

# ASR AND TRANSLATION FOR UNDER-RESOURCED LANGUAGES

L. Besacier, V-B. Le, C. Boitet, V. Berment

CLIPS/IMAG Laboratory, UJF, BP53, 38041 Grenoble cedex 9, France

{laurent.besacier, viet-bac.le, christian.boitet, vincent.berment}@imag.fr

## ABSTRACT

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages for which large resources are available or which have suddenly become of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities have been less worked on in the past years. One way of improving this "language divide" is do more research on portability of HLT for multilingual applications.

In this paper, we concentrate on speech-to-speech translation. We present here our methodology for fast development of ASR systems for under-resourced languages or, as they are called now,  $\pi$ -languages (poorly equipped). We present the resources collected for Vietnamese, and the experimental results of our first Vietnamese ASR system. The current validation of our methodology for Khmer is described next. We also discuss some issues related to machine translation and present first contributions of our laboratory in this context of " $\pi$ -languages".

## 1. INTRODUCTION

Nowadays, computers are used to write and communicate. Text processing tools, electronic dictionaries, and even more advanced systems like text-to-speech or dictation are now available for some languages. Measuring the availability of these services for a given language allows to define its "computerization level". An example of such a metric is presented in [1]: a list of services is evaluated for a given language by an expert and a mean score is calculated (marks for each service are weighted by the criticity or importance of the service). *Table 1* gives an example of this metric applied to Khmer, a language mainly spoken in Cambodia (6.2/20). The same metric evaluated for Vietnamese (*table 2*) gives 10/20. In [1], an under-resourced language (or  $\pi$ -language) is thus defined as a language which has a score below 10/20.

From these tables, we can note that for both languages, speech processing services do not exist at all. The reason is mainly that for developing such systems, a large amount of

resources is needed (text, transcribed speech corpora, phonetic dictionaries). Such resources are not available for languages like Vietnamese and Khmer. One may also face other problems like the absence of linguistic or phonetic descriptions, few standards (character coding, IPA, ...). In a perspective of speech-to-speech translation (STST), we have thus developed a methodology and tools to collect, process and model linguistic and acoustic resources in order to quickly develop STST systems for new target languages. This paper presents an overview of our activities concerning ASR and translation for under-resourced languages. *Section 2* presents our methodology, which is then evaluated (ASR only) for Vietnamese and Khmer languages in *section 3*. Finally, *section 4* concludes this work.

Services / ressources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
<b>Text processing</b>			
Basic input	10	16	160
Visualization / printing	10	14	140
Search and Replace	8	12	48
Text selection	6	12	72
Lexicographical sort	5	0	0
Spelling Correction	2	0	0
<b>Speech processing</b>			
Text-to-speech	5	0	0
Automatic Speech Recognition	5	0	0
<b>Translation</b>			
Tools for Automatic translation	8	4	32
<b>OCR</b>			
Optical Character Recognition	9	0	0
<b>Ressources</b>			
Bilingual dictionary	10	4	40
Usability dictionary	10	0	0
<b>Total</b>			540 / 1760
<b>Mean</b>			6,2 / 20

Table 1: Computerization level for the Khmer language

Services / ressources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
<b>Text processing</b>			
Basic input	10	16	160
Visualization / printing	10	16	160
Search and Replace	8	17	136
Text selection	6	17	102
Lexicographical sort	5	6	30
Spelling Correction	2	6	12
<b>Speech processing</b>			
Text-to-speech	5	0	0
Automatic Speech Recognition	5	0	0
<b>Translation</b>			
Tools for a utomatic translation	8	6	48
<b>OCR</b>			
Optical Character Re cognition	9	12	108
<b>Ressources</b>			
Bilingual dictionary	10	13	130
Usability dictionary	10	0	0
<b>Total</b>			886 / 1760
<b>Mean</b>			10 / 20

Table 2: Computerization level for the Vietnamese language

## 2. METHODOLOGY

### 2.1. Automatic Speech Recognition

#### 2.1.1. Speech and Text Resources Collection

Given the statistical nature of the methods used in ASR, the amount of resources collected is crucial for building an ASR system for a new target language.

Concerning text resources, a new methodology for fast text corpora acquisition for  $\pi$ -languages was already proposed and applied in [2]. Documents are gathered from Internet by web robots. Then, these web pages are filtered and analyzed for building a text corpus. An open source toolkit is currently proposed by CLIPS<sup>1</sup>. The amount of data collected with this approach for Vietnamese and Khmer will be presented in *section 3*.

