

# Vietnamese word and syllabeme (syllable-morpheme) frequencies: A corpus and lexical decision study

Hien Pham, Patrick Bolger & R. Harald Baayen

University of Alberta  
[hpham@ualberta.ca](mailto:hpham@ualberta.ca)

SEALS 22  
Agay - France  
May 31, 2012

- 1 Introduction
- 2 Corpora survey
- 3 Vietnamese language
- 4 Aims
- 5 Data and methods
- 6 An overview of the corpora
- 7 Results
- 8 Discussions
- 9 References

# Distance measurement



How far is the flagged hole from the golf player?

# Distance measurement



How far is the flagged hole from the golf player?



Some ethnic minorities may use a throw of this kind of knife as a unit to measure.

Does it matter?

Does it matter?

The answer is Yes!

Does it matter?

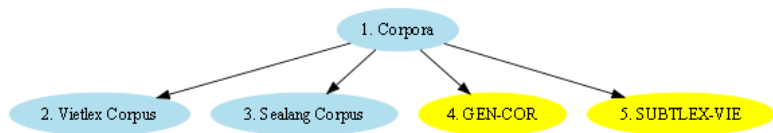
The answer is **Yes!**

Because these reflexes show how people divide chaos into patterns or categories.





# Vietnamese corpora



## Characteristics of Vietnamese language

- Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number and tense. Its grammar relies on word order and function words.

## Characteristics of Vietnamese language

- Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number and tense. Its grammar relies on word order and function words.
- Modern script uses the Vietnamese alphabet *chữ quốc ngữ*, or 'national script' based on a Romanized script expanded with diacritics to mark vowels, consonant, and tones.

## Characteristics of Vietnamese language

- Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number and tense. Its grammar relies on word order and function words.
- Modern script uses the Vietnamese alphabet *chữ quốc ngữ*, or 'national script' based on a Romanized script expanded with diacritics to mark vowels, consonant, and tones.
- Vietnamese is a shallow transparent orthographic language, with a nearly one-to-one grapheme-to-phoneme correspondence.

# Characteristics of Vietnamese language

- Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number and tense. Its grammar relies on word order and function words.
- Modern script uses the Vietnamese alphabet *chữ quốc ngữ*, or 'national script' based on a Romanized script expanded with diacritics to mark vowels, consonant, and tones.
- Vietnamese is a shallow transparent orthographic language, with a nearly one-to-one grapheme-to-phoneme correspondence.
- Single syllables are separately written, the one-to-one mapping of syllable (*âm tiết*) and morpheme (*hình vị* - separated by two spaces) leads to the concept of *syllabeme* (*tiết vị* or *tiếng*) in Vietnamese linguistics.

# Characteristics of Vietnamese language

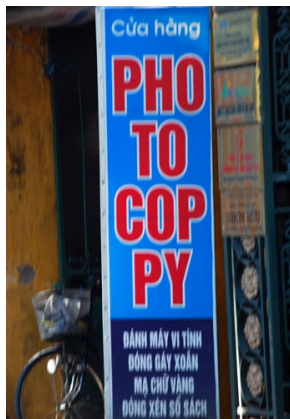
- Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number and tense. Its grammar relies on word order and function words.
- Modern script uses the Vietnamese alphabet *chữ quốc ngữ*, or 'national script' based on a Romanized script expanded with diacritics to mark vowels, consonant, and tones.
- Vietnamese is a shallow transparent orthographic language, with a nearly one-to-one grapheme-to-phoneme correspondence.
- Single syllables are separately written, the one-to-one mapping of syllable (*âm tiết*) and morpheme (*hình vị* - separated by two spaces) leads to the concept of *syllabeme* (*tiết vị* or *tiếng*) in Vietnamese linguistics.
- Dealing with the syllable-word illusion is not only difficult, but also an interesting topic of natural language processing but also in psycholinguistics since more than 70% of words are compounds in the language.

