SPOKEN AND WRITTEN LANGUAGE RESOURCES FOR VIETNAMESE

Viet-Bac Le*, Do-Dat Tran*, **, Eric Castelli, Laurent Besacier*, Jean-François Serignat***

* CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE

** International Research Center MICA
1 Dai Co Viet, Hanoi, VIETNAM

Email: (Viet-Bac.Le, Do-Dat.Tran, Eric.Castelli, Laurent.Besacier, Jean-Francois.Serignat)@imag.fr

ABSTRACT

This paper presents an overview of our activities for spoken and written language resources for Vietnamese implemented at CLIPS-IMAG Laboratory and International Research Center MICA. A new methodology for fast text corpora acquisition for minority languages which has been applied to Vietnamese is proposed. The first results of a process of building a large Vietnamese speech database (VNSpeechCorpus) and a phonetic dictionary, which is used for automatic alignment process, are also presented.

1. INTRODUCTION

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 which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this “linguistic divide” is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years, for instance in the SALTMIL[1] (Speech and Language Technology for Minority Languages) group. However, in SALTMIL, “minority language” mostly means “language spoken by a minority of people”. We rather focus, in our work, on languages which have a “minority of resources usable in HLT”. These languages are mostly from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

- Portability of acoustic models: this can be achieved, for example by using tools for performing a fast collection of speech signals (Vaufreydaz et al., 2000) or by using Language Adaptive Acoustic Modeling (Schultz & Waibel, 2001).
- Language modeling for new languages: we propose to use web-based techniques which have already shown ability to collect large amount of text corpora. For languages in which no usable text corpora exist, this is moreover the only viable approach to collect text data (Le et al., 2003).
- Dictionaries: collaborative approaches like in (Berment, 2002) could be also proposed for ASR.

This paper presents an overview of our activities for spoken and written language resources for Vietnamese. Firstly, a new methodology for fast text corpora acquisition for minority languages is proposed. With more than 800MB of text size, our Vietnamese text corpus will be used for many different applications in language processing domain: language modeling, spoken corpora definition, information retrieval...

Secondly, we describe the characteristics of a large Vietnamese language speech database (called VNSpeechCorpus) which will contain about 100 hours recorded in both quiet and office environment from 50 native speakers. VNSpeechCorpus will be available on CD-ROM and could be distributed by the ELRA-ELDA association. Automatic alignments could be provided too.

Finally, a phonetic dictionary was obtained by using our *VNPhoneAnalyzer*. It is based on the phone concatenation of the initial part, final part and tone of a syllable. The symbolic representation for the various sounds and a description of articulatory features is provided by the International Phonetic Alphabet – IPA (IPA, 1999).

2. TEXT CORPORA ACQUISITION

In this section, we will describe some techniques for fast text corpora acquisition and evaluation. First, by using a web-robot (or web-spider), we can collect and store web pages in the given language. And then, these web pages were filtered and analyzed for building a text corpus. Finally, a language model was estimated from this text corpus.

2.1. Web pages collection and data preparation

Documents were gathered from Internet by some web robots (among them, one was developed in our lab[2]). From some starting points on the Web, the robots can reach and find all the text documents and web pages which have a direct or indirect link with these starting points. However we must manage the Web sites (Internet domain names) accessed by the robots because we want to collect the pages and the documents in a given domain and in a given language only.

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL

[2] http://slmg-index.imag.fr

Some filtering techniques are needed to construct the text corpus from HTML pages. Firstly, we also noticed that the Vietnamese Web resources contain some redundant information (menus, references, advertisements, announcements ...) which is repeated in different pages. This is due to the day by day collecting of daily news which may have a direct influence on the quality of the text corpus and also on the performance of the language modeling. By filtering the redundant information contained in the web pages collected from the same site, the corpus size is reduced by 54% in our experiments. On the other hand, the corpus perplexity evaluation value is significantly improved by 26%.

Secondly, the text parts from the rest of these HTML pages were extracted and some document separators were inserted. The tokens <s> and </s> signal respectively the begin and the end of a sentence. Web texts contain also a variety of "non-standard" token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL's and e-mail addresses... Normalizing or rewriting such text using ordinary words is a first important issue.

Thirdly, because the collected documents were encoded in many encoding system (TCVN3, VNI, UCS, UTF-8 …), we must choose a unique character set for encoding all the documents. Universal Character Set (UCS) which is a part of Unicode international standard[1] provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. We have chosen Unicode standard for encoding all characters of our corpora and we constructed a tool to convert a character in several character sets to the Unicode (UTF-8 encoding system).

