Visual Processing of Vietnamese Compound Words: A Multivariate Analysis Using Corpus Linguistic and Psycholinguistic Paradigms

by

Hien Pham

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

University of Alberta

Abstract

This dissertation investigates disyllabic compound words in Vietnamese, an isolating tone language, using corpus linguistics and psycholinguistic experimental paradigms. Chapter 2 reports the construction of two corpora and a database of wide range of lexical variables. Chapter 3 discusses a visual lexical decision experiment with Vietnamese speakers, and Chapter 4 details the results of a visual lexical naming experiment, again with Vietnamese speakers. This dissertation supports the psychological status of the word as a single whole in non-decompositional model for reading Vietnamese compounds. The dissertation also documents the advantages of working with behavioral data from a single-subject with wide range of items. Finally, this dissertation demonstrates the involvement of phonology in silent reading through an analysis of the effects of lexical tone in the visual lexical decision task.

Preface

This thesis is an original work by Hien Pham. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "Lexical processing and representation of disyllabic compound words in reading Vietnamese", No. Pro00010236, March 22, 2011 and No. Pro00010236_REN1, February 28, 2012.

Some of the research conducted for this thesis forms part of a research collaboration with Dr. R. Harald Baayen and Dr. Benjamin V. Tucker at the University of Tübingen and University of Alberta. The technical apparatus referred to in chapters 2, 3, and 4 were designed by myself. The data analysis in chapter 2, 3 and 4 are my original work, as well as the literature review in chapter 1.

A revised version of chapter 3 of this thesis has been submitted for publication as Hien Pham and Harald Baayen (2014) "Vietnamese compounds show an anti-frequency effect in visual lexical decision". I was responsible for the data collection and analysis as well as the manuscript composition. Harald Baayen assisted with the data analysis and contributed to manuscript edits. Harald Baayen was the supervisory author and was involved with concept formation and manuscript composition.

Chapters 2 and 4 are currently being prepared for submission for publication. Harald Baayen and Benjamin Tucker assisted with the data analysis and contributed to manuscript edits. The dissertation author was the primary investigator and author

of this material.

Cho cha mẹ tôi

Acknowledgements

The research in this dissertation would not have been possible without the mentoring, guidance and support of my supervisors Dr. R. Harald Baayen and Dr. Benjamin V. Tucker, and the support of my former supervisor Dr. Patrick Bolger and the members of the CCP Lab. Dr. Harald Baayen, Dr. Benjamin Tucker, Dr. John Newman, Dr. Christina Gagné and Dr. Antti Arppe provided advice and inspiration throughout the development of this dissertation.

Funding for this dissertation was provided by the Project 322 Scholarship from Vietnam Government, the Graduate Student Research Assistant Award, the GSA Doctoral Travel Fund Award, numerous Travel Awards from the Department of Linguistics at the University of Alberta and Travel Fund Awards from SEALS.

Huge gratitude is owed to my family, especially my wife, who has taken care of our son and daughter while I was not there with them. They have been immensely supportive and patient with me throughout the long separation, that is this dissertation.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Psycholinguistic research, including behavioral, neurophysiological, and computational paradigms, has found that models of sequential processing have considerable explanatory power (see e.g., Frank et al., 2012, for a review). Further, there is ample evidence indicating that language users are sensitive to the frequency of linguistic units on many levels. For example, Hay et al. (2004) observed frequency effects in the phonotactics of language use; while Fidelholz (1975) and Jurafsky et al. (2001) documented effects of frequent one and two-word sequences. In recent studies on multi-word phrases, phrase frequency (e.g., n-grams) is found to affect the way in which language users interact with language (Arnon and Snider, 2010; Bannard and Matthews, 2008; Conklin and Schmitt, 2007; Janssen and Barber, 2012; Siyanova-Chanturia et al., 2011; Tremblay and Tucker, 2011). These effects of multi-word phrases are also found beyond the bigram level in children's language knowledge (Bannard et al., 2009; Lieven et al., 2009). This behavioral data has been found to be consistent with usage-based approaches of grammar (Bod, 1998; Bybee, 2006; Goldberg, 2005), where sequences coexist with their parts and recognition is influenced by many linguistic factors, such as n-gram frequency, mutual information between parts, and their constituents, etc.

Although many studies have been carried out, the structure of lexical representation and the mechanism of lexical (de)composition remain controversial. The present dissertation focusses on several general aspects of this research. First, what is the psychological reality of morphological complexity in the representation of compound words? Second, do models of lexical processing based largely on Indo-European languages apply to other languages, such as isolating languages? To address these general questions, research from a comparative psycholinguistic perspective needs to be carried out.

This dissertation investigates the processing of compound words in Vietnamese, an isolating tone language of the Mon-khmer language family, using corpus linguistic and psycholinguistic paradigms. Vietnamese appears to be a good testbed for studying compounds since this language is characterized by the use of a limited number of syllables to construct compound words and phrases. Interestingly, there are no word boundaries in the orthography of this language. The fundamental architecture of the Vietnamese lexical system is the combination of morphemes, syllables, and orthographic graphemes that are mapped to a single form. These characteristics can be used to test models of language processing for compounds and word sequences because of the strict way in which these forms are represented in the language. In English, for instance, compounds can be written without a space between two constituents such as *bedroom*, *football*, *strawberry*, *moonshine*; but can also be written with a space such as *foster father*; or be written with a hyphen between two constituents such as *blue-green*, *people-carrier*. In Vietnamese, every single syllable-morpheme unit is written between two spaces without any extra formal marker for word boundary. This could be viewed as an ideal test-bed for probing models of storage versus computation (or nondecomposition versus decomposition) in the processing of compounds and sequences of words.

The aim of this dissertation is to investigate the language-specific mental processes and to understand the mechanisms involved in reading an alphabetic-isolating language: Vietnamese. Specifically, I seek: 1) to investigate the advantages and disadvantages of single versus multiple-subject experiments; 2) to investigate the role of morphological decomposition in reading and the role of phonology in accessing meaning; 3) to examine the interactions of independent variables in reading and how they influence lexical processing times; and 4) to investigate the interactivity between different types of lexical knowledge in visual word recognition and to identify a set of factors that may be critical to the lexical processing.

In the remainder of this chapter I discuss in greater detail the main research questions addressed in this dissertation. This is followed by a brief description of each chapter.

Key research questions

This work will provide insight into the processing of compound words during reading and the role of compositional integration in reading of syllabemes in compounds. Vietnamese offers an ideal language for investigating these questions because there are no word boundaries to separate out words as larger units. I hope to shed further light on the nature of conceptual combination in visual word recognition and word sequence reading.

I also hope to contribute to several areas of morphological inquiry that are either under-researched or controversial. I also hope to propose revisions to existing models of compound processing. Specifically, I aim to address the following questions on the role of morphology in lexical processing.

What is the role of morphological decomposition?

- *Is detection of the constituent syllabemes facilitating or inhibiting recognition of compounds?*

Psycholinguistic research has focused on the representation and processing of many forms of morphologically complex words, such as derivational, inflectional, and compound words. These types of morphology have been treated separately in the literature, perhaps due to the assumption that these morphological types are processed differently. Indeed, these morphological types have different underlying characteristics. Inflectional morphology does not result in a new lexical entity, while derivational and compound morphology do. Inflection, the morphological indication of the grammatical subclass to which a word belongs (e.g., the *-s* in *trees* marks the plural subclass), involves a consistent and predictable semantic change. Inflection never results in a change of grammatical class, e.g., the *-s* in *books* marks the plural subclass (see e.g., Baayen et al., 1997a, 2003; Bertram et al., 2000b, for the regular and irregular inflectional morphology). Derivational morphology is an affixal process that forms a word with a meaning and/or category distinct from that of its base, e.g., the suffix *-er*, which combines with a verb to form a noun with the meaning 'one who does X', e.g., *writer*, and *reader*. Recent studies have investigated pseudo-derivational constructs using words such as derived-stems priming of opaque morphological pairs (e.g., *corn* and *corner*), non-morphological form pairs (e.g., *broth* and *brother*), and transparent morphological pairs (e.g., *hunt* and *hunter*) (see e.g., Feldman et al., 2009; Longtin et al., 2003; Longtin and Meunier, 2005; Rastle et al., 2004; Rastle and Davis, 2008; Rueckl and Aicher, 2008, among others). Semantically, pseudo-derivational words are unrelated, whereas truly derivational words are related, e.g., *work – worker*, *garden – gardener*. Compounding involves the com-

bination of two or more existent roots or words to form a larger word e.g., *moon walk*, *honeymoon*, whereas inflection and derivation are both commonly marked by affixation with one root. Considering morpho-orthographic segmentation, however, significant similarities between morphological effects are observed between derived and compound words. Fiorentino and Fund-Reznicek (2009) reported equivalent and significant masked priming effects for both transparent (teacup–TEA) and opaque compounds (carpet–CAR, honeymoon–HONEY) compared with orthographic, non-morphological controls (penguin–PEN).

One of the fundamental questions of psycholinguistics is how people store words or even word sequences, and how they retrieve those units when using language. This question is even more challenging for morphologically complex words, e.g., *teacup*, *table cloth* or *moon-shine*. How are these words represented and accessed in the mind? The role of morphological complexity in the representation and processing of compound words and inflectionally affixed words has been the topic of many studies (e.g., Domínguez et al., 2000; Forster, 1989; McQueen and Cutler, 1998; Seidenberg and Gonnerman, 2000; Taft, 1991). Over the last 40 years, the experimental literature on processing of compound words has investigated many languages. Structurally, these compound words are built from two stems in which the semantic relation between two constituents of the whole word may be either transparent (e.g., *teatable*), or opaque (e.g., *moonshine*) (Bauer, 1983; Downing, 1977; Levi, 1978; Spencer, 1991). Research on morphological complexity in the psycholinguistic literature has supported both decompositional and non-decompositional types of models.

First, *the obligatory decomposition model* has been proposed, according to which morphologically structured words are automatically and obligatorily decomposed into their morphemic subunits, which then trigger the lexical representation of the whole word (Taft, 1994, 2003). Within this model, prelexical processing is regarded

as semantically 'blind', i.e., the parsing process only relies on the orthographic features of the morphemes and decomposes any letter string, i.e., *farmer* and also *corner*.

The second model, a *supralexical model* of morphological decomposition, proposes that the decomposition of a letter-string occurs only after the whole word has been accessed in the lexicon (Giraudo and Grainger, 2001, 2003). After that, the morphemic representations activate higher level semantic representations, which send back activation to the corresponding form representations. This approach thus differs from the first approach in that morphological decomposition involves a semantically based search for morphemes (morphosemantic decomposition), which is only successful for true morphological structures, i.e., *farmer* is parsed into *farm-er* but *corner* is not.