Concerning speech resources, we have first built local collaborations (MICA/Hanoi<sup>2</sup>, ITC/Phnom-Penh<sup>3</sup>) and performed local recordings with our data collection tool EMACOP (Multimedia Environment for Acquiring and Managing Speech Corpora) [3].

#### 2.1.2. Pronunciation Dictionaries

The pronunciation dictionary is a key component of any ASR system. No such dictionary is available for any under-resourced language. To get one, we can consider the following options:

- knowledge-based approaches used to build a phonetizer - this option is time-consuming (design of a phonetizer for a new language) and necessitates a good knowledge of the target language,

- automatic approaches using a phone recognizer [4] – in that case, the pronunciation dictionary quality will highly depend on the quality of the phonemic decoder and on the phonemic coverage,

- grapheme-based approaches: this option, firstly proposed in [5] for English and German, does not use any pronunciation dictionary since the words of the vocabulary are described in terms of grapheme units.

In *section 3*, we will present results obtained with the first and third options described here.

#### 2.1.3. Acoustic modeling

Some methods of phone mapping can be used to evaluate acoustic similarities across languages. The core of these methods is the phone mapping table that describes the similarity of sounds between two different languages (a source language for which we have an acoustic model, and the target language). Both knowledge-based and data-driven methods [6] have been already proposed in our work to

manually or automatically obtain these phone mapping tables.

The phone mapping table can be obtained by using acoustic/phonetic knowledge to categorize phonetic units with similar features of the individual languages (knowledge-based method). In a knowledge-based method, we find the IPA counterpart of target phonemes among phonemes in source language. This kind of method can be used if no data at all is available in the target language but a good knowledge of the target language is needed.

By using a small amount of annotated acoustic data in the target language, the phone mapping table can be also automatically created with data-driven methods. In our approach, a confusion matrix is calculated by applying a source language phoneme recognizer on a target language speech corpus already labeled with the target language acoustic units.

After obtaining a pronunciation dictionary in terms of the symbols associated with the acoustic units of the source language, we can directly use acoustic models in source language to decode the speech of an utterance in the target language. In that case, it is interesting to use several source languages instead of one to better cover the phone inventory of the target language. Thus, this approach can be extended and used with multilingual acoustic models [7]: for instance, the phonemic coverage between French and Vietnamese is 67%, and increases to 87% between a multilingual model (7 languages<sup>4</sup>) and Vietnamese.

While both the knowledge-based and data-driven phone mappings can be used without modification of the original source language acoustic models, HMM adaptation using MLLR or MAP techniques may also be used to improve the systems using a small amount of target language adaptation data if available.

### 2.2. Machine Translation

#### 2.2.1. Translation support tools and resources building

Considering that a high or medium quality translation system is very long and complex to build and that the populations who speak a  $\pi$ -language are generally not able to undertake the design of such a system from scratch, less ambitious — although quite useful — systems have to be imagined. For example, on our website <http://www.laosoftware.com>, designed by V. Berment, a Lao<sup>5</sup> language user can get online the word for word translation in French of a short Lao text. Look-up of Lao-French and French-Lao dictionaries is also available.

Such systems are considerably simpler and cheaper than completely automated translation systems, they can be very useful if readers accept to spend a few hours to learn the

<sup>1</sup> <http://www-clips.imag.fr/geod/User/viet-bac.le/outils/>

<sup>2</sup> <http://www.mica.edu.vn>

<sup>3</sup> <http://www.itc.edu.kh/fr/>

<sup>4</sup> chinese, croatian, french, german, japanese, spanish, turkish

<sup>5</sup> The Lao language is spoken in Laos and in Thailand.

basics of the  $\pi$ -language in question, but they also need a dictionary which can have a very simple structure, but must have wide coverage. It is then a good strategy to design dictionaries for active reading systems so that they can be extended later to contain the information necessary for MT systems, and so that they can grow through contributions by the users of the “language services” (active reading and then MT, in this case).

In our example of online word for word translation, when a user finds that a word is wrongly (or not) translated, s/he can contribute to the dictionary by providing another or a new translation. Such a collaborative work has been undertaken for the Lao language by the *LaoLex* project, in which each contributor owns a personal dictionary and bi-texts (for more details, see [1]).

