taking into account the differences in its internal population, people origins, religions and racial differences, i.e. all those aspects that may influence a language and its linguistic variations. Furthermore an overview of the economic aspects of each country in terms of Gross Domestic Product, Industrial production growth rate, and other economical indexes as well as status of fixed and mobile telephone network (number of networks, nodes and subscribers) has been given. When available, the economical/technical forecasts for next years have been supplied too.

The following countries have been analysed: Burma (Myanmar), China – Taiwan – Hong Kong, Japan, India, Indonesia, Laos, Malaysia, Philippines, South Korea and Thailand. All reports are available in the WEB site of the consortium [3] in the documents section.

The reports have obviously described the well known presence on English in the old British colonies that in

some cases is not only used as official language (or unofficial lingua franca) but also in TV and radio advertisements in spite of the fact that the official language is different from English. Furthermore English is often used to communicate between people belonging to different racial groups within the same country.

In some countries it is still not clear which is the official language because of the colonialism and its internal story; so far, any language has imposed over the others and, as in the case of India, a number of different languages coexist in the same area because spoken by people belonging to different racial groups. Nevertheless some languages are more concentrated in some particular area and this has helped the consortium in compiling a first list of recording languages and regions.

Based also on the specific interest of the companies involved into the project, the following countries / languages have been selected:

Country	Languages	# speakers to record
China + Taiwan	Mandarin	2000
	Mandarin (Guoyu)	1000
Japan	Japanese	2000
India	Hindi (first)	3000
	Hindi (second)	3000
	English	3000
South Korea	Korean	1000
Thailand	Thai	1000
Australia	English	1000
New Zealand	English	1000

Table 1: Number of speakers to record per country and per language

In spite of its size, we have decided to recruit a reduced number of speakers in **China** because some speech databases in different languages are already available on the market (see above). So we have decided to concentrate on **Mandarin** by recording a first database in **Mainland China** and a smaller one in **Taiwan**.

Because its supposed high cost, we have decided to record only a 2000-speaker speech database in **Japan** from three main dialectal areas (Tokyo, Kansai and Kyushu).

India is the most selected country because of its size, the number of the different languages and his potentiality in terms of economic growth. At the same time there aren't so many linguistic resources. We intend to record three databases: two 3000-speaker **Hindi** speech database and one 3000-speaker **English** speech database. Details on this country are just below in the next section.

A 1000-speaker **Korean** speech database is supposed to be recorded by recruiting speakers from the 5 major dialects spoken in **South Korea**.

In **Thailand** we intend to record a 1000-speaker database in **Thai language** that is the official language of this country. People from four different areas will be recruited: North, North East, Centre and South.

Recently two other databases have been added to the above list by including two important countries although exactly belonging to the Asian Pacific area; because of their importance we have decided to record a 1000-speaker speech database in **Australian English** as well as another 1000-speaker speech database in **New Zealand English**.

At the moment we have discarded a number of countries/languages because of less relative interest or simply because lack of potential partners wishing joining this ambitious project but, obviously, its members are opened to evaluate any request useful to increase the overall value of the project. Some technical aspect are not clear at the moment, such us the phonetic structure of the languages, the script to use and were to find good written linguistic material for some minor languages but some of the partners have experience in managing non-Latin languages; in fact, they have participated recruiting campaign in Arabic country were similar problems have been faced, so we suppose to complete the specifications quickly.

3. Showcase India

Up to now no speech databases are publicly available which could be used to train and test successfully performing speech recognizers for any language spoken in India. This fact and the sheer size of the potential market for speech enhanced products prompted the LILA consortium to have India highly prioritized on the list of countries for which appropriate speech databases will be collected.

3.1 Some relevant facts about the country

India is subdivided in 28 states and 7 union territories. The size of the population is 1.05 billion and the age structure is as follows: 0-14 years 32.2%, 15-64 years 63%, 65 years and over 4.8% (2003 estimation). The religions are Hindu (81.3%), Muslim (12%), Christian (2.3%), Sikh (1.9%) and other groups including Buddhist, Jain, Parsi (2.5%).
Concerning GDP the purchasing power parity is USD 2.664 trillion. Per capita the purchasing power parity is USD 2,600, and the real growth rate is 4.3% (all 2002 estimation). 25% of the population lives below the poverty line. (For these and related facts cf. [4].)
Liberalization of the Indian telecom sector began in 1994 and become more serious in 2000 under the regulation of the Ministry of Communications. Mobile penetration is still 1%, while fixed-line is now around 4%-5%. These still very low numbers show the stage of the telecommunications industry in general, but also India's vast potential (cf. [5]).