The third approach, *form-then-meaning mapping model*, postulates that morphological decomposition is initiated by a purely orthographic type of analysis and is then followed by a decomposition mechanism relying on the syntactic properties of a word (Crepaldi et al., 2010). Similar to the first model type, this type of model proposes that morphologically complex words are always initially recognized on the basis of semantically independent decomposition at prelexical stages in visual word recognition (Longtin and Meunier, 2005; McCormick et al., 2008, 2009; Rastle et al., 2004). Nevertheless, it differs from the first model in that the initial morpho-orthographic processing stage is followed by a lemma level at which inflected word forms (e.g., *trees, told*, etc.) are mapped onto their infinitives (e.g., *tree, tell*, etc.), which are then mapped onto the semantic level.

Fourth, the *parallel model* (or *hybrid model*) proposes parallel mapping onto both prelexical form representations and supralexical semantically dependent representations (Baayen et al., 1997a; Diependaele et al., 2009; Feldman et al., 2009). This

approach explicitly claims that morphologically complex words are listed in the lexicon, i.e., morphemes do not necessarily mediate access to the whole-word form. Contrary to the form-then-meaning mapping model, in the parallel model, morpho-orthographic and morpho-semantic decomposition can occur in parallel at early initial processing stages in visual word recognition. The form-then-meaning and parallel models are both based on the assumption that morphological decomposition does not exclusively rely on one single segmentation mechanism.

Of the models mentioned above, the two-staged process (an early blind morphemic decomposition process preceding a late process of semantic recombination) has an influential place in the current literature on morphological processing. The form-then-meaning account argues that in reading, complex words undergo obligatory and automatic decomposition into morphemes (Rastle and Davis, 2008; Rastle et al., 2004; Taft, 2004; Taft and Ardasinski, 2006). This account is challenged, however, by experimental evidence from current eye-tracking studies of Dutch, English, Finnish, Italian, which reveal that lexical information associated with word semantics reliably affect the earliest eye fixation on the word (Juhasz and Berkowitz, 2011; Kuperman et al., 2008, 2009; Marelli and Luzzatti, 2012). Given the phonotactic restrictions on syllabemes in Vietnamese, do syllabemes in compound reading show strong inhibitory effects, instead of the facilitatory effects usually observed for English and related Indo-European languages?

- *What is the relative importance of phrase frequency in reading Vietnamese?*

Recent studies in psycholinguistics provide evidence for frequency effects for multi-word sequences in both perception (Arnon and Snider, 2010; Bannard and Matthews, 2008; Siyanova-Chanturia et al., 2011; Tremblay and Baayen, 2010; Tremblay et al., 2011) and production (Alario et al., 2002; Janssen and Barber, 2012; Tremblay and Tucker, 2011) in English. Are phrase frequency effects involving compounds and

classifiers detectable in Vietnamese?

- *What is the role of syllabic effects in visual word recognition?*

The syllable has been proposed as a sublexical unit in visual word recognition (Alvarez et al., 2004; Carreiras et al., 1993; Colé et al., 1999; Goslin et al., 2006; Perea and Pollatsek, 1998; Taft and Forster, 1976). However, other studies argued that there is no need to posit intermediate representations between orthogrpahic and lexical units (see e.g., Seidenberg, 1987). Evidence to support the syllabic processing in visual word recognition was obtained from languages with clear syllable boundaries and shallow orthography. A number of experiments have observed that syllable frequency predicts RTs for words in Spanish (Alvarez et al., 2001; Carreiras et al., 1993; Perea and Pollatsek, 1998), French (Mathey and Zagar, 2002) and German (Conrad and Jacobs, 2004; Conrad et al., 2006, 2007, 2008). With an inventory of syllabemes, what type of syllabic effects will be found in Vietnamese? Are they inhibitory or facilitatory?

Is dispersion a better measure than frequency?

In addition to frequency of occurence, *contextual diversity* (also known as *dispersion* in corpus linguistics and statistics) can be considered an important variable in psycholinguistics (Adelman et al., 2006). The dispersion measure gauges to what extent words are used uniformly or non-uniformly across corpora. The dispersion statistic was proposed over forty years ago by Juilland, one of the pioneers of corpus linguistics (Juilland et al., 1970). This type of measure has been reviewed and extended in more recent studies by Gries (2008, 2009).

From the psycholinguistic perspective, McDonald and Shillcock (2001) and Adelman et al. (2006) have documented that the number of times a word occurs in a corpus is

less informative than the number of documents in which the word occurs. It has also been confirmed by other studes (e.g., New et al., 2007; Brysbaert and New, 2009; Keuleers et al., 2010) for subtitles in which dispersion measures accounted for 1% – 3% more of the variance in lexical decision performance than did word frequencies. Baayen (2010, p. 456) finds that "most of the variance in lexical space is carried by a principal component on which contextual measures (syntactic family size, syntactic entropy, BNC dispersion, morphological family size, and adjectival relative entropy) have the highest loadings. Frequency of occurrence, in the sense of pure repetition frequency, explains only a modest proportion of lexical variability." Our prediction is that dispersion measures (introduced in Adelman et al., 2006) will better predict responses in visual word recognition in Vietnamese.

Semantic interpretation of Vietnamese compounds

Where do the meanings of words exist? Are they in dictionaries? Are they in our mind? These simple questions turn out to be very crucial. Landauer and Dumais (1997) proposed to quantify word meaning by means of vector spaces with latent semantic analysis (LSA). With this model, a word's meaning is derived from its co-occurences across different contexts/documents. The latent semantic vector spaces have been used in several studies on morphological processing to characterize degrees of semantic transparency (e.g., Gagné and Spalding, 2009; Jones et al., 2006; Moscoso del Prado Martín and Sahlgren, 2002; Moscoso del Prado Martín et al., 2005; Pham and Baayen, 2013; Rastle et al., 2004). In this dissertation, I expect that LSA and the related Hyper Analog to Language (HAL) (Burgess and Lund, 1997) measures for Head-Compound similarity will also help predict the semantic interpretation of compounds.

Are more productive tones read faster?

Vietnamese is a tonal language with 6 tones that are obligatory components of Vietnamese syllabemes. The distribution of tones across all the syllabemes indicates that some tones are more productive than others: *ngang* mid level (1285), *huyền* low falling (breathy) (1078), *ngã* mid rising, glottalized (468), *hỏi* mid falling(-rising)(793), *sắc* mid rising, tense (1704), *nặng* mid falling, glottalized, short (1309). Several questions arise from the fact that tones are marked orthographically: (1) Are more productive tones recognized faster? (2) Are certain combinations of tones easier to read than others? (3) Do the lexical tones affect morphological processing?

What are the effects of lexical connectivity in Vietnamese?

- *Are words with large neighborhoods easier to read?*

In visual word recognition, *orthographic neighborhood* refers to groups of words that resembles one another, e.g., *make, take, sake, cake, bake* and *lake* are all neighbours (Andrews, 1997; Coltheart et al., 1977; Glushko, 1979; Yarkoni et al., 2008). According to *Coltheart's neighborhood metric*, a word's neighborhood consists of all the other words that differ by one letter compared with the original word, e.g., *cord, ford, work, ward, worm* and *woad* are all orthographic neighbors of the word *word*. In word reading and lexical decision tasks, words with many neighbors are easier to respond to and to recognize (Andrews, 1989; Balota et al., 2004; Coltheart et al., 1977; Grainger, 1990; Mulatti et al., 2006). Some researchers suggest that neighborhood effects really reflect phonological similarity rather than orthographic similarity (Mulatti et al., 2006; Yates et al., 2004), because when orthographic similarity is controlled, phonological similarity still influences lexical decision times. Using only about 6000 syllabemes to create thousands of compounds, with a number of very productive syllabemes, Vietnamese might be one of the best testbeds for this phenomena. Is word recognition in Vietnamese co-determined by similarity neighbor-

hoods? If so, this would provide evidence that neighbors become co-activated and slow comprehension.

- *Strongly connected component*[1] *effects*: In an exploratory study on lexical connectivity, Baayen (2010) observed that the distant lexical neighbors (second family size) have an inhibitory effect in visual lexical decision and word naming. An inhibitory effect was also recorded from the shortest path from head to modifier. With these findings, the study "provides a partial solution" to the issue of how individual words are indentified in a massive spreading of activation in the lexical network raised by Balota and Lorch (1986). However, as noted by Baayen (2010, p. 400–401) "replication studies for larger data sets, and different languages, will be required before the present results can be established as more than a promising window on the pros and cons of morphological connectivity in the mental lexicon". Since Vietnamese syllabemes are very productive the connections are very dense. I predict that Vietnamese compounds which are part of the strongly connected component of the compound graph will be responded to more quickly in visual lexical decision.

- *Primary and secondary family size effects:* Among the morphological relationships between words in the lexicon, the *morphological family size*, the number of morphologically related complex words in which a given word occurs as a constituent, has been shown to be predictive of responses (Schreuder and Baayen, 1997). For example, *landfall* and *landless* are family members of the word *land*. Schreuder and Baayen (1997) observed that Dutch words with larger morphological families were processed faster and more accurately than the ones with smaller families in a Dutch visual lexical decision task. The facilitatory effects of family size have been replicated for Dutch (Bertram et al., 2000a; De Jong, 2002; De Jong et al.,

[1] In a directed graph, a strongly connected component is a maximal subset of vertices in which there is a directed path from any vertex to any other.

2000; Kuperman et al., 2009), English (Baayen et al., 1997b; De Jong et al., 2002; Juhasz and Berkowitz, 2011), Chinese (Feldman and Siok, 1997), Finnish (Kuperman et al., 2008; Moscoso del Prado Martín et al., 2004) and Hebrew (Moscoso del Prado Martín et al., 2005). Interestingly, the family size effect is found to be more predictive than other lexical predictors such as word frequency, word length, morpheme frequency, bigram frequency, and orthographic neighborhood size (De Jong et al., 2000; Schreuder and Baayen, 1997), and age of acquisition (De Jong, 2002).

Morphological connectivity is observed to go beyond the primary morphological family in compound processing (Baayen, 2010). Specifically, a question arises as to whether family members of the activated primary morphological family of the target (known as *secondary family size*) influence target word processing. This measure is obtained by summing the positional family sizes of their family members over both constituents of a compound. Baayen (2010) found that the secondary family size of compounds had an inhibitory effect on response latencies in visual lexical decision and word naming data acquired from the English Lexicon Project (Balota et al., 2007); see also Mulder et al. (2014). In other words, the activation of the secondary family size in the mental lexicon does not facilitate the processing of the target word, but rather inhibits it. Baayen observes that a lexical decision apparently involves discrimination of semantically relevant and semantically irrelevant lexical activation. This thesis also investigates whether secondary family size effects are also observed for Vietnamese.

Are there polysyllabic words in Vietnamese?

Traditionally, Vietnamese philology did not have notion comparable to the "word". Vietnamese philology distinguishes between the orthographic form of the syllabeme

refered to as 'chữ' and its pronunciation refered to as 'tiếng'. *Tiếng*, 'the sound of a syllable', is the only unit of speech, and *chữ* is the written symbol for a *tiếng*. A similar distinction can also be found in Chinese philology (see Chao, 1946, for brief discussion). Thus in Vietnamese (and in Chinese also) the syllable is generally regarded as the fundamental psychological unit of speech (see e.g., Ngô, 1984; Hoosain, 1992).

