Finite-State Description of Vietnamese Reduplication

Le Hong Phuong
LORIA, France
lehong@loria.fr

Nguyen Thi Minh Huyen
Hanoi Univ. of Science, Vietnam
huyenntm@vnu.edu.vn

Azim Roussanaly
LORIA, France
azim@loria.fr

Abstract

We present for the first time a computational model for the reduplication of the Vietnamese language. Reduplication is a popular phenomenon of Vietnamese in which reduplicative words are created by the combination of multiple syllables whose phonics are similar. We first give a systematical study of Vietnamese reduplicative words, bringing into focus clear principles for the formation of a large class of bi-syllabic reduplicative words. We then make use of optimal finite-state devices, in particular minimal sequential string-to string transducers to build a computational model for very efficient recognition and production of those words. Finally, several nice applications of this computational model are discussed.

1 Introduction

Finite-state technology has been applied successfully for describing the morphological processes of many natural languages since the pioneering works of (Kaplan and Kay, 1994; Koskenniemi, 1983). It is shown that while finite-state approaches to most natural languages have generally been very successful, they are less suitable for non-concatenative phenomena found in some languages, for example the non-concatenative word formation processes in Semitic languages (Cohen-Sygal and Wintner, 2006). A popular non-concatenative process is reduplication – the process in which a morpheme or part of it is duplicated.

Reduplication is a common linguistic phenomenon in many Asian languages, for example Japanese, Mandarin Chinese, Cantonese, Thai, Malay, Indonesian, Chamorro, Hebrew, Bangla, and especially Vietnamese.

We are concerned with the reduplication of Vietnamese. It is noted that Vietnamese is a monosyllabic language and its word forms never change, contrary to occidental languages that make use of morphological variations. Consequently, reduplication is one popular and important word formation method which is extensively used to enrich the lexicon. This follows that the Vietnamese lexicon consists of a large number of reduplicative words.

This paper presents for the first time a computational model for recognition and production of a large class of Vietnamese reduplicative words. We show that Vietnamese reduplication can be simulated efficiently by finite-state devices. We first introduce the Vietnamese lexicon and the structure of Vietnamese syllables. We next give a complete study about the reduplication phenomenon of Vietnamese language, bringing into focus formation principles of reduplicative words. We then propose optimal finite-state sequential transducers recognizing and producing a substantial class of these words. Finally, we present several nice applications of this computational model before concluding and discussing the future work.

2 Vietnamese Lexicon

In this section, we first present some general characteristics of the Vietnamese language. We then give some statistics of the Vietnamese lexicon and introduce the structure of Vietnamese syllables.

The following basic characteristics of Vietnamese are adopted from (Đoàn, 2003; Đoàn et al. , 2003; Hữu et al. , 1998; Nguyễn et al. , 2006).

2.1 Language Type

Vietnamese is classified in the Viet-Muong group of the Mon-Khmer branch, that belongs to the Austro-Asiatic language family. Vietnamese is also known to have a similarity with languages in the Tai family. The Vietnamese vocabulary features a large amount of Sino-Vietnamese words.

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 63–69,
Suntec, Singapore, 6-7 August 2009.

Moreover, by being in contact with the French language, Vietnamese was enriched not only in vocabulary but also in syntax by the calque of French grammar.

Vietnamese is an isolating language, which is characterized by the following properties:

- it is a monosyllabic language;
- its word forms never change, contrary to occidental languages that make use of morphological variations (plural form, conjugation, *etc.*);
- hence, all grammatical relations are manifested by word order and function words.

2.2 Vocabulary

Vietnamese has a special unit called "*tiếng*" that corresponds at the same time to a syllable with respect to phonology, a morpheme with respect to morpho-syntax, and a word with respect to sentence constituent creation. For convenience, we call these "*tiếng*" syllables. The Vietnamese vocabulary contains

- simple words, which are monosyllabic;
- reduplicative words composed by phonetic reduplication;
- compound words composed by semantic coordination and by semantic subodination;
- complex words phonetically transcribed from foreign languages.

The Vietnamese lexicon edited recently by the Vietnam Lexicography Center (Vietlex[1]) contains $40,181$ words and idioms, which are widely used in contemporary spoken language, newspapers and literature. These words are made up of $7,729$ syllables. Table 1 shows some interesting statistics of the word length measured in syllables. $6,303$ syllables (about 81.55% of syllables) are words by themselves. Two-syllable words are the most frequent, consisting of nearly 71% of the vocabulary.