When a bilingual dictionary involving a  $\pi$ -language has been obtained, it can then be added to the *Papillon* multilingual dictionary<sup>6</sup>. As this latter dictionary relies on an interlingual “hub” of interlingual links (the *axies*) to interconnect the word senses (or *lexies*) of the various languages, the integration of a new language can be done easily, simply by creating new links and possibly some new axes, thereby making the words (rather, the word senses) of the new dictionary translatable into and from any “connected” dictionary of the *Papillon* database.

### 2.2.2. Text Segmentation

In one of the area we are focusing on (Cambodia, Lao, Thailand ...), most of the languages are written without spaces between the words (unsegmented writing systems). Hence, segmenting texts is a primary task for a number of language processing systems. To do that, we developed a particular segmentation technique based on the syllables regularity. Among the possible applications of this technique, are the pre-processing of the texts collected from Internet (see section 2.1.1), phonetizers (see [1] at section II.2.4.3 for Lao, and [8] or [9] for Thai), basic text to speech engines (see [9] for Thai), as well as translation support tools (see [1] at section II.3.1 for a description to Lao). Described in detail in [1] at section III.2.4, the production of such a syllabic segmenter relies on a context-free grammar description of the syllables — a form that is intuitive for the linguists who do the actual description — and on its compilation. The result is either a syntactic analyzer or an automaton, depending on the grammar compiler we use, but in both cases, C++ code is produced<sup>7</sup>. A syllabic segmenter for Khmer has been produced in order to evaluate different segmentation methods. It is a 69 states automaton that tells if a given string of bytes is a UTF-8 encoded Khmer

syllable or not. This segmenter is currently being applied to our Khmer corpus<sup>8</sup>.

A word segmenter was also built for Khmer using dictionary-based longest matching algorithm. We have tested this word segmenter on a Khmer text corpus of 520 sentences (6400 reference words). The preliminary results obtained show 4.5% of segmented word error and 24.8% of segmented sentence error. In the future, we plan to use an N-gram language model to reduce this segmentation error.

## 3. EXPERIMENTS AND RESULTS FOR VIETNAMESE AND KHMER ASR

### 3.1. ASR System

All recognition experimentations use the JANUS Speech Recognition Toolkit (JRTK) [10] developed by the ISL Laboratories. The model topology is 3 states left-to-right, 64 Gaussian mixtures. The pre-processing of the system consists in extracting a feature vector every 10 ms. The feature vector of 43 dimensions contains zero-crossing, 13 MFCC, energy and their first and second derivatives. A LDA transformation is used to reduce the feature vector size to 24. For the moment, we deal with context-independent acoustic models only. Vietnamese language is a monosyllabic and tonal language with 6 tones. Thus, in our system, the selected recognition unit is syllable instead of word. Furthermore, the acoustic phone units are modeled independently of the tones. The decision between two different syllables corresponding to a same phone sequence but to different tones is made by the language model.

### 3.2. Data collected

For Vietnamese, we collected 2.5 Go of Web data which was then filtered to get 400 Mo of usable data for language modelling (for comparison purpose, 1 year of « Le Monde » French Journal corresponds to 120 Mo). 35h of speech signal were also collected (see [11] for details). Our Vietnamese pronunciation dictionary contains 6,492 monosyllables (either phonetized with a rule-based phonetiser or described in terms of graphemes). For Khmer, we have currently 80 Mo of Web data (~6000 pages) and 3 hours of speech signal.

### 3.3. ASR Performance

*Table 3* shows the performance obtained with Vietnamese adaptation data (1h or 2h to adapt the source acoustic models) aligned with acoustic models obtained with either knowledge-based (IPA) or data-driven phone mapping methods. The source acoustic models are either French or

<sup>6</sup> [www.papillon-dictionary.org](http://www.papillon-dictionary.org)

<sup>7</sup> Part of this effort has been done together with Claude Del Vigna, CNRS, Paris (Centre d'Analyse et de Mathématique Sociales)

<sup>8</sup> Note that this Khmer syllabic segmenter is already used in the MSWord add-in called GMSWord in order to provide a convenient mouse selection for this unsegmented writing.

multilingual (the multilingual acoustic model was obtained with the courtesy of Tanja Schultz from CMU).

From these results, it is interesting to note the potential of the data-driven method for phone mapping which leads to approximately the same results as the knowledge-based (IPA) one. Obviously, we also see that the phonemic coverage of the source acoustic models is important, which is illustrated by the better results obtained with the multilingual source models than with the French source models. Finally, we have also trained a grapheme-based system with 3 hours of Vietnamese speech, and it gives 57% syllable accuracy on the same test database, which is promising since no phonetic knowledge at all is used in such an approach.