The notion of 'word' is much less important in traditional Vietnamese grammar. In languages with alphabetic scripts, a word can be generally defined as a string of letters bounded by spaces. In Vietnamese, however, there are spaces between syllabemes, but there are no boundary markers for words. The question of what constitutes a word was widely debated in the 1980s (see e.g., Cao, 1985; Nguyễn, 1984; Phan, 1984).

In fact, the 'word' is a notion that is not without its problems in general linguistics. The word is conventionally defined as "the minimum free form" (Bloomfield, 1933). It is also described as having some sort of structural stability, which is required to pass the traditional insertion and affixation tests, and as being an onomasiological unit. Although this conventional approach to the 'word' fits well with Vietnamese, the understanding of the concept word in Vietnamese linguistics is diverse. To date, there are at least four main theories explaining the nature of the Vietnamese word, along with the extraordinary claim of Hockett (1944, p. 255) that "there are no words in Chinese" (and Vietnamese). These approaches can be grouped into two main camps, the monosyllabic view and the polysyllabic view.

The monosyllabic view

- The *traditional monosyllabic view* holds that the syllable is the minimum free

form and that there is no word in the Vietnamese language. This position has been defended by Chao (1946) and Trương (1970), who contend that all words in Vietnamese are monosyllabic, without exception. In this tradition, Nguyễn (1996) claims that Vietnamese has a second higher order onomasiological unit, namely, the 'phrase' (*ngữ*). Nguyễn (1985) uses the concept of *ngữ* to range over compounds, duplicatives, identification phrases, and idioms.

- The *neo-monosyllabic view* of Trần (1968), Emeneau (1951), and Nguyễn (1976a) holds that all words in Vietnamese are "one syllable long", except in cases of reduplication, and cases of compounding, where the word is polysyllabic.

- The *syllabeme view* of Stankevich and Nguyen (1976) and Nguyễn (1981) argues that the syllabeme is the basic unit and that only borrowings can be polysyllabic.

Polysyllabic view

- *The polysyllabic view* of Thompson (1963, 1965); Nguyễn (1963); Hồ (1976); Nguyễn (1976b); Đái (1978); Mai et al. (1997) among others, holds that there are morphemes that are not necessarily monosyllabic, and that there are words which are not necessarily monomorphemic. In this tradition, the 'word' is defined as the smallest meaning unit, having a stable phonetic form, serving a naming function, that can be used independently, and that recurs freely in speech to form sentences, e.g., *nhà* 'house', *người* 'person', *nếu* 'if', *thì* 'then', *sân bay* 'airport', *đen sì* 'very black', etc. The unit for creating words is the *tiếng* (syllable).

The pollysyllabic view of the word in Vietnamese is currently more popular. All of the positions discussed above simply agree with the position of Lyons (1968, p. 187–188), which states that "all words [in Chinese and Vietnamese] are invariable", in the sense that they are never inflected. This dissertation seeks to investigate the psychological status of words in Vietnamese using disyllabic compound words. Specifically, are Vietnamese compounds processed as wholes or are Vietnamese words made up of individual syllables?

Structure of the dissertation

In the sections described above, I described the main research questions related to the processing of Vietnamese compound words, including discussion of robust predictors such as frequency, dispersion, family size, and secondary family size. I also briefly presented the major characteristics of Vietnamese which are relevant to word processing. I also reviewed models of morphological processing. In the following sections, I present a brief summary for each chapter of the dissertation.

Chapter 2.

There are clear practical and theoretical motivations for constructing a Vietnamese lexical database. To begin with, there is as yet no such database for the language. Chapter 2 in this dissertation reports the construction of two Vietnamese corpora: a general corpus (including recent online newspaper articles and short stories) and a movie subtitle corpus (including recent movie subtitle translations). From these corpora I extracted a wide range of lexical variables relevant to psycholinguistic and corpus analysis. Specifically, I calculated word frequency, syllable frequency, the corresponding dispersions for words and syllables, and Inverse Document Frequency.

Furthermore, mutual information between syllables of disyllabic compounds, Hyper Analog to Language - HAL (Burgess and Lund, 1997), and Latent Semantic Analysis - LSA (Landauer and Dumais, 1997) statistics were also computed. In addition, various dispersion measures and Gries's Deviation of Proportion (Gries, 2009) were also obtained. These measures themselves and comparisons of the different predictors serve as the results of this study, further they serve as the predictors for the following studies where some of these measures are used to fit lexical processing models. With a comprehensive data set like this, we will be able to investigate the interactions between predictors for language comprehension and production, to gauge the complex dynamic system that is language.

Chapter 3.

Research on compound word processing and the role of compound constituents has long utilized the lexical decision and lexical naming tasks. The lexical decision task has recently also been used successfully in mass experiments on large numbers of items and subjects (see e.g., Keuleers et al., 2013). Chapter 3 investigates the processing of compound words using the visual lexical decision task. I predict that I will find that more frequent words are recognized faster than less frequent words as has been shown in many experiments. This frequency effect is regarded as one of the more robust effects in the literature on word recognition. If we strictly distinguish three types of morphologically complex words as mentioned above, then by examining the processing of compound words with the specific attention for the predictivity of constituents' frequencies and whole-word frequency, we can address the question of decompositional versus full form processing. Investigating the representation and processing of compound words can play an important role in the study of language mechanism: for instance, the way the mental-computational processes serve active

language comprehension and production; the interplay between storage and computation; the manner in which morphological and semantic factors impact on the nature of storage. Libben (2006, p. 3) points out that "when we study compounds, we examine the fundamental characteristics of morphology in language and the fundamentals of the human creative capacity for morphological processing and representation." Vietnamese compounding is highly productive in forming new words. The unusual writing system, in which each syllable coincides with a morpheme and is written between two spaces in a Roman writing system, places Vietnamese in a unique position in the cross-linguistic panorama of compound processing. In this language, the syllable boundary is alway clear while the word boundary has to be infered. This characteristic is not only of interest to psycholinguistics, but also to computational linguistics and natural language processing.

Two rival families of theories introduced earlier, the second model, the *prelexical theories*, also taken as models for decompositional theories (Taft and Forster, 1976), and *the theories of lexical processing*, will be tested against Vietnamese compound data. The *parallel dual route model* of Baayen et al. (1997a) will also be considered. Although these theories have different predictions, they have in common an assumption that a frequency effect observed in an experiment for head, modifier, or whole compound indicates the activation of a lexical representation for that lexical unit. The disagreement between these theories lies with the temporal order in which these representations are activated. According to the *prelexical* theories, the compound representation is activated only after its constituents have been accessed. In *supralexical* theories (e.g., Giraudo and Grainger, 2001), the constituents are accessed after the whole word is activated. On the other hand, the fourth model, *parallel dual route model*, considers the representation for compound, head and modifier are accessed in parallel, and the winning route depends on the distributional

properties of all three forms, including their frequencies.

Chapter 4.

In chapter 4, I make use of a lexical naming experiment to investigate how we proceed from visual processing to speech production. In this type of experiment, a participants was presented with printed words on a computer screen and asked to pronounce these words as quickly and accurately as possible. In this process, it is expected that orthographic and phonological representations play a prominent role. The role of semantics in lexical naming, however, has been at the center of an ongoing debate[2]. In this study, I carried out a mega-study of word naming latency and production duration in Vietnamese, again using a single-subject experimental design (see Chapter 3) with responses to 13,999 disyllabic compound words. Using lexical measures derived from two Vietnamese corpora, as reported in chapter 2, the effects of predictors such as `constituent frequency`, `compound frequency`, `family size`, `neighborhood density`, `secondary family size`, `orthographic complexity`, `word length`, and `connectivity in the compound graph` are investigated.

Chapter 5.

In chapter 5, I summarize the findings presented in Chapters 2 through 4. This chapter discusses how these findings come together and provide a clearer picture on the processing of Vietnamese compounds. I then discuss some of the challenges and limitations of the current work. This discussion is then followed by a description of proposed topics for further research.

[2]In single-route models of reading mapped orthography directly onto phonology; in recent models, Harm and Seidenberg (2004) introduced that the orthography-to-phonology mapping is mediated by semantic representations some of the time. Dual-route models of reading, on the other hand, have proposed that while reading real words requires lexico-semantic representations, non-words are processed through a direct orthography-to-phonology mapping.

References

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

Alario, F., Costa, A., and Caramazza, A. (2002). Frequency effects in noun phrase production : Implications for models of lexical access. *Language and Cognitive Processes*, 17:299–319.

Alvarez, C. J., Carreiras, M., and Perea, M. (2004). Are syllables phonological units in visual word recognition? *Language & Cognitive Processes*, 19(3):427–452.

Alvarez, C. J., Carreiras, M., and Taft, M. (2001). Syllables and morphemes: contrasting frequency effects in spanish. *Journal of experimental psychology. Learning, memory, and cognition*, 27(2):545–555.

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:802–814.

Andrews, S. (1997). The effects of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychological Bulletin & Review*, 4:439–461.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Baayen, R. H. (2010). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In Olsen, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997a). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 36:94–117.

Baayen, R. H., Lieber, R., and Schreuder, R. (1997b). The morphological complexity of simplex nouns. *Linguistics*, 35:861–877.

Baayen, R. H., McQueen, J., Dijkstra, T., and Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In Baayen, R. H. and Schreuder, R., editors, *Morphological structure in language processing*, pages 355–390. Mouton de Gruyter, Berlin.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133:283–316.

Balota, D. A. and Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(3):336–345.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Bannard, C., Lieven, E., and Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41):17284–17289.

Bannard, C. and Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, 19(3):241–248.

Bauer, L. (1983). *English word formation.* Cambridge University Press, New York.

Bertram, R., Baayen, R., and Schreuder, R. (2000a). Effects of family size for complex words. *Journal of Memory and Language*, 42:390–405.

Bertram, R., Schreuder, R., and Baayen, R. H. (2000b). The balance of storage and computation in morphological processing: the role of Word Formation Type, Affixal Homonymy, and Productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:419–511.

Bloomfield, L. (1933). *Language.* University of Chicago Press, Chicago.

Bod, R. (1998). *Beyond grammar: An experience-based theory of language.* CSLI publications, Stanford, CA.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Burgess, C. and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2-3):177–210.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Cao, X. H. (1985). Về cương vị ngôn ngữ học của tiếng [On the status of syllable]. *Ngôn ngữ [Language]*, 2:30 – 40.

Carreiras, M., Alvarez, C. J., and De Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language*, 32(6):766–780.

Chao, Y. (1946). The logical structure of Chinese words. *Language*, 22:4–13.

Colé, P., Magnan, A., and Grainger, J. (1999). Syllable-sized units in visual word recognition: Evidence from skilled and beginning readers of French. *Applied Psycholinguistics*, 20(04):507–532.