3.2 The languages in India

The languages spoken on the Indian subcontinent mainly belong to 3 major families: Indo-Aryan (a branch of Indo-European), Dravidian and Tibeto-Burman. But there are also a few languages which belong to the Austro-Asiatic family. The Indo-Aryan languages are mainly spoken in the northern and central parts of India, whereas Dravidian

languages are spoken in South India with some isolated groups of speakers in the north. Speakers of Tibeto-Burman languages live along the Himalayan fringe. The Austro-Asiatic languages are spoken by groups of tribal peoples from West Bengal to Madhya Pradesh (cf. [6, 7]). The approximate spreading of these language families is indicated on the map below by different shades of grey (cf. [8]).

The Indian census records over 200 different "mother tongues" among which are the 18 "scheduled" languages, i.e. officially recognized by the constitution. About 75% of all Indians speak an Indo-Aryan language, and about 23% speak a Dravidian language. The Tibeto-Burmese and the Austro-Asiatic group embrace about 1% of the speakers each (cf. [9]).

The language which by far is spoken by most of the Indians is Hindi (40.2%), which belongs to the Indo-Aryan family, followed by Bengali (8.3%), Telugu (7.9%), Marathi (7.5%), Tamil (6.3%), Urdu (5.2%) etc. (source: 1991 census of India).

Hindi is the majority language in several states of India (cf. map below indicating the majority language spoken in each state). Some other states were created based on language boundaries. It is not surprising that the main languages have different regional variants, e.g. Hindi spoken in Rajasthan is different from Hindi spoken in Bihar of Himachal Pradesh. Linguistic diversity does not only reflect regional differences but also intricate levels of social hierarchy and caste. Nevertheless, most of the official languages have reached a standard of speaking language which on the whole has become the accepted style of speaking. For Hindi e.g. this "standard" is based on the dialect of the Delhi-Agra region.

According to the 1991 census of India about 50 million speakers have Hindi as second language and about 21 million speakers have it as third language. Depending on the first language the command of Hindi as second (or third) language varies considerably, e.g. from 0.7% for Tamil speakers to 44.4% for Sanskrit speakers.

And according to the same census about 65 million speakers (8.0%) have English as second language and about 25 million speakers (3.1%) have it as third language. Depending on the first language the command of English as second (or third) language varies considerably, e.g. from 1.5% for Gujarati speakers to 22.0% for Malayalam speakers.

The choice of the language which could or should serve as medium of common communication for whole India – namely Hindi or English − is an unresolved controversial issue of Indian language policy. To take Hindi, an Indo-Aryan language, could mean the Dravidian speaking population and its culture would be dominated. English on the other hand remains a foreign tongue left over from the British rule, and it is used fluently only by a relatively small, privileged segment of the population.

For the official languages 10 Indic scripts are applied, where the most commonly used script Devanagari is shared by several Indo-Aryian languages (cf. [10]).

MAJORITY LANGUAGES OF INDIAN STATES
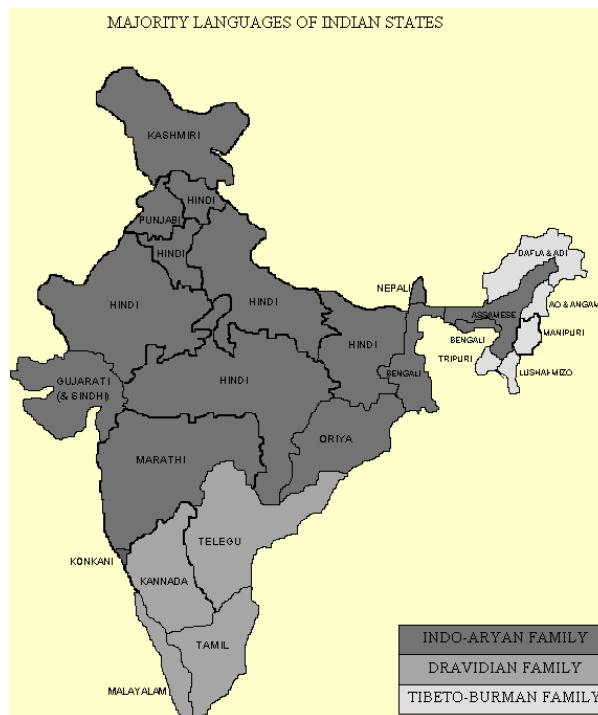
3.3 Indian Speech databases to be collected

In the year 1991-92 the Technology Development for Indian Languages (TDIL) program was launched by the Ministry of Information Technology, Govt. of India. This program aims at promotion of IT tools for Indian Languages. And it also has established thirteen Resource Centres for Indian Language Technology Solutions covering all the eighteen constitutionally approved official languages (cf. [11]). These activities, including the EMILLE project [12], are focused on collecting and processing Indian text corpora and lexica. Speech processing seems to be dealt with only marginally. Since no appropriate Hindi speech databases are available at ELRA or LDC either, no speech databases of Indian languages are publicly available which could be used to develop performant speech recognizers for any Indian language.

Given this situation the LILA consortium decided to collect in the first project phase three speech databases in India which can be used to develop recognizers for mobile telephone applications:

- Hindi as first language, 3000 speakers, to be collected in 8 states with Hindi as official language;

- Hindi as second or third language, 3000 speakers, to be collected in the remaining states of India;

- English as second or third language, 3000 speakers, to be collected all over India.