2.3 Syllables

In this paragraph, we introduce phonetic attributes of Vietnamese syllables. In addition of the monosyllabic characteristic, Vietnamese is a tonal language in that each syllable has a certain pitch characteristic. The meaning of a syllable varies with its tone. This phonetic mechanism can also be found in other languages such that Chinese or Thai.

[1] *http://www.vietlex.com/*

Length	#	%
1	$6,303$	15.69
2	$28,416$	70.72
3	$2,259$	5.62
4	$2,784$	6.93
≥ 5	419	1.04
Total	$40,181$	100

Table 1: Length of words measured in syllables

No.	Tones	Notation
1.	low falling	à
2.	creaky rising	ã
3.	creaky falling	ạ
4.	mid level	a
5.	dipping	ả
6.	high rising	á

Table 2: Vietnamese tones

There are six tones in Vietnamese as specified in Table 2. The letter *a* denotes any non-accent syllable. These six tones can be roughly classified into two groups corresponding to low and high pitches in pronunciation. The first half of the table contains three low tones and the second half contains three high tones. In addition, the difference in the tone of two syllables are distinguished by flat property of tones. The 1st and 4th tones in Table 2 are flat (*bằng*), the other tones are non-flat (*trắc*).

The structure of a Vietnamese syllable is given in Table 3. Each syllable can be divided into three parts: onset, rhyme and tone. The onset is usually a consonant, however it may be empty. The rhyme contains a vowel (nucleus) with or without glide /w/, and an optional consonant (coda). It is noticed that the initial consonant of a syllable does not carry information of the tone, the Vietnamese tone has an effect only on the rhyme part of the syllable (Tran et al., 2006). This result reinforces the fact that a tone is always marked by the nucleus composant of the rhyme which is a vowel. Readers who are interested in detail the phonetic composition of Vietnamese syllables may refer to (Tran et al., 2006; Vu et al., 2005).

3 Reduplication in Vietnamese

Reduplication is one of the methods for creating multi-syllable words in Vietnamese. A reduplica-

Tone			
Onset	Rhyme		
	Glide	Nucleus	Coda

Table 3: Phonetic structure of Vietnamese syllables

tive word is characterized by a phenomenon called phonetic interchange, in which one or several phonetic elements of a syllable are repeated following a certain number of specific rules.

From the point of view of sense, the reduplication in Vietnamese usually indicates a diminutive of adjectives, which can also be found in Hebrew, or a pluralization in Malay, in Thai and in Indonesian, or an intensivity as the use of partial reduplication in Japanese, Thai, Cantonese and Chamorro (an Austronesian language spoken on Guam and the Northern Mariana Islands). In this aspect, Vietnamese reduplication serves similar functions as those of reduplication in several Asian languages, as reported in an investigation of Asian language reduplication within the NEDO project (Tokunaga et al. , 2008a; Tokunaga et al. , 2008b).

The Vietnamese reduplication creates an expressional sense connecting closely to the phonetic material of Vietnamese, a language of rich melody. Consequently, there are many Vietnamese reduplicative words which are difficult to interpret to foreigners, though in general, native Vietnamese speakers always use and understand them correctly (Diệp, 1999).

Vietnamese reduplicative words can be classified into three classes basing on the number of syllables they contain: two-syllable (or bi-syllabic) reduplicative words, three-syllable (or tri-syllabic) reduplicative words and four-syllable reduplicative words. The bi-syllabic class is the most important class because of two reasons: (1) bi-syllabic reduplicative words make up more than 98% amount of reduplicative words, that is, almost reduplicative words has two syllables; and (2) bi-syllabic reduplicative words embody principle characteristics of the reduplication phenomenon in both phone aspect and sense formation aspect. For these reasons, in this paper, we address only bi-syllabic reduplicative words and call them reduplicative words for short, if there is no confusion.

As presented in the previous section, a syllable has a strict structure containing three parts: the onset, the rhyme and the tone. Basing on the phonetic interchange of a syllable, we distinguish two types of reduplication:

- full reduplication, where the whole syllable is repeated;
- partial reduplication, where either the onset is repeated or the rhyme and the tone are repeated.