Coltheart, M., Davelaar, E., Jonasson, J. T., and Besner, D. (1977). Access to the internal lexicon. In Dornick, S., editor, *Attention and performance*, volume VI, pages 535–556. Erlbaum, Hillsdale, New Jersey.

Conklin, K. and Schmitt, N. (2007). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1):72–89.

Conrad, M., Carreiras, M., and Jacobs, A. (2008). Contrasting effects of token and type syllable frequency in lexical decision. *Language & Cognitive Processes*, 23(2):296–326.

Conrad, M., Grainger, J., and Jacobs, A. M. (2007). Phonology as the source of syllable frequency effects in visual word recognition: evidence from French. *Memory & Cognition*, 35(5):974–983.

Conrad, M. and Jacobs, A. (2004). Replicating syllable frequency effects in spanish in german: One more challenge to computational models of visual word recognition. *Language and Cognitive Processes*, 19(3):369–390.

Conrad, M., Stenneken, P., and Jacobs, A. M. (2006). Associated or dissociated effects of syllable frequency in lexical decision and naming. *Psychonomic bulletin review*, 13(2):339–345.

Crepaldi, D., Rastle, K., Coltheart, M., and Nickels, L. (2010). 'Fell' primes 'fall', but does 'bell' prime 'ball'? Masked priming with irregularly-inflected primes. *Journal of Memory and Language*, 63(1):83–99.

De Jong, N. H. (2002). *Morphological families in the mental lexicon*. PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., and Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language*, 81:555–567.

De Jong, N. H., Schreuder, R., and Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, 15:329–365.

Diependaele, K., Sandra, D., and Grainger, J. (2009). Semantic transparency and masked morphological priming: The case of prefixed words. *Memory & Cognition*, 37(6):895–908.

Domínguez, A., Cuetos, F., and Segui, J. (2000). Morphological processing in word recognition: A review with particular reference to Spanish data. *Psicológica*, 21:375–401.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 53:810–842.

Đái, X. N. (1978). *Hoạt động của từ tiếng Việt [The usage of Vietnamese words]*. Nhà xuất bản Khoa học Xã hội, Hà Nội.

Emeneau, M. (1951). *Studies in Vietnamese (Annamese) grammar*. University of California at Berkeley, Berkeley.

Feldman, L. B., O'Connor, P. A., and del Prado Martin, F. (2009). Early morphological processing is morpho-semantic and not simply morpho-orthographic: Evidence from the masked priming paradigm. *Psychonomic Bulletin & Review*, 16(4):684–691.

Feldman, L. B. and Siok, W. W. T. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23:778–781.

Fidelholz, J. (1975). Word frequency and vowel reduction in English. In *Papers from the 75th meeting of the Chicago Linguistics Society*, pages 200–213. University of Chicago.

Fiorentino, R. and Fund-Reznicek, E. (2009). Masked morphological priming of compound constituents. *The Mental Lexicon*, 4(2):159–193.

Forster, K. I. (1989). Basic issues in lexical processing. In Marslen-Wilson, W., editor, *Lexical representation and process*, pages 75–107. MIT Press, Cambridge, MA.

Frank, S., Bod, R., and Christiansen, M. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531.

Gagné, C. L. and Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, 60(1):20 – 35.

Giraudo, H. and Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin and Review*, 8:127–131.

Giraudo, H. and Grainger, J. (2003). On the role of derivational affixes in recognizing complex words: Evidence from masked priming. In Baayen, R. H. and Schreuder, R., editors, *Morphological Structure in Language Processing*, pages 211–234. Mouton, Berlin.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5(4):674–91.

Goldberg, A. (2005). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Goslin, J., Grainger, J., and Holcomb, P. J. (2006). Syllable frequency effects in french visual word recognition: An ERP study. *Brain Research*, 1115(1):121–134.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, 29:228–244.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437.

Gries, S. T. (2009). Dispersions and adjusted frequencies in corpora: Further explorations. *Language and Computers*, 71(1):197–212.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111:662–720.

Hay, J., Pierrehumbert, J., and Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In Local, J., Ogden, R., and Temple, R., editors, *Phonetic Interpretation: Papers in Laboratory Phonology VI*, pages 58–74. Cambridge University Press, Cambridge.

Hockett, C. (1944). Reviews. *Language*, 23:252–255.

Hoosain, R. (1992). Psychological reality of the word in Chinese. In Hsuan-Chih Chen and Ovid J.L. Tzeng, editor, *Language processing in Chinese*, volume 90, pages 111–130. North-Holland, Amsterdam.

Hồ, L. (1976). *Vấn đề cấu tạo từ tiếng Việt hiện đại [Word formation in modern Vietnamese]*. Nhà xuất bản Khoa học Xã hội, Hà Nội.

Janssen, N. and Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE*, 7(3):e33202.

Jones, M., Kintsch, W., and Mewhort, D. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55:534–552.

Juhasz, B. J. and Berkowitz, R. N. (2011). Effects of morphological families on english compound word recognition: A multitask investigation. *Language and Cognitive Processes*, 26(4-6):653–682.

Juilland, A., Brodin, D., and Davidovitch, C. (1970). *Frequency dictionary of French words*. Romance languages and their structures. Hague, Paris.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. L. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 229–254. Benjamins, Amsterdam.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behaviour Research Methods*, 42(3):643–650.

Keuleers, E., Mandera, P., and Brysbaert, M. (2013). Lexical decision with the masses. In *Annual meeting of the Belgian Association for Psychological Sciences*.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23:1089–1132.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, 35:876–895.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. Academic Press, New York.

Libben, G. (2006). Why study compound processing? An overview of the issues. In Libben, G. and Jarema, G., editors, *The representation and processing of compound words*, pages 1–22. Oxford University Press, New York.

Lieven, E., Salomo, D., and Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3).

Longtin, C., Segui, J., and Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, 18(3):313–334.

Longtin, C. M. and Meunier, F. (2005). Morphological decomposition in early visual word processing. *Journal of Memory and Language*, 53(1):26–41.

Lyons, J. (1968). *Introduction to theoretical linguistics.* Cambridge University Press, Cambridge.

Mai, N. C., Vũ, D. N., and Hoàng, T. P. (1997). *Cơ sở ngôn ngữ học và tiếng Việt [Fundamentals of linguistics and Vietnamese].* Giáo Dục, Hà Nội.

Marelli, M. and Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, 66(4):644–664.

Mathey, S. and Zagar, D. (2002). Lexical similarity in visual word recognition: The effect of syllabic neighborhood in French. *Current Psychology Letters: Behaviour, Brain & Cognition*, pages 107–121.

McCormick, S. F., Rastle, K., and Davis, M. H. (2008). Is there a 'fete' in 'fetish'? Effects of orthographic opacity on morpho-orthographic segmentation in visual word recognition. *Journal of Memory and Language*, 58(2):307–326.

McCormick, S. F., Rastle, K., and Davis, M. H. (2009). Adore-able not adorable? Orthographic underspecification studied with masked repetition priming. *European Journal of Cognitive Psychology*, 21(6):813–836.

McDonald, S. A. and Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3):295–323.

McQueen, J. M. and Cutler, A. (1998). Morphology in word recognition. In Zwicky, A. M. and Spencer, A., editors, *The Handbook of Morphology.* Basil Blackwell, Oxford.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The

case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., and Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch. *Journal of Memory and Language*, 53(4):496–512.

Moscoso del Prado Martín, F. and Sahlgren, M. (2002). An integration of Vector-Based Semantic Analysis and Simple Recurrent Networks for the automatic acquisition of lexical representations from unlabeled corpora. In Lenci, A., Montemagni, S., and Pirrelli, V., editors, *Proceedings of the LREC'2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Paris. European Linguistic Resources Association.

Mulatti, C., Reynolds, M. G., and Besner, D. (2006). Neighborhood effects in reading aloud: new findings and new challenges for computational models. *Journal of experimental psychology. Human perception and performance*, 32(4):799–810.

Mulder, K., Dijkstra, T., Schreuder, R., and Baayen, H. R. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, 72:59 – 84.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677.

Ngô, T. N. (1984). The syllabeme and patterns of word formation in Vietnamese. New York University dissertation.

Nguyễn, K. T. (1963). *Nghiên cứu về ngữ pháp tiếng Việt [Studies in Vietnamese grammar]*. Nhà xuất bản Khoa học, Hà Nội.

Nguyễn, P. P. (1976a). Vấn đề láy từ trong tiếng việt [Reduplication in Vietnamese]. *Tập san Khoa học Xã hội*, 3:73–81.

Nguyễn, T. C. (1981). *Ngữ pháp tiếng Việt: Tiếng – Từ ghép – Đoản ngữ [A Vietnamese grammar: Syllable – Compound – Phrase]*. Nhà xuất bản Đại học và Trung học chuyên nghiệp, Hà Nội.

Nguyễn, T. G. (1984). Về mối quan hệ giữa "từ" và "tiếng" trong Việt ngữ [On the relation between "word" and "syllable" in Vietnamese]. *Ngôn ngữ [Language]*, 3:41 – 57.

Nguyễn, T. G. (1985). *Từ vựng học tiếng Việt [Vietnamese lexicology]*. Đại học và Trung học chuyên nghiệp, Hà Nội.

Nguyễn, T. G. (1996). *Từ và nhận diện từ tiếng Việt [Words and recognizing words in Vietnamese]*. Giáo Dục, Hà Nội.

Nguyễn, V. T. (1976b). *Từ và vốn từ trong tiếng Việt hiện đại [Words and word stocks in modern Vietnamese]*. Nhà xuất bản Đại học và Trung học chuyên nghiệp, Hà Nội.

Perea, M. and Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):767–779.

Pham, H. and Baayen, H. R. (2013). Semantic relations and compound transparency: A regression study in CARIN theory. *Psihologija*, 46(4):455–478.

Phan, T. (1984). Hình vị và âm tiết [Morpheme and syllable]. *Ngôn ngữ [Language]*, 2:29 – 40.

Rastle, K. and Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, 23(7):942–971.

Rastle, K., Davis, M. H., and New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11:1090–1098.

Rueckl, J. and Aicher, K. (2008). Are CORNER and BROTHER morphologically complex? Not in the long term. *Language & Cognitive Processes*, 23(7):972–1001.

Schreuder, R. and Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, 37(1):118–139.

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In Coltheart, M., editor, *Attention and Performance XII*, pages 245–264. Lawrence Erlbaum Associates, Hove.

Seidenberg, M. S. and Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, 4(9):353–361.

Siyanova-Chanturia, A., Conklin, K., and Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

Spencer, A. (1991). *Morphological theory: An introduction to word structure in generative grammar*. Cambridge University Press, Cambridge.

Stankevich, N. and Nguyen, T. C. (1976). The problem of the word in its relationship to the grammatical system in Vietnamese. *Vietnamese Studies*, 40:218–243.

Taft, M. (1991). *Reading and the mental lexicon*. Psychology Press, Hove.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, 9(3):271–294.

Taft, M. (2003). Morphological representation as a correlation between form and meaning. In Assink, E. and Sandra, D., editors, *Reading complex words*, pages 113–137. Amsterdam: Kluwer, New York.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57A:745–765.