# Phở vương



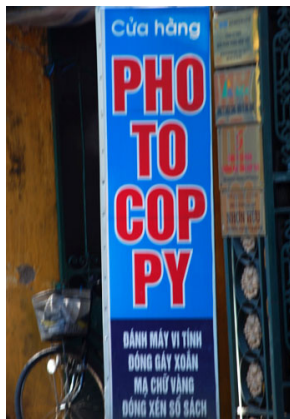
<http://phovuong.vn/>

What does this tablet say?





## What does this tablet say?



“We thought that this was a Vietnamese phrase. It took us awhile to realize it said *Photocopy*. We thought it was some kind of special phở dish.” said a tourist.

## Aims of this study

- We present a new database of Vietnamese word frequencies based on film and television subtitles (SUBTLEX-VIET) and a general newspapers and books (GEN-COR) corpora.

# Aims of this study

- We present a new database of Vietnamese word frequencies based on film and television subtitles (SUBTLEX-VIET) and a general newspapers and books (GEN-COR) corpora.
- We validated these frequencies with lexical decision times of about 20,000 monosyllabic and disyllabic Vietnamese words.

# Data & Methods

- Subtitles are freely available online and are a good source of spoken language.

# Data & Methods

- Subtitles are freely available online and are a good source of spoken language.
- More than 13,000 Vietnamese subtitles have been collected and processed that gave us a corpus of about 80 million words.

# Data & Methods

- Subtitles are freely available online and are a good source of spoken language.
- More than 13,000 Vietnamese subtitles have been collected and processed that gave us a corpus of about 80 million words.
- A general corpus (newspaper articles, stories and novels) has been built and used as a reference corpus with about 100 million words.

# Data & Methods

- Subtitles are freely available online and are a good source of spoken language.
- More than 13,000 Vietnamese subtitles have been collected and processed that gave us a corpus of about 80 million words.
- A general corpus (newspaper articles, stories and novels) has been built and used as a reference corpus with about 100 million words.
- The corpora is tokenized and tagged by vnTokenizer [Lê *et al.*2008] and vnTagger [Lê *et al.*2010].

# Data & Methods

- Subtitles are freely available online and are a good source of spoken language.
- More than 13,000 Vietnamese subtitles have been collected and processed that gave us a corpus of about 80 million words.
- A general corpus (newspaper articles, stories and novels) has been built and used as a reference corpus with about 100 million words.
- The corpora is tokenized and tagged by vnTokenizer [Lê *et al.*2008] and vnTagger [Lê *et al.*2010].
- We computed *word frequencies* (the number of time each word was encountered) and *dispersion* (the number of films or documents in which it appeared). All frequencies were transformed to  $\log_{10}$ .



## A sketch of plain text corpora

Các em Nam sinh không được gây mất trật tự.  
Không được đánh nhau.  
Hôm nay là lễ khai giảng  
Đừng tự phá hỏng buổi lễ của chính mình.  
Đừng làm ba mẹ các em thất vọng.  
Học sinh mới năm nay ghê quá.  
Rắc rối rồi! Yakuza  
đang ở trong sân trường!  
Ai đó gọi cảnh sát đi! Nhanh lên!  
Ề

## A sketch of tokenized corpora

Các em Nam\_sinh không được gây mất trật\_tự .  
Không được đánh nhau .  
Hôm\_nay là lễ khai\_giảng  
Đừng tự phá hỏng buổi lễ của chính mình .  
Đừng làm ba\_mẹ các em thất\_vọng .  
Học\_sinh mới năm nay ghê quá .  
Rắc\_rối rồi ! Yakuza  
đang ở trong sân trường !  
Ai đó gọi cảnh\_sát đi ! Nhanh lên !  
Ê