Finally, there are many different solutions to extract the relevant sentences from a text corpus. Classically, we can keep all the sentences exclusively made with words of the task-specific vocabulary preliminary defined.

2.2. Experiments and evaluation

Text corpus cannot be collected easily in the minority languages for some reasons:

- There are less websites than in the majority languages.
- The transmission rate is often very low.

Consequently, we can not crawl all of the websites but we must focus on some which have more pages and higher debit than the others. So, a non negligible time was used to find out the websites to collect.

There are about 2500 Vietnamese websites in Vietnam which publish: daily news, information, entertainment, e-commerce, forum... The daily news web pages introduced a constraint in the data collection, since we had to regularly access the same sites to get an acceptable amount of data. This is the major difference with web data collection for a majority language like French or English where there are enough web pages that can be collected at a given time.

The Vietnamese data collection is composed of more than 2.5 GB of web pages. After data preparation, the text corpus is made of 868 MB, i.e. 10,020,267 sentences.

We construct a Vietnamese language model for estimating this text corpus. In these experiments, we tried 4 different solutions to filter the text corpora: *all-sentences* (without sentence filtering)*; block-, sentence-based* and *hybrid* (take blocks and sentences containing only in-vocabulary words). In all cases, we selected sentences without size restriction (for more detail, see Le et al., 2003). The perplexity measures for Vietnamese are given and compared to a same size as French text corpus in table 1.

	VN: Web original filter		**VN: Web** redun. filter		**FR: Web**	
Expe.	**Size** (MB)	**PPL**	**Size** (MB)	**PPL**	**Size** (MB)	**PPL**
all	868	260	402	201	686	539
block	667	359	357	282	366	637
sent.	370	**252**	226	**195**	156	580
hybrid	729	259	373	199	411	**509**

Table 1: Perplexity of the language models

Table 1 also shows that our Web-based methods can be successfully applied to majority and minority languages like French and Vietnamese even if the correspondence between perplexities of Vietnamese and French language models is not very significant here because each language have a particular characteristic.

3. VIETNAMESE LANGUAGE SPEECH DATABASE (VNSPEECHCORPUS)

3.1. Text Corpus

Two phases of collecting text data were implemented in our project. In the first phase, data is collected by some experts in order to ensure the desired requirements. And in the second phase, data is extracted automatically with one desired distribution of acoustic units from the web corpora obtained in part 2.

Beside database of phonemes, digits, application words, other data including sentences and paragraphs were collected from different resources such as stories, books, and web documents... The selected data covers different fields and contains many dialogs and short paragraphs. This initial data then was manipulated and was divided into smaller paragraphs and conversations (about 4-6 lines/ paragraph or conversation) that help speaker to utter or read easily.

3.2. Corpus Organization

The VNSpeechCorpus contains 5 different kinds of data:

- Phonemes.
- Tones.
- Digits and string of digits.
- Application words.
- Sentences and paragraphs.

The phonemes are read by all speakers. The vowels can be read independently except two vowels *ă* /ă/ and *â* /ɤ̆/, because their sounds are only represented completely in words in which they appear, such as *ngắn (short), tân*

[1] http://unicode.org

(new)… The consonants are combined with vowel *ơ* /ɤ/ and falling tone for pronouncing.

Vietnamese is a tonal language with 6 tones (Doan, 1997), for example: *ba* (three)*, bà* (grandmother)*, bá* (king)*, bả* (bane)*, bã* (waste)*, bạ* (any)*.* The speakers are asked to read the words with different tones. These words have almost the same initial and final part, but they have different tones.

The digit corpus consists of isolated digits, connected digits and natural numbers. In Vietnamese digital system, most of the digits and numbers are read or uttered with the unique sound. However, there are some synonyms, especially, the numbers ended by digit 4 and 5; they could be read in different ways. In order to cover all cases, the corpus consists of all of the variants (synonyms) of theses digits and numbers.

A set of more than 50 application words is defined in the corpus. Each word corresponds to an action which is useful in several applications such as telephone services, measurement, human-machine interface …

After selecting and processing selected paragraphs and conversations, our sentences corpus is divided into two parts, a common part and a private part. The common part contains 33 conversations and 37 paragraphs. They were read by all speakers. The private part includes about 2,000 short paragraphs, each speaker was asked to read 40 paragraphs.