Taft, M. and Ardasinski, S. (2006). Obligatory decomposition in reading prefixed words. *The Mental Lexicon*, 1(2):183–199.

Taft, M. and Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, 15:607–620.

Thompson, L. (1963). The problem of the word in Vietnamese. *Word*, 19:39–52.

Thompson, L. (1965). *A Vietnamese grammar*. The University of Washington Press, Seattle.

Trần, T. K. (1968). *Việt Nam văn phạm trung học [A Vietnamese grammar for High School]*. Tân Việt, Saigon, 8th edition.

Tremblay, A. and Baayen, R. H. (2010). Holistic Processing of Regular Four-word Sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In Wood, D., editor, *Perspectives on formulaic language: Acquisition and communication*, pages 151–173. The Continuum International Publishing Group, London.

Tremblay, A., Derwing, B., Libben, G., and Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(2):569–613.

Tremblay, A. and Tucker, B. V. (2011). The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon*, 6(2):302–324.

Trương, V. T. (1970). *Structure de la langue Viêtnamienne*. Imprimerie Nationale: P. Geuthner, Paris.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Yates, M., Locker, Lawrence, J., and Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, 11(3):452–457.

CHAPTER 2

Constructing two Vietnamese corpora and building a lexical database

This chapter is a manuscript for publication as Hien Pham, Benjamin Tucker, and Harald Baayen. (2014) "Constructing two Vietnamese corpora and building a lexical database." *Manuscript.*

Abstract

The present paper reports the creation of two corpora of contemporary Vietnamese. This paper describes the construction of two equally sized Vietnamese corpora (a corpus from Vietnamese film subtitles, SUBTLEX-VIET, and a general corpus of varieties of online newspapers and stories, GENLEX-VIET). The corpora are open to further extension: more material can be added as the opportunity arises. We have attempted to document every step of the construction and extraction of linguistic information from the language corpora to provide a potential road map for others

who would like to create similar corpora. The resultant corpora are available in three versions: plain text, tokenized, and POS tagged. In the second half of the paper, the construction of a lexical database derived from the corpora is described. The database includes measures such as *frequency of occurrence, dispersion, Mutual Information, Inverse Document Frequency* , as well as vector space measures based on *Latent Semantic Analysis* and *Hyperspace Analogue to Language*. We conclude with reporting a check of the corpora and the lexical predictors, including a comparison of the predictors and a validation using psycholinguistic data from visual lexical decision experiments.

Keywords: written corpus, film subtitle corpus, frequency, dispersion, LSA, HAL, Vietnamese, validation

Introduction

Vietnamese is an understudied language and to the best of our knowledge, there are only a few small scale Vietnamese corpora such as the Corpora of Vietnamese Texts (Pham et al., 2008) with approximate 1 million words (precisely this is syllable counts), the Vietnamese Corpus (Trung tâm từ điển học, 1998) with about 80 million syllables, the Vietnamese Text Corpus (Southeast Asian Languages Library, 2009) with about 50 million syllables. These corpora, however, are only available to the public for searching words in a concordance view. Importantly, it is not possible with these corpora to calculate frequency and dispersion measures, nor is it possible to estimate semantic similarity of words using word co-occurrence across documents (LSA) or within text windows (HAL). This renders these corpora insufficient for most types of linguistic research. Corpora and corpus-derived lexical resources that are fully accessible and manipulable is not only useful for investigations of the lexicon but also for studying syntax and discourse on word sequences. The current study addresses this gap by creating two corpora of Vietnamese, calculating lexical statistical predictors, and validating the corpora with psycholinguistic data.

To date, there are several works on Vietnamese word frequency. However, they are either out-of-date or are based strictly on monomorphemic/monosyllabic frequencies, disregarding compound frequencies. The first Vietnamese frequency dictionary (not available in electronic version) is based on a collection of texts with 524,500 words by Nguyễn and Lê (1980). It consists of newspaper articles, poetry, theatrical works, children's literature, and Hồ Chí Minh's writings from 1956 to 1972. This dictionary lists separate lexical frequencies under such categories as nouns, verbs, adjectives, numbers, connecting words, proper nouns, and so on. Pham et al. (2008) constructed an electronic corpus of 1,063,912 words, drawn from newspapers

and children's literature (newspaper articles from 2006 and children's picture books from 1976 to 2006). Although this corpus contains double the number of words in the Nguyễn and Lê (1980) corpus, the 'words' are again syllables and do not take into account that often syllables combine into multi-syllabic compound words. This characterization of words in Vietnamese leads to a confound of syllable frequency and word frequency. The present paper reports on the construction of two new corpora of Vietnamese: a corpus of newspaper articles, novels and short stories, and a corpus of film subtitles.

The present paper also describes the calculation of a wide range of lexical statistical predictors. Lexical statistical predictors such as frequency and dispersion measures are validated with lexical decision response times as suggested by Keuleers et al. (2010), using the lexical decision data reported in Pham and Baayen (2014).

This paper is organized as follows: (1) brief background, (2) construction of the corpora, (3) calculation of lexical statistical predictors, (4) psycholinguistic validation of the corpora, and (5) discussion of the findings. In what follows, we briefly describe some background information on the *characteristics of Vietnamese*, *Subtitle corpora* and the *Lexical predictors* that will be considered in the current study.

Vietnamese. Vietnamese (tiếng Việt), the official language of Vietnam, is a tone language spoken by approximately 90 million people in Vietnam (based on figures in 2013 of the General Statistics Office of Vietnam) and about 4 million speakers living abroad (in places like the U.S.A., France, Australia, Canada, Germany, the Netherlands). It belongs to the Việt-Mường sub-branch of the Vietic branch of the Mon-Khmer family, which is itself a part of the Austro-Asiatic family. In this isolating language, the syllabeme is a bi-functional unit which is a contiguous string of letters that forms a syllable or morpheme in Vietnamese (all syllables are monomor-

phemic and all morphemes are monosyllabic). In written Vietnamese, syllabemes are separated by spaces. Thus, Vietnamese words may consist of one syllabeme (e.g., *cây* 'tree', *cơm* 'rice', *mắt* 'eye') or multiple syllabemes (e.g., *cơm đen* 'opium', *tư tưởng* 'ideology').

Historically, Vietnamese has made use of four different writing systems: `Khoa đẩu văn` were the first symbols to be used to represent *Vietnamese* that existed thousands of years ago during the Hồng Bàng (Hùng Vương) dynasty (Đỗ, 2012). They can be found on the Đông Sơn Bronze Drums and in inscriptions. According to Chinese history, during the Yao dynasty of China, a Vietnamese King's envoy offered a sacred turtle (Thần Kim Quy) on which there were some inscriptions using the `Khoa đẩu` script, describing the creation of Earth and Heaven. `Chữ Nho` was the second script, which used Chinese written symbols. It served as the medium of education and official communication before A.D. 939. `Chữ Nôm` is the third script. It is largely based on the Chinese characters used by Buddhist monks and Confucian scholars from the 11th century to record eight-line stanzas or long narratives in native verse. `Nôm` writings prospered under the Trần dynasty (1225–1400). `Chữ Quốc ngữ` is the currently used script. It is based on Roman letters introduced by Catholic missionaries from Portugal, France, Spain and Italy since the 16$^{\text{th}}$ century. This script enables Vietnamese speakers to learn how to read and write within a few months and currently serves as the official orthography nation-wide (Nguyễn, 1997).

Subtitle corpora. Most written corpora from Vietnamese, and most other languages, come from newspaper texts and other types of literature. Recently, a number of corpora have been built based on film and television subtitles. One of the first corpora using subtitles (New et al., 2007) was created using 9,500 film and television subtitle documents which generated a corpus of approximately 50 million words. The

authors validated the subtitle frequencies by comparing them to those of a French written corpus - *Lexique 2* (New et al., 2004) and to those of a spoken language corpus, *Corpus de Référence du Français Parlé* (Delic, 2004). A striking finding reported by New et al. (2007) is that word frequencies from the subtitle corpus were as good as those from the written language corpus and better than the frequency from the *Corpus de Référence du Français Parlé* corpus in predicting reading behaviour. These results have encouraged other researchers to build similar subtitle corpora in different languages: American-English with 51 million words (Brysbaert and New, 2009), Chinese with 33.5 million words (Cai and Brysbaert, 2010), Dutch with 43.7 million words (Keuleers et al., 2010), Greek with 27.7 million words (Dimitropoulou et al., 2010), and Spanish with 41 million words (Cuetos et al., 2011).

Lexical predictors. In language comprehension and processing, words are key building blocks in computational modeling (McClelland and Rumelhart, 1981; Baayen et al., 2011), cognitive neuroscience (e.g., Petersen et al., 1988, 1989), and psycholinguistics (e.g., Pinker, 1999). Due to the fundamental status of words, it is widely recognized that researchers need to collect an enormous wealth of information about individual lexical items to investigate their complexity and interaction. Many variables have been identified that are important such as: *word frequency*, *dispersion*, *HAL*, *LSA*, *letter length*, *phoneme length*, *syllable length*, *number of morphemes*, *syntactic class*, *orthographic neighborhood*, *orthographic complexity*, *phonological neighborhood*, *frequency of orthographic*, *phonological neighborhoods*, *spelling-to-sound consistency*, *number of meanings*, *familiarity*, and *age of acquisition*, among many others. The present paper computes these corpus-based lexical predictors including collocation measures (e.g., mutual information - MI, MI3, z-score, and log-likelihood). We describe how each of these predictors was calculated in the section on lexical predictors

below.

Corpus Construction

The GENLEX-VIET

The first corpus is a general corpus of Vietnamese which we will call the GENLEX-VIET corpus.

Materials. Data in the GENLEX-VIET corpus comprise news texts, e.g., newspaper articles and short stories, published from 2006 to 2010. These texts were extracted from top ranked online newspapers, such as www.vietnamnet.vn, www.vnexpress.net, www.nhandan.com.vn, www.suutap.com, evan.com.vn, dactrung.net, vnthuquan.net. Between December 10 – 24, 2010, the data was collected using the web crawling HTTrack tool. The program processed 82,263,474 words coming from 164,526 HTML files.

Processing. When each HTML file was encountered, the actual text and accompanying illustrations were extracted by removing navigation panels, banners, footers, tables of links, etc. The resulting document is then stored as a plain text file. All the resulting text files are then used as the input of the further processing phases (see the "All processing" section below), word segmentation and annotation.

The SUBTLEX-VIET

The second corpus, which is made up of film and TV subtitles has been dubbed the SUBTLEX-VIET corpus.