## A sketch of tagged corpora

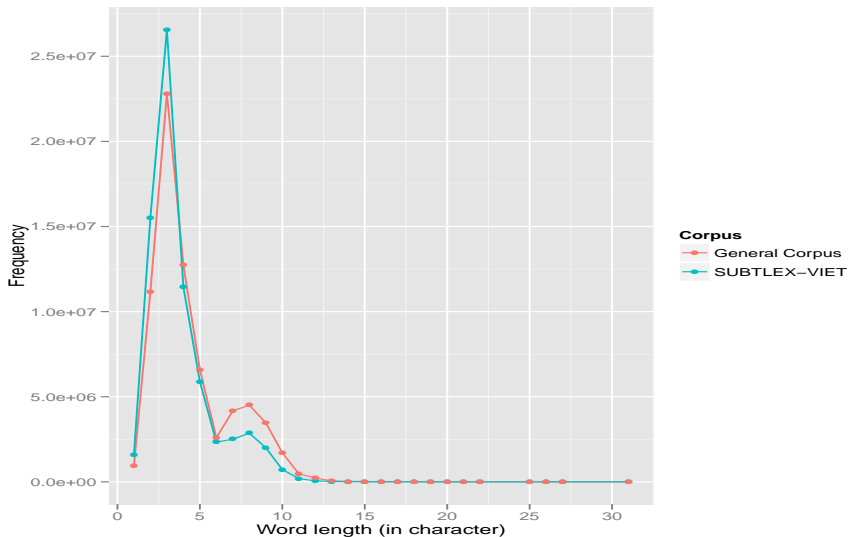
Các/L em/N Nam\_sinh/N không/R được/V gây/V mất/V trật\_tự/N ./.  
Không/R được/V đánh/V nhau/N ./.  
Hôm\_nay/N là/V lễ/N khai\_giảng/V  
Đừng/R tự/P phá/V hỏng/A buổi/N lễ/N của/E chính/T mình/P ./.  
Đừng/R làm/V ba\_mẹ/N các/L em/N thất\_vọng/V ./.  
Học\_sinh/N mới/A năm/N nay/P ghê/A quá/R ./.  
Rắc\_rối/A rồi/C !! Yakuza/Np  
đang/R ở/V trong/E sân/N trường/N !!  
Ai/P đó/P gọi/V cảnh\_sát/N đi/V !! Nhanh/Np lên/R !!  
Ê/I

## Words and their equivalences

	Word	NoTone	NoDiacritics	North2
1	ốm nghén	ôm n̄ghen	om n̄ghen	om5 ɲɛn5
2	âu phục	âu phuc	au phuc	ǎw1 fukp̄6
3	giá bìa	gia bia	gia bia	za5 biə2
4	giả bộ	gia bô	gia bo	za4 bo6
5	giã biệt	gia biêť	gia biet	za3 biət6
6	gia bình	gia binh	gia binh	za1 biŋ1
7	giá mà	gia ma	gia ma	za5 ma2
8	già mồm	gia môm	gia mom	za2 mom2
9	gia cố	gia cô	gia co	za1 ko5
10	già cõi	gia côi	gia coi	za2 koj3

**Figure:** Words and their equivalences. The IPA has been computed with the vPhon tool [Kirby2008].

# A comparison of SUBTLEX-VIET and GEN-COR



Distribution of summed word frequency as a function of word length (measured in number of characters)

## Visual lexical decision experiment

- *Stimuli*: The study involved mono- and disyllabic words. We took all the mono- and disyllabic words based on a Vietnamese Dictionary (except for the one character words) [Vien Ngon ngu hoc2000]. This resulted in a total of 21,498 words.

## Visual lexical decision experiment

- *Stimuli*: The study involved mono- and disyllabic words. We took all the mono- and disyllabic words based on a Vietnamese Dictionary (except for the one character words) [Vien Ngon ngu hoc2000]. This resulted in a total of 21,498 words.
- The Wuggy pseudoword generator [Keuleers Brysbaert2010] was used to construct a corresponding pseudoword for each word in the experiment, i.e., 21,498 nonwords generated.

## Visual lexical decision experiment

- *Stimuli*: The study involved mono- and disyllabic words. We took all the mono- and disyllabic words based on a Vietnamese Dictionary (except for the one character words) [Vien Ngon ngu hoc2000]. This resulted in a total of 21,498 words.
- The Wuggy pseudoword generator [Keuleers Brysbaert2010] was used to construct a corresponding pseudoword for each word in the experiment, i.e., 21,498 nonwords generated.
- *Participant*: The single-subject participant in this study is a native Vietnamese speaker.