3.3. Distribution of Acoustic units

To evaluate our corpus, we used several modules to analyze the distributions of acoustic units including mono-words, base syllables, Initial-Final parts, phonemes and tones in the corpus and compare their distribution with the distribution obtained on a larger corpus which was collected in section 1 (Web corpus). We consider the acoustic units distribution obtained on this large web corpus, as the reference distribution of what can occur on Vietnamese language. Vietnamese is a monosyllabic and tonal language. Besides analyzing the distributions of phonemes (mono-phone, diphone and triphone) like other language (figure 1), we carried out an analysis of the distributions of tones (figure 2), initial and final parts.

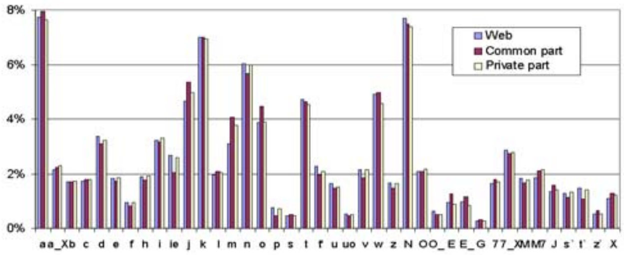

Figure 1: Distribution of mono-phones in common part, private part and Web corpora

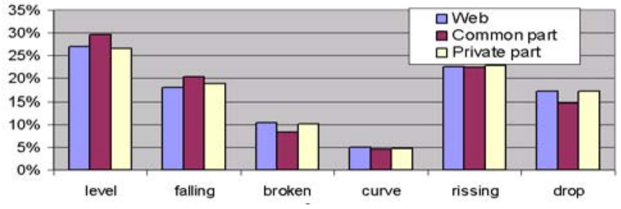

Figure 2: Distribution of six tones in common part, private part and Web corpora

In addition, we calculated the correlation coefficients between the distributions of the common part and the private part with the web reference corpora. Table 2 shows that the correlation coefficients are near to 1. So we can conclude that our corpus is acceptable and correctly balanced in terms of acoustic units and tones.

Acoustic Units	Correlation Coefficients	
	Private part	Common part
Mono-phone	0.9962	0.9885
Di-phone	0.9821	0.9458
Tri-phone	0.9811	0.9420
Tone	0.9984	0.9885
Initial Part	0.9904	0.9670
Final Part	0.9910	0.9706

Table 2: Correlation coefficients of acoustic units between common part, private part and Web data

3.4. Speaker selection and recording

Our speakers are the employees of the International Research Center MICA, teachers and students of Hanoi University of Technology and their friends. They are from four big cities and provinces, Hanoi, Nghe An, Ha Tinh, HCM city, which represent 3 major dialect regions: the South, the North, and the Middle. The age of the speakers ranges from 15 to 45 years old, among the 50 speakers, 25 are females and 25 are males.

For the acquisition, and managing of speech signals during recording, we use the EMACOP system, developed at CLIPS. EMACOP is a Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 9x and Windows NT. EMACOP meets SAM specifications on input and output (Vaufreydaz, 2000).

In our project, recordings will take place with a SENNHEISER HMD 410-6 head microphone and a microphone pre-amplifier Soundcraft Spirit Folio FX8. The sampling frequency is 16 kHz.

At this time, 15 speakers have been recorded in the studio of the International Research Center MICA, Hanoi University of Technology, Vietnam. Each speaker has been asked for recording about 60 minutes, which includes 45 common minutes of phonemes, tones, digits and strings of digits, application words and common sentences and paragraphs corpus, and 15 private minutes of about 40 short paragraphs.

4. DESIGNING A VOCABULARY AND A PHONETIC DICTIONARY

4.1. Vocabulary

In order to build the language model for ASR, it is necessary to have a vocabulary. This list of words will be also useful to filter out the text documents before training a language model. We can use a bilingual or a multilingual dictionary for generating this vocabulary.

In fact, there are many methods to construct a bi-lingual or multi-lingual dictionary. In the context of the Papillon[1]

[1] http://bushido.imag.fr/papillon

project, the construction of a lexical base for a new language may take several different ways depending on where the author has to start. The Papillon project aims at creating a multilingual lexical database covering among others English, French, Japanese, Malay, Lao, Thai and Vietnamese.

From this Papillon project, we got a dictionary for Vietnamese language (French-Vietnamese and Vietnamese-French). Then, we filtered this dictionary to have a list of more than 40,000 unique words in Vietnamese: compound words, borrowed words and isolated words. By taking only the most frequent words, we can discount this size of vocabulary to 20,000 words. These were the highest frequency words which occur in the documents of our Web text corpus.