Materials. Many subtitles are freely available on the Internet. They are basically text files with time stamps and occasional metadata about the translators. The translation is usually produced and double checked by highly proficient volunteers working as members of a volunteer group for subtitle websites. We obtained permission to download all the Vietnamese subtitle files from one of the largest websites in Vietnam and two other large websites providing subtitles in many languages. Between December 10–24, 2010, software written specifically to download and process these files (Subscene Downloader and Wget) were utilized to process subtitles from the following internet sites: http://subscene.com, www.phudetiengviet.com, www.opensubtitles.org. Disregarding duplicates, the program processed 79,757,504 words coming from 13,349 subtitle documents, of which the majority (12,668) were translated subtitles of English films and television series (we used the Internet Movie Database <www.imdb.com> to determine the countries of origin)[1].

Processing. In what follows, we sketch the main steps in the editing phase. First, all the zip files were extracted into the original format. Second, all the files were converted into TXT files. Third, all the time-frame markers and the names of the translator, proofreader, director, actors, email addresses, extra information such as advertisements, etc. that were not part of the film contents were removed. Finally, to avoid the repetition of subtitle files of the same film and to check files for content and technical errors (e.g., bad translation or wrong codepage), all file names were checked both automatically and manually by the first author.

The number of words in this corpus (roughly 80 million) on which these word frequencies were based is slightly larger than what has been previously gathered (New

[1] We will return to the possible disadvantages of using translated subtitles below

et al., 2007; Brysbaert and New, 2009; Keuleers et al., 2010; Cai and Brysbaert, 2010; Dimitropoulou et al., 2010). According to Brysbaert & New's calculation, it is well above the required 16 million words, the size that is sufficient for validating with psycholinguistic response latency data (Brysbaert and New, 2009) and is large enough to allow estimates per million with 1-digit precision.

All processing

The files were first standardized so that they had the same coding and form. Since Vietnamese texts can be encoded in many different encoding types, such as VNI-Times, VnTimes, UTF-16LE, UTF-8, etc., all the text files were converted into UTF-8 encoding so that they would work well with corpus processing programs, e.g., tokenizer and tagger programs. Standardization was applied for some typos and misspellings as well as incorrectly placed tones, e.g., *hỏa*, *lòe*, and *túy* changed to *hoả*, *loè*, and *tuý* respectively. Regular expressions were used to find and correct these types of errors.

As noted above, Vietnamese is a typical isolating language and as a result there are many compound words. We used the vnTokenizer program (Lê et al., 2008) to segment words in the corpus. We also used the vnTagger program (Lê et al., 2010) to tag the corpus for part of speech. Since these compounds almost always contain words themselves (like the English compound "shower curtain"), all the multi-syllabeme words were connected by underscore (_) marks. The results of these two processes were stored separately resulting in three versions of the corpus, the plain untokenized version, the tokenized version, and the syntactic tagged version. Following other corpora, including CELEX (Baayen et al., 1995), the French subtitle corpus and the Dutch subtitle corpus, we wanted to provide information about the various

grammatical functions of words in addition to the frequencies of the word forms themselves.

The resulting corpora offer a wealth of potential data for analysis by corpus linguists. However, quantitative studies require more information from the corpora than just concordancing lines in KWIC - Key Word In Context - format provided by most current corpus tools. In what follows, we present the construction of a lexical database reporting a large number of lexical predictors that we calculated from the corpora.

Lexical Predictors

Materials

We extracted frequency measures from the untokenized-version of the corpora and from the tokenized-version. The untokenized corpora are used for calculating syllable frequencies; the tokenized-version corpus for calculating word and phrase frequencies. In the variable calculation section, I present the formulae for calculating the relevant measures, such as all kinds of collocation measures. The raw corpora were processed with corpus tools and merged into sections of the database and stored in R (R Core Team, 2013) dataframe format. This format is preferable for further data manipulation and data analysis. All calculations were performed for both the GENLEX-VIET and SUBTLEX-VIET corpora, examples in the description of predictor calculation are given using the SUBTLEX-VIET corpus.

Variable calculation

Frequency extraction

Frequency. Word frequency is one of the most well-studied corpus variables. It is predictive for many aspect of human behavior in language processing (Hasher and Zacks, 1984). Word frequency is a simple count of the number of instances of a word in a corpus, i.e., in frequency of occurrence within that corpus. Let N denote the sample size in word tokens (the number of individual words in the corpus) and let i denote the i-th unit (w_i) in a list of word types (the number of types in a word frequency list is the number of unique word forms, rather than the total number of tokens in the corpus). We specify word frequency as $f(i, N)$, i.e., frequency of w_i in a sample of N tokens (see e.g., Baayen, 2001, for further discussion). One of the most frequently-used statistics in corpus linguistics is the frequency list of occurrence of linguistic items, such as syllables, words, or n-grams. A frequency list is a sorted list of words with the total number of occurrences of a repeating unit in the corpus (see e.g., Baayen, 2001, for further discussion).

Calculation of frequency. Due to the isolating character of Vietnamese, we calculated three measures of frequency. The first one presents the frequency of syllabemes. The second frequency measure counts lexicographical word frequencies (both monosyllabic and multisyllabic words). The third measure reports the frequencies of words together with their part of speech in the sentences in which they appeared. The resultant three files were saved in convenient formats, e.g., CSV and RDA, for both text editor programs and statistical programs. For the purposes of our own analyses, all the files were stored in RDA format encoded in UTF-8 for manipulation and analysis in the R (R Core Team, 2013) statistical programming language.

The first file (SUBTLEX-VIET-SYLL.rda) includes the syllabemes and their corresponding frequencies. There were 6176 different syllabemes in the corpus. We calculated the frequencies of each syllabeme based on the total count (SyllFreq) and the total number of subtitles in which a syllabeme appeared (SyllDisp).

Table 2.1: Layout of the SUBTLEX-VIET-SYLL file.

Syllabeme	HPCo	SyllFreq	SyllDisp	SyllDisp%	SyllPerMil	Log10SyllFreq	Log10SyllDisp
thuỷ	78	13177	4001	29.91	166.80	4.12	3.60
thuỵ	2	1100	618	4.62	13.92	3.04	2.79
thuyên	3	676	433	3.24	8.56	2.83	2.64
thuyền	28	17282	3492	26.10	218.76	4.24	3.54
thuyết	59	12346	5699	42.60	156.28	4.09	3.76
ti	27	1629	1056	7.89	20.62	3.21	3.02

Table 2.1 illustrates the information available for a sample of six syllabemes. The order of the items in the file was based on the frequency of the syllabemes from high frequency to low frequency.

Table 2.2: Layout of the SUBTLEX-VIET-WF file.

Word	WordFreq	WordDisp	WordDisp%	Log10SubtFreq	Log10SubtDisp	WFPerMillion
á à	87	82	0.61	1.94	1.92	1.10
gian ác	46	38	0.28	1.67	1.59	0.58
tội ác	2422	1568	11.57	3.38	3.20	30.66
hiểm ác	64	61	0.45	1.81	1.79	0.81
bạc ác	2	2	0.01	0.48	0.48	0.03
quái ác	40	40	0.30	1.61	1.61	0.51

The column names of Table 2.1 and Table 2.2 can be interpreted as follows:

- `Syllabeme`: the syllabeme itself, e.g., *hoa, tân, tuyến, thủ, công, etc.*.
- `Word`: the word forms, e.g., *á à, gian ác, quân tử, etc.*.

- `HPCo`: the total number of times the syllabeme is a headword or part of a headword in the Vietnamese dictionary (Hoàng, 2000).
- `SyllFreq`: is the total number of times the syllabeme occurred in the corpus
- `WordFreq`: is the total number of times the word occurred in the corpus
- `SyllPerMil`: the syllabeme frequency per million with 4-digit precision.
- `Log10SyllFreq`: the value based on $\log_{10}$(SyllFreq+1) with 4-digit precision. Calculating the log frequency based on the raw frequencies is straightforward. Following Keuleers et al., we give words that are not in the corpus a value of 0 (see Keuleers et al., 2010, for further discussion).
- `Log10SubtFreq`: the value based on $\log_{10}$(WordFreq+1) with 4-digit precision.
- `SyllDisp`: the number of films/documents in which the syllabeme occurred
- `SyllDisp%`: the percentage of films/documents in which the syllabeme occurred
- `Log10SyllDisp`: the value of $\log_{10}$(LetDisp+1) with 4 digit precision.
- `Log10SubtDisp`: the value of $\log_{10}$(WordDisp+1) with 4 digit precision.
- `WFPerMillion`: the frequency per million with 4 digit precision.

Similarly, the second file for word frequencies (SUBTLEX-VIET-WF.rda) contains the word form frequencies and dispersion information. In total, our corpus included 244,648 distinct word types. Table 2.2 shows the layout of the information, which is essentially the same as that for the syllabeme frequencies.

Plot rank vs. frequency. Eighty years ago, Zipf (1935) observed the distribution of English word frequencies and found that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Zipf's findings can be expressed in Formula 2.1 where C is a constant:

$$frequency = \frac{C}{rank} \tag{2.1}$$

Zipf's Law states that the most frequent word will occur approximately twice as often as the second most frequent word, which occurs two times as often as the fourth most frequent word, and so forth. Another way to state Zipf's Law is that the product of mutiplying the frequency of each word with its rank is a value that is more or less constant throughout the whole corpus, as illustrated in Formula 2.2 below.

$$constant \approx frequency \times rank \tag{2.2}$$

Another way to look at the data is through plotting the rank versus frequency on a graph sorted by rank, with the most frequently appearing word first as shown in Figure 2.1. The Zipf curves are plotted on linear scales and a logarithm-to-the-base-2 diagrams with a slope of -1. Zipf's Law characterizes the use of words in natural language, such as Vietnamese, so a language typically has a few words (e.g., *có* 'have', *của* 'belong, of', *chó* 'dog', *mắt* 'eye', etc) that are used very often, and a lot of words (e.g., *xoải* '(bird) stretch', *quéo* 'bachang mango', etc.) that are very rarely used. As can be seen from Figure 2.1, the shape of the Zipf curve plotted on linear scales tends to hug the axes (the left panel). Whereas the shape of the Zipf curve plotted on log scales appears as a straight line with a slope of -1 (the right panel).

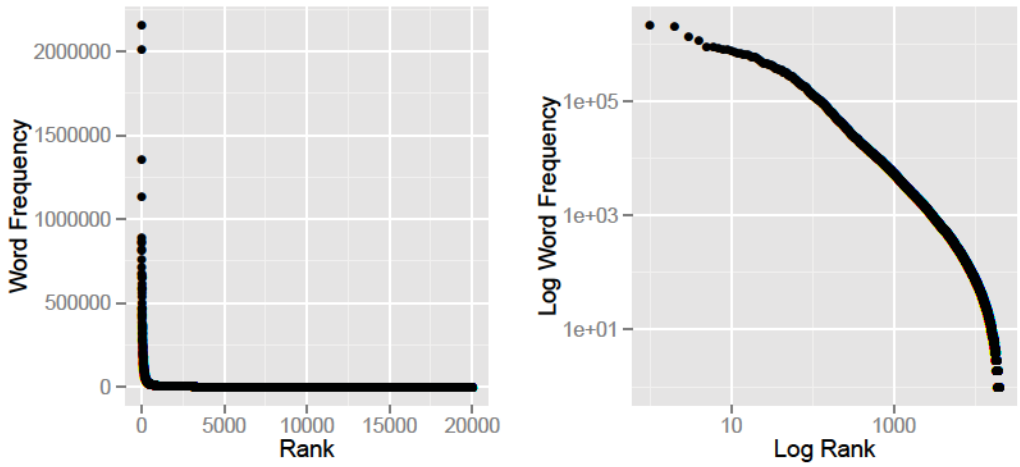

Figure 2.1: Frequency versus rank graphs of the SUBTLEX-VIET corpus. The left panel plots the data ordered by rank on linear scales. The right panel plots the data ordered by rank on logarithmic scales.