In this work, we constraint ourselves by focusing only on the construction of an efficient computational model applied for reduplicative words which have clear and well-defined formation principles. These words can be classified into three types investigated in detail in the following subsections. In given examples, the base syllables (or root syllable, or root for short) are the ones which are underlined. The reduplication that has undefined or incomplete formation rules will be tackled in future works.

3.1 Full Reduplication

In this type of reduplication, the root is identically repeated; there is only a slight difference on stress in pronunciation. For example, *hao hao* (a little similar), *lăm lăm* (intentional), *đùng đùng* (accidentally dertermined), *lừ lừ* (silently). In the Vietnamese lexicon there are 274 reduplicative words of this type.

In principle, there appears to be many reduplicative words of this type whose their roots may be whatever syllables bearing whatever tone, for instance *đỏ đỏ*, *hớ hớ*, *sững sững*, *chậm chậm*. However, in consequence of the difference of stress between the root and the reduplicant, the tone of the reduplicant is changed in order to be in harmony with the root, for the sake of more readability and audibility ("easier to read, easier to hear"). This consequence leads to the formation of reduplicative words of the second type which we call reduplication with tone according.

3.2 Reduplication with Tone According

As presented above, the difference between tone of the root and the reduplicant is a consequence of the difference between their stress which is expressed by their tones. This creates reduplicative words of the second type; for example, *đo đỏ* (reddish), *hơ hớ* (in the bloom of youth), *sừng sững* (statly, high and majestic), *chầm chậm* (rather slow). The tone properties (low or high pitch, flat or non-flat) are now put into use.

Reduplicant	Root	#
a	*ả*	72
a	*á*	128
à	*ã*	27
à	*ạ*	80
	Sum	307

Table 4: Statistic of the second type reduplication

The prosodic influence is responsible for the creation of the reduplicant from its root. As a result, the combination of tones between two syllables is realized in the following principle: *non-flat tones of the roots are matched against a corresponding flat tones of their reduplicants*. That is, the non-flat root has to select for it the flat reduplicant belonging to the same pitch, *i.e.*, in the same row. In this type of reduplicative words, the root is stressed in pronunciation.

A detailed statistic about these reduplicative words with respect to the combination of tones is given in Table 4. There are 307 reduplicative words of the second type.

3.3 Reduplication with Final Consonant According

In this type of reduplication, there is not only the difference between tones of the root and the reduplicant but also the difference between their final consonants (hence their penultimates). Some examples of this type of reduplication which we call the third reduplication type are:

- *cầm cập* (clatter, shiver), *lôm lốp* (pop pop), *xăm xắp* (a little full), *thiêm thiếp* (fall asleep), *nơm nớp* (be in a state of suspense)
- *giôn giốt* (sourish), *ngùn ngụt* (burn violently), *phơn phớt* (light red), *hun hút* (profound), *san sát* (be very close to, adjoining)
- *vằng vặc* (very clear), *nhưng nhức* (a little ache), *rừng rực* (brightly), *phăng phắc* (very silent), *chênh chếch* (a little oblique), *anh ách* (feeling bloated).

The practical observation shows that the modification of final consonant from the root to the duplicate also has a clear rule: *the noisy phone of the root is transformed to a nasal phone of the reduplicant* as shown in Table 5.

Example	At root	At reduplicant
	Noisy phone	Nasal phone
ăm ắp	*-p*	*-m*
phơn phớt	*-t*	*-n*
vằng vặc	*-c*	*-ng*
anh ách	*-ch*	*-nh*

Table 5: Transformation rules of final consonants

Root	Reduplicant	#
-p	*-m*	52
-t	*-n*	96
-c	*-ng*	56
-ch	*-nh*	28
	Sum	232

Table 6: Statistic of the third type reduplication

The transformation of final consonant occurs only with the roots having as final consonant *p*, *t*, or *c*. The principle of tone combination is the same as that of the second reduplication type.

A detailed statistic about these reduplicative words is given in the Table 6. There are 232 reduplicative words of the third type.

Briefly, the total number of reduplicative words of all the three types of reduplication is 813, making up about $813/28,416 \approx 2.86\%$ of the number of two-syllable words.

4 Implementation

We report in this section the construction of a computational model for recognition and production of the three types of reduplication presented in the previous section. We have implemented finite-state sequential transducers (FSTs) which are able to recognize and produce corresponding types of reduplicative words. These devices operate on the same input and output alphabets, say Σ, containing all Vietnamese characters.