In a second phase other languages like Bengali, Urdu, etc. could be collected as well.

The specification will be based on SALA-II with respect e.g. to content, format, sampling frequency, recording environment, speaker profile, etc. (cf. [13]).

4. LILA Validation

In a consortium where each partner is responsible for the production of part of the SLR (i.e. in LILA one or two languages per partner), it is important that each partner provides SLR of equal quality ('E-quality'). Only E-quality in the final LR allows a fair exchange between partners at the end of the project. Therefore, an independent validation centre will check the databases against the specifications. This strategy was already adopted in SpeechDat-II, SpeechDat-Car, SALA, Orientel and SpeeCon [2]. In all these projects, the Speech Processing Expertise Centre (SPEX) [14] from the Netherlands was chosen as the validation centre, as it will be in LILA.

The validation of a SpeechDat like database typically consists of checks on documentation, database format, design, speech files, label files, phonemic lexicon, speaker & environment distributions, and orthographic transcriptions [15].

Over the years speech database validation has evolved [16]. In earlier projects validation consisted of two parts: in the early stage of the project a pre-validation in which a mini database of the first 10 recorded speakers was checked in order to be able to correct design errors and at the end a validation of the complete database. In the more recent projects Orientel and SpeeCon, three parts have been added: firstly a prompt sheet validation in which the reading scripts, used for the recordings, are checked for design and completeness. Secondly, the phonemic lexicon is validated by a phonetic expert who is a native speaker of the language involved. Finally, a pre-release database validation is conducted, to make sure that after changes have been made in between final validation and production of the final database no repairable errors are left. For the LILA project, the two basic validation parts will be done in any case, but probably the expanded procedure will be implemented.

SPEX has quite some experience in validating speech databases in languages which are not mastered by staff members like Arabic and Hebrew from Orientel and Japanese, Thai and Mandarin from SpeeCon. A network of native language experts was developed over the years. The LILA project allows SPEX to expand both its expert network and its validation skills for a new range of Asian languages.

5. Current approach to LR production

With the very rapid growth of the market of HLT, that led to more demand of language resources, a number of initiatives have been implemented in order to fulfill the demand. Within the last decade, the "partnership" scenario has been set up to cut production costs: partners get together and formed consortia, each member of the consortia would produce one data base according to the same specifications. If the database is qualified by an external validation center, then the producing partner could exchange it against all other partners databases under an usage license. In some cases European partners managed to obtain public support from the European Commission (shared-cost projects), in other cases partners relied in private investment. With this approach each partner retained the ownership of the database it produced.

Some agreed to distribute the database to parties outside the consortium (often via ELRA) and hence recoup (some of) the production costs.

We do not mention here the resources developed internally exclusively with private funds that are unlikely to end up in any distribution catalogue

With these approaches, partners assumed that all resources are of equal value and of equal cost. Experience has proven this assumption to be inaccurate.

In order to better address such issue, LILA partners are envisaging other scenarios which aim at equalizing the production costs as well as the revenues generated by the distribution of such resources to third parties not involved in the production process. Two open issues are being debated currently:

1. the funds required for the production of such large number of resources may be based on an investment-fund-like schema to ensure that enough financial resources are available to the project.

2. the use of a "production" consortium that will be responsible for the concrete production work of the various databases within each country. This method would ensure a high level of capitalization on experiences learnt within each country to cut costs but also to improve the production process. Such consortium will rely on a number of local experienced production centers.

6 References

[1] http://www.elra.info

[2] http://www.speechdat.org

[3] http://gps-tsc.upc.es/veu/lila Provisional web page

[4] http://www.cia.gov/cia/publications/factbook/ geos/in.html

[5] Winter, Margot: India's Telecommunications Market: Implementation Priorities and Differences with China. Siemens Business Information Report ST030401, March 2003, p. 1.

[6] http://adaniel.tripod.com

[7] http://www.india4world.com/indian-language /index. html

[8] Baldridge, Jason: Reconciling Linguistic Diversity: The History and the Future of Language Policy in India. University of Toledo Honors Thesis, August 1996. http://www.ling.upenn.edu/~jason2/ papers/natlang.htm

[9] http://www.ciil.org/languages/map4.html

[10] Vikas, Om: Language Technology Development in India. Ministry of Information Technology, New Delhi, India. http://www.emille.lancs.ac.uk/lesal/omvikas.pdf

[11] http://tdil.mit.gov.in

[12] http://www.emille.lancs.ac.uk/home.htm

[13] http://www.sala2.org

[14] http://www.spex.nl

[15] H. van den Heuvel, The Art of Validation, The ELRA Newsletter, Vol. 5(4), pp. 4-6.

[16] H. van den Heuvel, D. Iskra, E. Sanders, F. de Vriend, SLR validation: current trends and developments, LREC 2004, Lisbon