Dispersion measures. Albeit introduced into psycholinguistics more recently, dispersion has been documented as an important predictor (see e.g., Adelman et al., 2006, for more details). The dispersion measure gauges to what extent words are used across corpora. In other words, dispersion (also known as *contextual diversity* (see e.g., Adelman et al., 2006)) is the number of texts in a corpus in which a word occurs. This measure was also confirmed by Brysbaert and New (2009) for subtitles in which dispersion accounted for 1% – 3% more of the variance in lexical decision performance than did word frequencies. McDonald and Shillcock (2001) and Adelman et al. (2006) have documented that the number of times a word occurs in a corpus is less informative than the number of documents in which the word occurs. The idea of a dispersion statistic was proposed over forty years ago by Juilland, one of the pioneers of corpus linguistics (Juilland et al., 1970). This measure has been reviewed and extended in more recent studies by Gries (2008, 2009) working in the corpus linguistic paradigm. Gries (2008, 2009) reviewed many dispersion measures

and adjusted frequency measures and proposed an alternative measure, *Deviation of Proportion*. This study emulated the mentioned methods and made use of the relevant open source tools offered by Gries (2008)[2] to compute frequency, types of dispersions, as well as the Deviation of Proportion.

Collocation extraction

Besides the calculation of frequency and dispersion from corpora, the frequency of co-occurrence of two or more linguistic variables, such as words can also be calculated. The concept of collocation indicates that words frequently co-occur with other words and the combination constitutes larger constituents within a context, e.g., an utterance or a sentence (Cantos Gómez, 2013, 196). Assuming a random distribution of forms, collocation is regarded as a probabilistic phenomenon comprising statistically significant collocations of actual pattern of occurrences. In collocation analysis, a statistically significant difference can be taken as having evidence that the presence of a word in a context affects the occurrence of another word in some manner. The following subsections describe the calculation of different collocational predictors.

Mutual information (MI). One of the core statistical measures in co-occurrence statistics in corpus linguistics is *mutual information* (*MI*), which is well established in information theory as expressed in Formula 2.3 (see Cantos Gómez, 2013; Church and Hanks, 1990; Oakes, 1998; Ooi, 1998, for detailed discussion).

$$MI = log_2 \frac{p(w_1, w_2)}{p(w_1) \times p(w_2)} \tag{2.3}$$

[2] http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/_dispersion1.r

Specifically, MI compares the probability of observing the words w_1 and w_2 jointly, $p(w_1, w_2)$, with the probabilities of observing w_1 and w_2 indenpendently (by chance), $p(w_1) \times p(w_2)$. If there is legitimate association between w_1 and w_2, then the joint probability $p(w_1, w_2)$ will be much larger than chance $p(w_1) \times p(w_2)$, consequently the MI score will be greater than 0. If there is no relationship between w_1 and w_2, then $p(w_1, w_2)$ and $p(w_1) \times p(w_2)$ will be equal, and then MI will be approxiate to 0. If w_1 and w_2 are in complimentary distribution, then the MI score will be negative.

MI3. Cantos Gómez (2013, 207) observed that "the more a word co-occurs with a node word, the smaller its *MI* score is." To give more weight to frequent events, the power of three has been applied to the numerator of the *MI* equation, as shown in Formula 2.4. This has been reported as "the most effective coefficient" for extracting collocation (Oakes, 1998; Cantos Gómez, 2013).

$$MI3 = log_2 \frac{(p(w_1, w_2))^3}{p(w_1) \times p(w_2)} \tag{2.4}$$

z-score. The *z*-score was introduced by Berry-Rogghe (1973) to compare the observed frequency between a lexical node and its collocate to the expected frequency and to calculate the difference between these values. Given that we have the total number of words of the whole corpus (T_c); the total occurences of the specific collocation within the whole text or corpus (F_c); the total occurrences of the certain collocation within the specific context span (O); the total number of words of the specific context span (T_s), the *z*-score can be calculated using the Formula 2.5.

$$z = \frac{(O - E)}{\sqrt{E(1 - p)}} \tag{2.5}$$

in which p is the probability of a given word obtained within the whole corpus with $p = \frac{F_c}{T_c}$ and E is the expected frequency of a given word obtained by $E = p \times T_s$.

Log-likelihood. The *Log-likelihood* (LL), or G^2, also known as LL test, proposed by Read and Cressie (1988) and Dunning (1993), is an alternative to the chi-square test. To calculate the *Log-likelihood*, we need to build a contingency table as in Table 2.3, where a denotes the observed occurrence of word w in corpus 1; b denotes the observed occurrence of word w in the reference corpus 2; c denotes the number of words in corpus 1; and d denotes the number of words in the reference corpus 2.

Table 2.3: Contingency table

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

The value of 'c' corresponds to the number of words in corpus 1, and 'd' corresponds to the number of words in corpus 2 (N values). Values 'a' and 'b' are called the observed values O, whereas the expected values E need to be calculated with Formula 2.6:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{2.6}$$

Table 2.4: Example in contingency table

	Corpus 1	Corpus 2	Total
Frequency of word	3	6	3+6
Frequency of other words	100-3	100-6	100+100-3-6
Total	100	100	100+100

In the example showns in Table 2.3, N1 = c, and N2 = d. Consequently, $E1 = \frac{c \times (a+b)}{(c+d)}$ and $E2 = \frac{d \times (a+b)}{(c+d)}$. In other words, in order to calculate E1 (the expected values) for corpus 1 we first need to multiply the total number of words (c) by the sum of observed values in corpus 1 (a) and corpus 2 (b) then divide by the sum of the total number of words in corpus 1 (c) and corpus 2 (d). Similarly, to calculate E2 (the expected values) for corpus 2 we first need to multiply the total number of words in corpus 2 (d) with the sum of the observations from both corpora (a+b) then divide by the sum of total number of words of corpus 1 (c) and corpus 2 (d). Using the hypothetical data in Table 2.4 we find: E1 = 4.5 and E2 = 4.5. The size of the two corpora are considered in the calculation for the expected values, so we do not need normalize the figures before applying the formula. We then use the calculated expected value to calculate the Log-likelihood value with Formula 2.7:

$$LL = 2 \sum_i O_i ln \left(\frac{O_i}{E_i} \right) \tag{2.7}$$

Specifically, the *Log-likelihood* equals 2 times the summation of O_i times the natural logarithm of O_i over E_i. This equates to calculating the *Log-likelihood* of the above hypothetical data in Table 2.4 as follows: 2*((a*ln(a/E1))+(b*ln(b/E2))). Resulting in LL=2*((3*ln(3/4.5)) + (6*ln(6/4.5))) = 1.02.

The *Log-likelihood* (LL), or G^2, can also be used to indicate the collocational strength between two words. This measure calculates whether word w_1 is a collocation of word w_2 by estimating the ratio between how often it occurs in a text (or a corpus) compared to how often it would be expected to occur based on random chance. In order to calculate LL we need to construct a contingency table as shown in Table 2.5, where a denotes the frequency of the co-occurrence; b denotes the number of instances where the unit does not co-occur with the collocate; c denotes the

number of instances where the collocate does not co-occur with the unit; and d denotes the number of words in the corpus minus the number of occurrences of the unit and the collocate. The motivation for us to calculate LL lies in the needs of investigating the orthographic and semantic relationships between constituents of compound words such as the semantic transparency (see e.g., Libben et al., 2003) and the semantic relations (Gagné and Shoben, 1997; Pham and Baayen, 2013).

Table 2.5: Contingency table for LL calculation

	W_2	Not-W_2	Total
W_1	a	b	a+b
Not-W_1	c	d	c+d
Total	a+c	b+d	a+b+c+d

The *Log-likelihood* then can be calculated with Formula 2.8, which is adapted from the one used by Cantos Gómez (2013).

$$\begin{aligned} LL = 2 \times (a \times log(a) + b \times log(b) + c \times log(c) + d \times log(d) - (a+b) \times log(a+b) \\ -(a+c) \times log(a+c) - (b+d) \times log(b+d) - (c+d) \times log(c+d) \\ +(a+b+c+d) \times log(a+b+c+d)) \end{aligned} \tag{2.8}$$

The higher the G^2 value, the more significant is the difference between two frequency scores. A G^2-value of 3.8 or higher is significant at the level of $p < 0.05$ and a G^2-value of 6.6 or higher is significant at $p < 0.01$. The G^2 value itself is always a positive number. However, we can compares relative frequencies between the two corpora in order to specify positive keyness (an indicator for '+' overuse) and negative keyness (an indicator for '-' underuse) of corpus 1 relative to corpus 2. Positive keyness indicates a specific keyword occurs more often in the compared corpus than would be expected by chance in comparison with the reference corpus. On the contrary,

negative keyness means that a specific keyword occurs less often in the compared corpus than would be expected by chance in comparison with the reference corpus.

Table 2.6: Sample of the database with different measures extracted from the GENLEX-VIET corpus.

Word	Word1	Freq1	Word2	Freq2	Texts	Gap	Joint	MI	Z	MI3	LogLik
bạch yến	bạch	7495	yến	716	9	1	12	7.85	15.90	-14.28	106.44
bà cô	bà	110126	cô	140005	1184	2	1817	3.60	8.24	7.45	5684.84
bà con	bà	110126	con	272411	4387	1	7475	4.68	84.56	13.57	33451.41
ba gác	ba	66504	gác	4971	158	1	268	6.38	41.63	-0.83	1826.95
bà già	bà	110126	già	23077	1225	1	1977	6.32	110.10	7.81	13228.83
bái đường	bái	796	đường	134375	8	1	11	3.41	0.19	-14.65	31.85
bái kiến	bái	796	kiến	108051	17	1	21	4.65	4.38	-11.86	94.57
bái phục	bái	796	phục	59531	58	1	70	7.25	30.50	-6.65	558.36
bài tập	bài	42675	tập	92863	1148	1	1695	5.46	66.82	7.15	9422.98
bài tiết	bài	42675	tiết	38378	450	1	588	5.21	34.02	2.57	3084.76
bài toán	bài	42675	toán	24627	816	1	1054	6.69	94.24	5.09	7618.58
bài trí	bài	42675	trí	50857	237	1	257	3.61	3.19	-1.02	811.00
bài xích	bài	42675	xích	3604	29	1	33	4.47	4.67	-9.90	140.98
ba lá	ba	66504	lá	28255	159	1	204	3.48	1.59	-2.02	611.51
ba má	ba	66504	má	14527	449	1	1114	6.89	105.32	5.33	8332.80
bẩm sinh	bẩm	10476	sinh	138692	1097	1	1609	6.83	123.68	6.92	11799.71
bán buôn	bán	90580	buôn	13282	480	1	627	5.75	47.19	2.84	3730.83
bán chác	bán	90580	chác	484	39	2	41	6.59	17.80	-8.96	290.02
bàn chân	bàn	82868	chân	59642	1420	1	2570	5.74	95.40	8.95	15230.08

Table 2.6 shows one collocation set in the database, which has been computed with the untokenized corpora (no word boundary). Word1 is the constituent to the left, followed by Freq1 (the first constituent's frequency in the whole index); Word2 is the constituent to the right, followed by Freq2 (the second constituent's frequency in the whole index); Texts is the number of texts this pair was found in; Gap is the most common distance between Word1 and Word2; Joint is their joint frequency; MI, *z*-score, MI3, and Log-likelihood are the collocational measures between two constituents as presented above. Based on the MI scores of any pairs of syllabemes, we can find out compound words and random syllabeme pairs. The syllabeme pairs with high MI score are almost likely compound words, while the one with low MI

score might be random pairs.