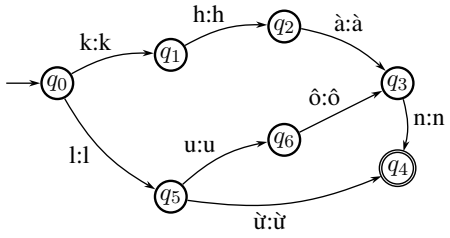
FSTs are formal devices for encoding regular relations. A regular relation is a mapping between two regular languages. In our cases, these languages are sets of Vietnamese root and reduplicant syllables.

We adapted nice and efficient algorithms developed by (Daciuk et al., 2000) to incrementally construct minimal transducers from a source of data. These algorithms are originally designed to build optimal deterministic finite-state automata on-the-fly but they can also be used to construct optimal

sequential transducers. We could consider simply that the alphabet of the automata would be $\Sigma \times \Sigma$; output strings of Σ^* are associated with the final states of the lexicon and they are only outputed once corresponding valid inputs from Σ are recognized. Interested readers are invited to refer to (Daciuk et al., 2000) for further detail of the algorithms for building optimal automata on-the-fly.

4.1 First Type Transducer

In the first type reduplication, the root and the reduplicant is completely identical in writing; they are only distinguished by a stress in pronunciation. We can simply construct a deterministic finite-state transducer (FST) f_1 that produces reduplicants from their roots in which the output string labeled on each arc is the same as its input character; that is $f_1(x) = x$ where x is a syllable in the first type duplication. As an illustration, the following minimal FST recognizes and generates three first type reduplicative words *luôn luôn* (always), *lừ lừ* (silently), *khàn khàn* (raucous).

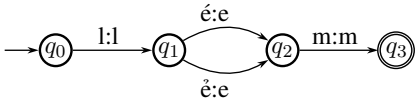


The minimal FST f_1 recognizing all 274 reduplicative words of the first type consists of 90 states in which 16 states are final ones. It has 330 transitions, the maximum number of outtransitions from a state is 28.

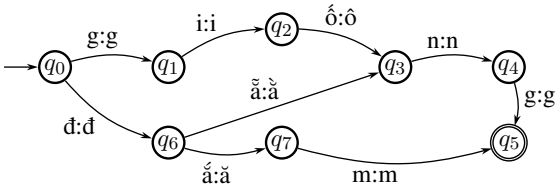
4.2 Second Type Transducer

In the second type reduplication, the root has an non-flat tone while its reduplicant has the corresponding flat tone. A root determines for it the unique reduplicant. Hence we can construct a sequential FST f_2 which is able to generate reduplicants from roots.

For instance, consider two reduplicative words of the second type *lem lém* (glib) and *lem lẻm* (voluble). They can be recognized by the minimal sequential FST f_2 such that $f_2(lém) = lem$ and $f_2(lẻm) = lem$ as depicted in the following figure:



Similarly, the minimal FST f_2 which generates three reduplicative words *giông giống* (a little similar), *đằng đẵng* (interminable) and *đăm đắm* (fixedly) is as follows:

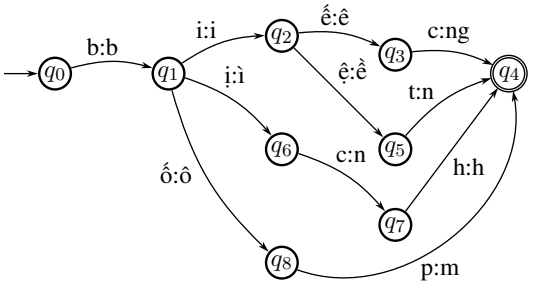


The minimal FST f_2 recognizing all 307 reduplicative words of the second type consists of 93 states in which 11 states are final ones. It has 371 transitions, the maximum number of outtransitions from a state is 22.

4.3 Third Type Transducer

The roots and reduplicants in the third type reduplication are not only combined by principles of flat and non-flat tones, they are also distinguised by last consonants. We know that in the case the root ends with *c*, its reduplicant is one character longer than it. The other three transformations of last consonants do not change the length of the reduplicants with respect to that of the roots.