4.2. Phonetic dictionary

Phonetic dictionary or pronunciation dictionary is a key part for acoustic modeling in ASR. However, there is not any official pronunciation dictionary in Vietnam which satisfies our requirement. Therefore, we decide to construct a dictionary which is not used only for our works but also for other requirements in spoken language processing.

As referred above, Vietnamese language is a mono-syllabic and tonal language with 6 tones. A syllable in full structure (a tonal syllable or an isolated word) has five parts: *initial sound* (consonant), *medial sound* (semi-vowel), *nucleus sound* (vowel or diphthong), *final sound* (consonant or semi-vowel) and tone (see figure 3). Except the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part.

<table>
<tr><td colspan="5">Tonal syllable (6,492)</td></tr>
<tr><td colspan="4">Base syllable (2,376)</td><td rowspan="3">Tone
(6)</td></tr>
<tr><td rowspan="2">INITIAL
(22)</td><td colspan="3">FINAL (155)</td></tr>
<tr><td>Medial(1)</td><td>Nucleus(16)</td><td>Ending(8)</td></tr>
</table>

Figure 3: The phonological hierarchy of Vietnamese syllables with the total number of each phonetic unit

Since Vietnamese is a monosyllabic language (each syllable is one isolated word), we decide to extract a vocabulary of only about 6,500 isolated-words from 40,000 words vocabulary obtained in the previous section for building a pronunciation dictionary.

From this isolated-words vocabulary, we have firstly extracted all **22 initial parts**, **155 final parts** and **6 tones**. And then the phonological transcripts of these parts (IPA Symbols) were manually built. They were called *IPA Reference Table for Sub-word Units (**IPATU**).*

Secondly, we have constructed an automatic syllable-based phonological analyzer (called ***VNPhoneAnalyzer***). This analyzer uses a phone concatenation algorithm described below.

Phone concatenation algorithm:

1. Separate a syllable into initial part, final part and tone.
2. Transcribe the initial part, final part and tone by their correspondent phonological representations looked up in the IPATU Table.
3. Concatenate all of these phonological transcripts into a complete phonological transcript for a syllable.

The *VNPhoneAnalyzer* could output many transcription forms: IPA-Unicode Symbolic, IPA number, SAMPA[1]…

Finally, a pronunciation dictionary for Vietnamese was also built by applying the *VNPhoneAnalyzer* for the isolated-word vocabulary and tested under the helping of Linguistic Institute of Vietnam.

5. CONCLUSIONS AND PERSPECTIVES

We have introduced the process of building the written and spoken resources for Vietnamese language at CLIPS-IMAG Laboratory (France) and International Research Center MICA (Vietnam). VNSpeechCorpus has been carried out in order to provide a great quantity of usable data for training and testing ASR systems. It could also be used for speech synthesis based on the acoustic units that are smaller than syllable in Vietnamese language. From the results, we can conclude that, with a suitable adaptation we can apply the methods which were developed for majority languages to minority languages. In our project, the Web-based methods built in CLIPS-IMAG which is useful for French and English is adapted for Vietnamese language. At present, the automatic alignment and recording process are being done for VNSpeechCorpus. In the future, VNSpeechCorpus will be available on CD-ROM. Furthermore, our methods will be used for other minority languages such as Lao, or Cambodian... for evaluating more precisely the efficacy of our methodology.

REFERENCES

Berment, V. (2002). Several technical issues for building new lexical bases. In Workshop Papillon, Tokyo, Japan, 2002.

Doan, T.T. (1997). Ngữ âm Tiếng Việt (Vietnamese Phonetics). Vietnam National University Publishing House, 1997.

IPA (1999). Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, 1999.

Le, V.B., Bigi, B., Besacier, L., Castelli, E. (2003). Using the Web for fast language model construction in minority languages. In Eurospeech 2003, pp. 3117-3120, Geneva, 1-4 Sept, 2003.

Schultz, T., Waibel, A. (2001). Language independent and language adaptive acoustic modeling for speech recognition. In Speech Communication, vol. 35, no. 1-2, pp. 31–51, 2001.

Vaufreydaz, D., Bergamini, C., Serignat, J.F., Besacier, L., Akbar, M. (2000). A new methodology for speech corpora definition from internet documents. In LREC, vol. I, pp. 423–426, Athens, Greece, 2000.

[1] http://www.phon.ucl.ac.uk/home/sampa