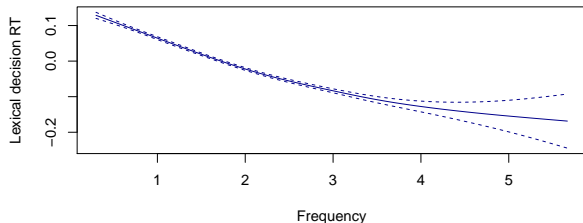


## Visual lexical decision experiment

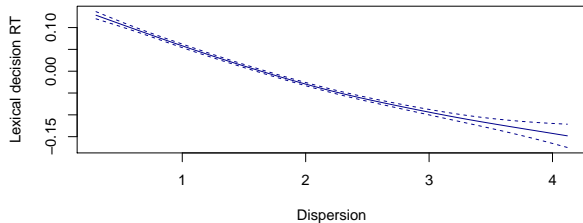
- *Stimuli*: The study involved mono- and disyllabic words. We took all the mono- and disyllabic words based on a Vietnamese Dictionary (except for the one character words) [Vien Ngon ngu hoc2000]. This resulted in a total of 21,498 words.
- The Wuggy pseudoword generator [Keuleers Brysbaert2010] was used to construct a corresponding pseudoword for each word in the experiment, i.e., 21,498 nonwords generated.
- *Participant*: The single-subject participant in this study is a native Vietnamese speaker.
- *Procedure*: Participant was tested in a noise-attenuated experimental room. Each visual stimulus was preceded by a fixation mark in the middle of the screen for 500 ms. After that the stimulus appeared at the same position. Each word remained on the screen until the participant's response or 2000 milisecond elapsed. A new trial was initiated 500 ms afterwards.

# Results: Partial effects of frequency and dispersion

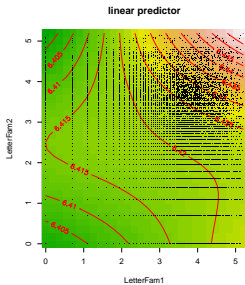
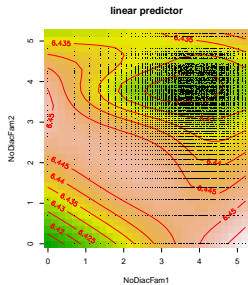
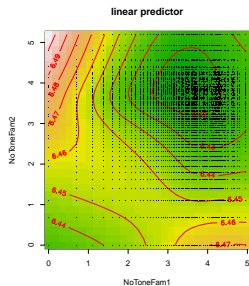
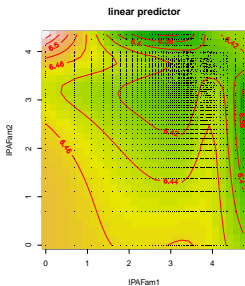
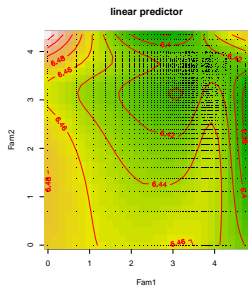
## Frequency effects



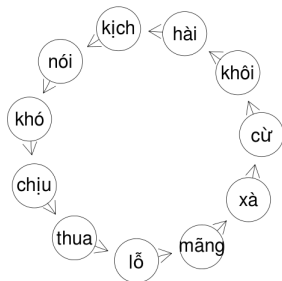
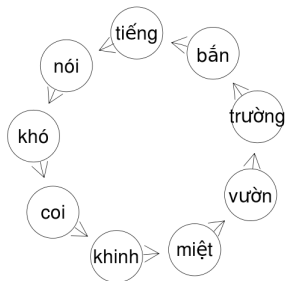
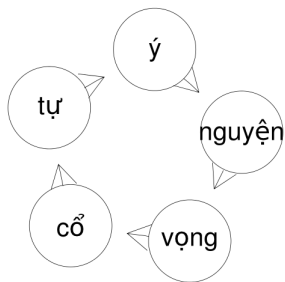
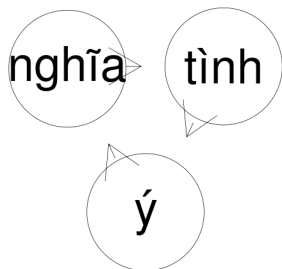
## Dispersion effects



# GAMs models



## Strongly connected compounds



# Discussions

- The study presents a frequency measure for Vietnamese, based on subtitles and general corpora, with a lexical decision validation study involving monosyllabic (except one-character words) and disyllabic Vietnamese words.

