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

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

1. Corpora
   - 2. Vietlex Corpus
   - 3. Sealang Corpus
   - 4. GEN-COR
   - 5. SUBTLEX-VIE
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.
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 tiếţ) and morpheme (hình vị - separated by two spaces) leads to the concept of syllabeme (tiếţ 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 vuông

http://phovuong.vn/
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.

Pham, Bolger, & Baayen (UofA) Corpus - Lexical decision study

Agay-France, May 31, 2012 9 / 26
“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.
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].

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 log10.
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 $\log_{10}$. 
Các em Nam sinh không được gây mắt tratamiento.
Không được đánh nhau.
Hôm nay là lễ khai giảng
Đứng tự phá hỏng buổi lễ của chính mình.
Đứng làm ba mẹ các em thật vọng.
Học sinh mới năm nay ghé quá.
Rắc rối rỗi! Yakuza
dạng ở trong sân trường!
Ai dớ gởi cảnh sát đi! Nhanh lên!
Ê
Các em Nam_sinh không được gây mất trata_tự.
Không được đánh nhau.
Hôm_nay là lễ khai_giang
Đừng tự phá hộng buơi lễ của chính mình.
Đừng làm ba mãe các em thất_vọng.
Học_sinh mới năm nay ghé quá.
Rác_rơi rơi ! Yakuza
dạng 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 trata_tụ/N ./. Không/R được/V đánh/V nhau/N ./. Hôm_nay/N là/V lẽ/N khai_giang/V
Đứng/R tự/P phá/V họng/A buổi/N lẻ/N của/E chính/T mình/P ./. Đứng/R làm/V ba_mẹ/N cát/L em/N thất_vọng/V ./. Học_sinh/N mới/A năm/N nay/P ghê/A quá/R ./. Rác_rỏi/A rỏi/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

<table>
<thead>
<tr>
<th>Word</th>
<th>NoTone</th>
<th>NoDiacritics</th>
<th>North2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ṽm nghén</td>
<td>Ṽm nghen</td>
<td>om nghen om5 ŋẹn5</td>
</tr>
<tr>
<td>2</td>
<td>Ṽu phuc</td>
<td>Ṽu phuc</td>
<td>au phuc ŋwl fukp6</td>
</tr>
<tr>
<td>3</td>
<td>gia bia</td>
<td>gia bia</td>
<td>gia bia za5 bia2</td>
</tr>
<tr>
<td>4</td>
<td>gia bô</td>
<td>gia bô</td>
<td>gia bo za4 bo6</td>
</tr>
<tr>
<td>5</td>
<td>gia biêt</td>
<td>gia biêt</td>
<td>gia biet za3 biêt6</td>
</tr>
<tr>
<td>6</td>
<td>gia bính</td>
<td>gia bính</td>
<td>gia bính za1 bìn1</td>
</tr>
<tr>
<td>7</td>
<td>gia mà</td>
<td>gia ma</td>
<td>gia ma za5 ma2</td>
</tr>
<tr>
<td>8</td>
<td>gia môm</td>
<td>gia môm</td>
<td>gia mom za2 mom2</td>
</tr>
<tr>
<td>9</td>
<td>gia cô</td>
<td>gia cô</td>
<td>gia co za1 ko5</td>
</tr>
<tr>
<td>10</td>
<td>gia cội</td>
<td>gia cội</td>
<td>gia co za2 koj3</td>
</tr>
</tbody>
</table>

**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.

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 millisecond elapsed. A new trial was initiated 500 ms afterwards.
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

Pham, Bolger, & Baayen (UofA)  
Corpus - Lexical decision study  
Agay-France, May 31, 2012
GAMs models

Partial regression surfaces modeled with tensor products for lexical decision response latencies for the interaction.
Strongly connected compounds

nghià  tinh
ý
nguyen
co
vong
tu

noi tieng ban
khoe
coi
khinh
miet
truong
vo
chu
thua
lo
manh
kich
hai
khoe
cu
xa
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

- 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.
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 cannot 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.
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 cannot 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.)

- 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 cannot 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.
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.
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 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).

Từ điển tiếng Việt [Vietnamese Dictionary].
Hà Nội - Đà Nẵng: Nhà xuất bản Đà Nẵng, Trung tâm Từ điển học.