Hence the FST f_3 which recognizes the third type reduplication is required to modify the tones of the reduplicants with respect to those of the roots on the one hand, and to transform last consonants of the roots on the other hand. For example, the minimal FST f_3 recognizing four reduplicative words *biêng biếc* (bluish green), *biền biệt* (leave behind no traces whatsoever), *bình bịch* (a series of thudding blows) and *bôm bốp* (pop pop) is given in the figure below:



The minimal FST f_3 recognizing all 232 reduplicative words of the third type consists of 59

states in which 2 states are final ones. It has 262 transitions, the maximum number of outtransitions from a state is 19.

Once all the three transducers have been constructed, we can unify them by making use of the standard union operation on transducers to obtain a sequential FST which is able to recognize all the three class of reduplication presented above (Mohri, 1996; Mohri, 1997).

4.4 A Software Package

We have developed a Java software package named **vnReduplicator** which implements the above-mentioned computational model of Vietnamese reduplication. The core component of this package is a minimal FST which can recognize a substantial amount of reduplicative bi-syllabic words found in the Vietnamese language.

The first application of this core model which we have developed is a reduplication scanner for Vietnamese. We use the minimal FST of the core model to build a tool for fast detection of reduplication. The tool scans a given input text and produces a list of all the recognized reduplicative words. The detection process is very fast since the underlying transducer operates in optimal time in the sense that the time to recognize a syllable corresponds to the time required to follow a single path in the deterministic finite-state machine, and the length of the path is the length of the syllable measured in characters.

As an example, given the following input text

"Anh đi *biền biệt*. Cô vẫn chờ anh hơn 20 năm *đằng đẵng*."[2],

the scanner marks two reduplicative words as shown in the italic face.

We are currently investigating another useful application of the core model for a partial spell checking of Vietnamese text. It is observed that people may make typhographical errors in writing like *đằng đẳng* instead of the correct word *đằng đẵng*. In such cases, the computational model can be exploited to detect the potential errors and suggest corrections.

The reduplication model could also help improve the accuracy of Vietnamese lexical recognizers in particular and the accuracy of Vietnamese word segmentation systems in general. The reduplication scanner will be integrated to **vnTokenizer**[3] - an open source and highly accurate tokenizer for Vietnamese texts (Le et al., 2008).

The software and related resources will be distributed under the GNU General Public Lisence[4] and it will be soon available online[5].

5 Conclusion and Future Work

We have presented for the first time a computational model for the reduplication of the Vietnamese language. We show that a large class of reduplicative words can be modeled effectively by sequential finite-state string-to-string transducers.

The analysis of the various patterns of reduplication of the Vietnamese language has twofold contributions. On the one hand, it gives useful information on identification of spelling variants in Vietnamese texts. On the other hand, it gives an explicit formalization of precedence relationships in the phonology, and as a result helps ordering and modeling phonological processes before transfer of the presentation to the articulatory interface.

It is argued that the relation between morphology and phonology is an intimate one, both synchronically and diachronically. As mentioned earlier, Vietnamese reduplication is always accompanied by a modification of phone and tone for a symmetric and harmonic posture. We thus believe that the compact finite-state description of a large class of reduplication would help connect morphosyntactic attributes to individual phonological components of a set of Vietnamese word forms and contribute to the improvement of Vietnamese automatic speech recognition systems.

As mentioned earlier, the current work does not handle partial reduplication in which either the onset is repeated or the rhyme and the tone of syllables are repeated, for example *bồng bềnh* (bob), *chúm chím* (open slightly one's lips), *lẩm cẩm* (doting), *lúng túng* (perplexed, embarrassed). Partial reduplication is a topic which has been well studied for a long time by Vietnamese linguists community. It has been shown that partial reduplicative words also have certain principle formation rules (Diệp, 1999; UBKHXH, 1983). Hence, partial reduplicative words could also be generated and recognized by an appropriate finite-state

[2] He has left behind no traces whatsoever. She has been waiting for him for 20 years.

[3] *http://www.loria.fr/∼lehong/tools/vnTokenizer.php*

[4] *http://www.gnu.org/copyleft/gpl.html*

[5] *http://www.loria.fr/∼lehong/projects.php*

model which encodes precisely their formation rules. This is an interesting topic of our future work in constructing a rather complete computational model for Vietnamese bi-syllabic reduplication.