# Discussions

- The study presents a frequency measure for Vietnamese, based on subtitles and general corpora, with a lexical decision validation study involving monosyllabic (except one-character words) and disyllabic Vietnamese words.
- The study reveals that *dispersion* can be used to better predict word recognition performance. Therefore, we reckon that the SUBTLEX-VIET frequencies will be of valuable use for language research, especially in the psycholinguistic study, such as word recognition research.

## Discussions (cont.)

- Validating the obtained frequencies with lexical decision times is a good practice. It is the motivation for us to collect a large amount of subtitles to build a frequency database.

## Discussions (cont.)

- Validating the obtained frequencies with lexical decision times is a good practice. It is the motivation for us to collect a large amount of subtitles to build a frequency database.
- The results show that the mean RTs were faster for disyllabic words in comparison with mono-syllabic words (644 and 703 milliseconds, respectively).



## Discussions (cont.)

- Validating the obtained frequencies with lexical decision times is a good practice. It is the motivation for us to collect a large amount of subtitles to build a frequency database.
- The results show that the mean RTs were faster for disyllabic words in comparison with mono-syllabic words (644 and 703 milliseconds, respectively).
- It might be because in Vietnamese, there are number of word formation units which can act as a morpheme in compounds but can not be a mono-syllabic word. In the context of isolated word recognition, in which no contextual information is provided, readers need to put the character into some contexts in their inner voice to figure out whether it is a word or not. It explains why time-course for recognizing mono-syllabic words is longer than that of disyllabic words.

## Discussions (cont.)


- High frequency words (based on the general corpus) were responded to 48 ms faster than low frequency words. Interestingly, *dispersion* (also known as *contextual diversity*), derived from the subtitle corpus, emerges as a better predictor over the observed frequency itself.


## Discussions (cont.)


- High frequency words (based on the general corpus) were responded to 48 ms faster than low frequency words. Interestingly, *dispersion* (also known as *contextual diversity*), derived from the subtitle corpus, emerges as a better predictor over the observed frequency itself.
- This finding supports the repeat effects in learning i.e., those words most often repeated in different contexts or sessions are best memorized and take a shorter time to retrieve.

Thank you!

# References I

 KEULEERS, EMMANUEL, MARC BRYLSBAERT.  
2010.  
Wuggy: A multilingual pseudoword generator.  
*Behav Res Methods* 42.627–633.

 KIRBY, JAMES.  
2008.  
vPhon: a Vietnamese phonetizer (version 0.2.4).  
Retrieved on April, 2011 from <http://home.uchicago.edu/~jkirby/>.

 LÊ, HONG PHUONG, THI MINH HUYEN NGUYEN, AZIM  
ROUSSANALY, VINH HO.  
2008.  
A hybrid approach to word segmentation of Vietnamese texts.  
In *Language and Automata Theory and Applications*, ed. by Carlos  
Martin-Vide, Friedrich Otto, Henning Fernau, volume 5196 of *Lecture  
Notes in Computer Science*, 240–249. Heidelberg: Springer Berlin.

## References II



LÊ, HONG PHUONG, AZIM ROUSSANALY, THI MINH HUYEN  
NGUYEN, MATHIAS ROSSIGNOL.

2010.

An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts.

In *Traitement Automatique des Langues Naturelles - TALN 2010*, p. 12, Montréal Canada. ATALA (Association pour le Traitement Automatique des Langues).



VIỆN NGÔN NGỮ HỌC.

2000.

*Từ điển tiếng Việt [Vietnamese Dictionary]*.

Hà Nội - Đà Nẵng: Nhà xuất bản Đà Nẵng, Trung tâm Từ điển học.