Concerning Khmer, our first ASR system (grapheme-based, acoustic models trained on 2.5h of speech, 16,000 words vocabulary) obtains 73.6% word accuracy on 200 test utterances (dialog). This first Khmer ASR system was obtained from scratch in 5 months.

Firstly, pronunciation dictionary for Khmer grapheme-based recognition was built by simply splitting a word into its graphemes. However, we know that the Khmer character is script. So, a character romanization procedure was needed to convert Khmer characters to a computer-readable form in the grapheme-based pronunciation dictionary. For this, the following options can be used: use of the Unicode Character Name<sup>9</sup>, use of the Khmer Romanization Table<sup>10</sup>,... Secondly, after having collected some Khmer speech data, with no graphemic/acoustic alignment available, an automatic labeling process was firstly employed to map the acoustic data with the acoustic models. In our work, we used a word boundary detector (not described here) to segment an utterance into words. Then speech parts labeled as words were linearly segmented into graphemes. This segmentation method (using a word segmenter) improves the performance compared to fully linear segmentation of a speech utterance in graphemes as done in [5]. Finally, for context dependent graphemic modeling, we used the singleton method proposed in [5] where each question consists in one single grapheme, to generate the model clustering questions.

Source system	Models	Adapt 1h	Adapt 2h
		SA	SA
French	IPA	60.4	63.6
	Data-driven	61.6	63.8
GlobalPhone Multilingual (7 languages)	IPA	64.6	66.3
	Data-driven	63.8	65.3

Table 3: ASR performance for Vietnamese (% syllable accuracy, dialog test corpus)

<sup>9</sup> <http://www.unicode.org/charts/PDF/U1780.pdf>

<sup>10</sup> [http://www.eki.ee/wgrs/rom1\\_km.pdf](http://www.eki.ee/wgrs/rom1_km.pdf)

## 4. CONCLUSION

In this paper, we have presented our methodology for fast development of ASR systems for under-resourced languages. The resources collected for Vietnamese, and the experimental results of our first Vietnamese ASR system were presented. The current validation of our methodology for Khmer was also described. Our first Khmer ASR system, designed in five months, obtained 73.6% word accuracy. Moreover, some issues related to machine translation and first contributions of our laboratory in this context of under-resourced languages were also discussed here.

## 5. REFERENCES

- [1] V. Berment "Méthodes pour informatiser des langues et des groupes de langues peu dotées" PhD Thesis, J. Fourier University – Grenoble I, May 2004.
- [2] V.B. Le, B. Bigi, L. Besacier, E. Castelli, "Using the Web for fast language model construction in minority languages", Eurospeech 2003, pp. 3117-3120, Geneva, September 2003.
- [3] D. Vaufraydaz, M. Akbar, J. Caelen, J.F. Serignat, "EMACOP: Environnement Multimédia pour l'Acquisition et la gestion de COpus Parole", JEP'98 (Speech Processing Workshop), Martigny, Switzerland, pp. 175-178, June 1998.
- [4] R. Bayeh, S.Lin, G.Chollet, C.Mokbel, "Towards multilingual speech recognition using data driven source/target acoustical units association", ICASSP 2004, vol. I, pp. 521-524, Montreal, Canada, May 2004.
- [5] M. Killer, S. Stüker, and T. Schultz, "Grapheme based Speech Recognition", Eurospeech 2003, Geneva, Switzerland, September 2003.
- [6] V-B. Le, L. Besacier. "First steps in fast acoustic modeling for a new target language. Application to Vietnamese" Proceedings IEEE ICASSP 2005. Philadelphia, USA. April 2005.
- [7] T. Schultz, A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition", Speech Communication, vol. 35, no. 1-2, pp. 31-51, 2001.
- [8] P. Mittrapiyanuruk, C. Hansakunbuntheung, V. Tesprasit & V. Sornlertlamvanich, "Issues in Thai text-to-speech synthesis: the NECTEC approach", Proceedings of NECTEC Annual Conference, Bangkok Thailand. 2000.
- [9] T. Karoonboonyanan, V. Sornlertlamvanich & S. Meknavin, "A Thai Soundex system for spelling correction". Proceeding of the National Language Processing Pacific Rim Symposium, pp. 633-636, Phuket, Thailand, ISBN 974-89570-9-8. 1997.
- [10] M.Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine", ICASSP'97, Munich, 1997.
- [11] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, J-F. Serignat, "Spoken and written language resources for Vietnamese", LREC 2004, Lisbon, May 2004.