Furthermore, in addition to the bi-syllabic reduplication forms, there exists also three or four syllable reduplication forms, for example *cỏn còn con* (very little), *tẹo tèo teo* (very small), or *vội vội vàng vàng* (hurry), *đủng đà đủng đỉnh* (deliberate). These reduplication forms involve the copying operation of morphological structures which is a non-regular operation. Non-regular operations are problematic in that they cannot be cast in terms of composition – the regular operation of major importance in finite-state devices, while finite-state devices cannot handle unbounded copying. However, the question of the possibility for an elegant account to reduce these specific kinds of reduplication to purely regular mechanisms would be of interest for further research to extend and improve the core reduplication components for Vietnamese. Unknown reduplicative word guessing is another interesting and useful topic since the lexicon can never cover all reduplicative words.

Acknowledgement

We gratefully acknowledge helpful comments and valuable suggestions from three anonymous reviewers for improving the paper.

References

Yael Cohen-Sygal and Shuly Wintner. 2006. *Finite-State Registered Automata for Non-Concatenative Morphology*. Computational Linguistics, Vol. 32, No. 1, Pages 49–82.

Jan Daciuk, Stoyan Mihov, Bruce W. Watson and Richard E. Watson. 2000 *Incremental Construction of Minimal Acyclic Finite-State Automata*. Computational Linguistics, Vol. 26, No. 1, 2000.

Le H. Phuong, Nguyen T. M. Huyen, Roussanaly A., Ho T. Vinh. 2008 A hybrid approach to word segmentation of Vietnamese texts. *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain*. Springer LNCS 5196, 2008.

Diệp Quang Ban and Hoàng Văn Thung. 1999 *Ngữ pháp Tiếng Việt (Vietnamese Grammar)*. NXB Giáo dục, Hà Nội, Việt Nam.

Đoàn Thiện Thuật. 2003 *Ngữ âm tiếng Việt (Vietnamese Phonetics)*. NXB Đại học Quốc gia Hà Nội, Hà Nội, Việt Nam.

Đoàn Thiện Thuật (Editor-in-chief) and Nguyễn Khánh Hà and Phạm Như Quỳnh. 2003 *A Concise Vietnamese Grammar (For Non-native Speakers))*. Thế Giới Publishers, Hà Nội, Việt Nam.

Hữu Đạt and Trần Trí Dõi and Đào Thanh Lan. 1998 *Cơ sở tiếng Việt (Basis of Vietnamese)*. NXB Giáo dục, Hà Nội, Việt Nam.

Ronald Kaplan and Martin Kay. 1994. *Regular Models of Phonological Rule Systems*. Computational Linguistics, Vol. 20, No. 3, Pages 331–378.

Koskenniemi Kimmo. 1983 *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. The Department of General Linguistics, University of Helsinki.

Mehryar Mohri. 1996 *On Some Applications of Finite-State Automata Theory to Natural Language Processing*. Natural Language Engineering, Vol. 2, No. 1, Pages 61–80.

Mehryar Mohri. 1997 *Finite-State Transducers in Language and Speech Processing*. Computational Linguistics, Vol. 23.

Nguyễn Thị Minh Huyền, Laurent Romary, Mathias Rossignol and Vũ Xuân Lương. 2006. *A Lexicon for Vietnamese Language Processing*. Language Resources and Evaluation, Vol. 40, No. 3–4.

Tokunaga T., Kaplan D., Huang C-R., Hsieh S-K, Calzolari N., Monachini M., Soria C., Shirai K., Sornlertlamvanich V., Charoenporn T., Xia Y., 2008. *Adapting international standard for Asian language technologies*. Proceedings of The 6th International Conference on Language Resources and Evaluation (LREC 2008)

Tokunaga T. et al. 2008. *Developing International Standards of Language Resources for Semantic Web Applications* Research Report of the International Joint Research Program (NEDO Grant) for FY 2007, *http://www.tech.nedo.go.jp/PDF/100013569.pdf*

Tran D. D. and Castelli E. and Serignat J. F. and Trinh V. L. and Le X. H. 2006. *Linear F_0 Contour Model for Vietnamese Tones and Vietnamese Syllable Synthesis with TD-PSOLA*. Proceedings of TAL2006, La Rochelle, France.

Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong and John-Paul Hosom. 2006. *Vietnamese Large Vocabulary Continuous Speech Recognition*. Proceedings of Eurospeech 2005, Lisboa.

Ủy ban Khoa học Xã hội Việt Nam. 1983. *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. Nhà xuất bản Khoa học Xã hội – Hà Nội, Việt Nam.