Table 2.7: Correlations between `Frequency`, `Texts`, `Gap`, `Joint`, `MI`, `z-score`, `MI3`, and `Log-likelihood`. The upper diagonal part of the table contains the estimates of the correlation coefficients, and the lower diagonal part contains the corresponding p-values. The figures in bold are the ones that are not significant.

	GenCorFreq	Texts	Gap	Joint	MI	z-score	MI3	LogLik
GenCorFreq	*****	0.895	-0.049	0.939	0.057	0.496	0.542	0.934
Texts	<0.001	*****	**-0.021**	0.968	**0.014**	0.485	0.638	0.919
Gap	0.007	**0.461**	*****	**-0.026**	-0.146	-0.081	**-0.003**	**-0.039**
Joint	<0.001	<0.001	**0.376**	*****	**0.027**	0.495	0.601	0.962
MI	0.001	**0.695**	<0.001	**0.376**	*****	0.607	-0.057	0.116
z-score	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.426	0.618
MI3	<0.001	<0.001	**0.854**	<0.001	0.001	<0.001	*****	0.561
LogLik	<0.001	<0.001	**0.053**	<0.001	<0.001	<0.001	<0.001	*****

Table 2.7 shows the results of a Spearman correlation analysis on the "naive" collocation ranks of two constituents of compounds based on the collocational strength obtained from the corpus by the various measures. The collocation ranks of the *z*-score significantly correlate with all the other measures. The *MI*, *MI3*, and *LogLik* values all correlate significantly with one another. The rank correlation between *MI* and *MI3* is not particularly high.

Semantic space models

Besides frequencies of occurrences and dispersion, psycholinguists have also considered quantifying semantic and associative relations between words. Some mathematical models of semantic memory put the emphasis on pure association and propose that co-occurrence in utterances is the basis upon which word meanings are built (Lund and Burgess, 1996; Burgess and Lund, 1998; Landauer and Dumais, 1997; Landauer et al., 1998). Among many other attempts, two prominent models of this type have been developed over the decades, Burgess and Lund's *Hyperspace Analogue*

to Language (HAL) and Landauer and Dumais' *Latent Semantic Analysis* (LSA). According to these two models, a word's meaning is determined by the words that it appears with[3]. If two words appear together more than they appear with other words, we expect that the meaning of those two words to be highly related. To detect whether two words are related, HAL and LSA both make use of corpora, which reflect authentic, random and representative samples of the utterances that appear in the language.

Latent Semantic Analysis (LSA). The underlying philosophy of LSA (Landauer and Dumais, 1997; Landauer et al., 1998) is that the total information about all the contexts of terms in which a given term occurs or does not occur provides a set of mutual constraints that largely establishes the meaning similarity of a set of words. LSA begins with a term-document matrix that count how often a word a word occurs in a number of contexts or documents. In LSA, a Term by Document matrix is built from a corpus of text where Term is the number of terms in the corpus and Document is the number of documents (LSA considers a document as a 'context' and words, or word phrases, are considered as terms). Each element in the matrix is transformed based on Formula 2.9:

$$M'_{t,d} = \frac{\ln(1 + M_{t,d})}{-\sum_{i=0}^{D} P(i|t) \ln P(i|t)} \tag{2.9}$$

in which P(i|t) is the probability that context i is active, given that item t has occurred (i.e. it is the result of dividing the raw frequency, $M_{t,d}$, by the total of the item vector, $\sum_{i=0}^{D} M_{t,i}$), which is the entropy of the item over all contexts.

The matrix D is subjected to a singular value decomposition, resulting in a new

[3]This idea is not new in linguistics. Firth (1957, 11) referred to this as "You shall know a word by the company it keeps."

matrix in a reduced space of 300 latent dimensions[4], similar to principal components analysis. The meaning of a word is represented in LSA as a *vector* in this 300-dimensional space. In a relative way, LSA approximates aspects of human language learning and comprehension. The semantic similarity between two items is evaluated through the cosine of the angle between their LSA vectors. This study made use of the `LSA` package (Wild, 2011) in R (R Core Team, 2013) to produce the LSA scores based on the two corpora. The processes, using the LSA package, comprise the following steps:

1. A textmatrix is constructed with the `textmatrix()` function from the input corpus.
2. The singular-value decomposition is then executed over this textmatrix and the resulting partial matrices are truncated and returned by the `lsa()` function.
3. The number of dimensions to keep can be set using various dimensionality calculation routines (e.g., `dimcalc_kaiser()`).
4. The resulting latent semantic space can be converted back to textmatrix format using the `as.textmatrix()` function.

Hyperspace Analogue to Language (HAL). The HAL model (see Lund et al., 1995; Lund and Burgess, 1996; Burgess and Livesay, 1998) of lexical semantics uses global word co-occurrence from a large corpus of text to calculate the distance between words in co-occurrence space. Burgess and colleagues collected raw text materials from online USENET groups with more than 130 million words and built a word frequency database. The HAL model defines context simply as the words that

[4]The original LSA divided its corpus into 30,000 episodes, and assessed the number of times each one of words appeared in each episodes. Instead of assigning 30,000 individual values to each word, factor analysis reduces the number of values to about 300.

immediately surround a given word. HAL computes an $N \times N$ matrix, where N is the number of words in its lexicon, using a 10-word reading frame that moves incrementally through a corpus of text. In HAL, anytime two words are simultaneously in the frame, the association between them is increased; that is, the corresponding cell in the $N \times N$ matrix is incremented. The amount by which the association is incremented varies inversely with the distance between the two words in the window, specifically, $\Delta = 11 - d$, where d is the distance between the two words in the window. For each word pair in the corpus, HAL assigns a co-occurrence value based on how close the two words are within the frame window of 10 words. In that window, words that appear adjacent to one another get a score of 10; words that are separated by one word get a score of 9; and so on. Each cell of the matrix represents the summed co-occurrence counts for a single word pair, which is direction sensitive. The similarity measurements is then applied to word vectors. This yields a measure of semantic similarity between any desired pair of words. Finally, HAL gives a word by word matrix (for instance, the original English HAL is a 70,000 x 70,000 matrix) that reflects word to word co-occurrence. The word's meaning is defined by the pattern of values, the *vector*, in each of the cells in the matrix for each word, i.e., each word has 140,000 numbers assigned to it. The distance metrics come from the Minkowski family of distance metrics, which, as stated by Lund and Burgess (1996, p. 204), involve the familiar Euclidean (as $r = 2$) and city-block (as $r = 1$) metrics as presented in Formula 2.10:

$$\text{distance} = \sqrt[r]{\sum(|x_i - y_i|)^r} \tag{2.10}$$

The HAL model is used for estimating the representation of linguistic units (e.g., words) as a function of the contexts in which linguistic units occur. The present study

made use of the tool provided by Shaoul and Westbury (2006) to compute the HAL scores for the relevant linguistic units. Shaoul and Westbury (2006) implemented *HiDEx* (High Dimensional Explorer) that extends HAL in two ways: It removes the unwanted influence of orthographic frequency from the measures of distance, and it finds the number of words within a certain distance of the word of interest (NCount, the number of neighbors). HiDEx allows users to build and analyze many variations of the HAL model. It has some features that make it very useful for measuring the similarity of words in terms of their contexts.

The results of the semantic space models in this study are two large matrices stored as two big RDA dataframes in the R statistical computing software (R Core Team, 2013) which are ready for calculating the measures of distance between words. The associative semantic relations between word pairs, or constituents of compound, can be fitted in to extract the desired resutls.

Corpus Comparison Results

Comparing the subtlex-viet corpus to the GENLEX-VIET corpus

We compared the word length distribution of the SUBTLEX-VIET corpus (78,817,521 words) with that of the GENLEX-VIET corpus (82,263,474 words). This comparison aimed to identify potential differences between the two databases resulting from the raw materials of the two corpora, and to test whether the SUBTLEX-VIET corpus is representative of contemporary Vietnamese. As we previously mentioned, lots of film subtitles were translated from English films. The current section tests the pros and cons of the GENLEX-VIET and SUBTLEX-VIET corpora. To test the normality of word length in character, we first plot the distribution of summed token frequency

of word length for monosyllabic and polysyllabic words of two corpora combined, see Figure 2.2. Interestingly, the distribution indicates that short words are more representative of oral language since the token frequencies of one two-letter words in the SUBTLEX-VIET corpus are 26% higher than those of the GENLEX-VIET corpus and the GENLEX-VIET corpus has a higher likelihood of longer words than the SUBTLEX-VIET corpus. This might also be due to the timing for reading subtitles having to be matched with the rapid transition of film and television scenes. Alternatively, it could be that Vietnamese tends to use shorter words in spoken language simply because such words are preferred in speech. Further research is required to disentangle this question.

In order to statistically analyze the relationship between the two corpora, the corpora were reduced to only those words that were common to both. All 21,498 words with a frequency of one per million greater than 0 were included in a correlation analysis which showed that the two frequency-per-million counts were highly correlated, r=0.784, p<0.001 (see also New et al., 2007; Dimitropoulou et al., 2010, for a similar correlation between subtitle and written text corpora). As shown in Figure 2.3, the distribution curves are highly similar. Figure 2.3 shows that, in both corpora, one- and two-syllabeme words have the highest summed frequency. The type frequency of mono- and disyllabic words in the GENLEX-VIET and SUBTLEX-VIET explained 92.8% and 94.6% respectively.

The frequency database from these two corpora offers a significant addition to Vietnamese corpora in part due to its large sample size (over 160 million words between the two corpora), its up-to-date current content (years 2006 – 2010), and to the inclusion of large samples of daily language use (i.e., newspapers, film subtitles). A further advantage of this database is that frequency counts were not only based on syllable forms but also based on word forms. It is a tool that allows systematic inves-

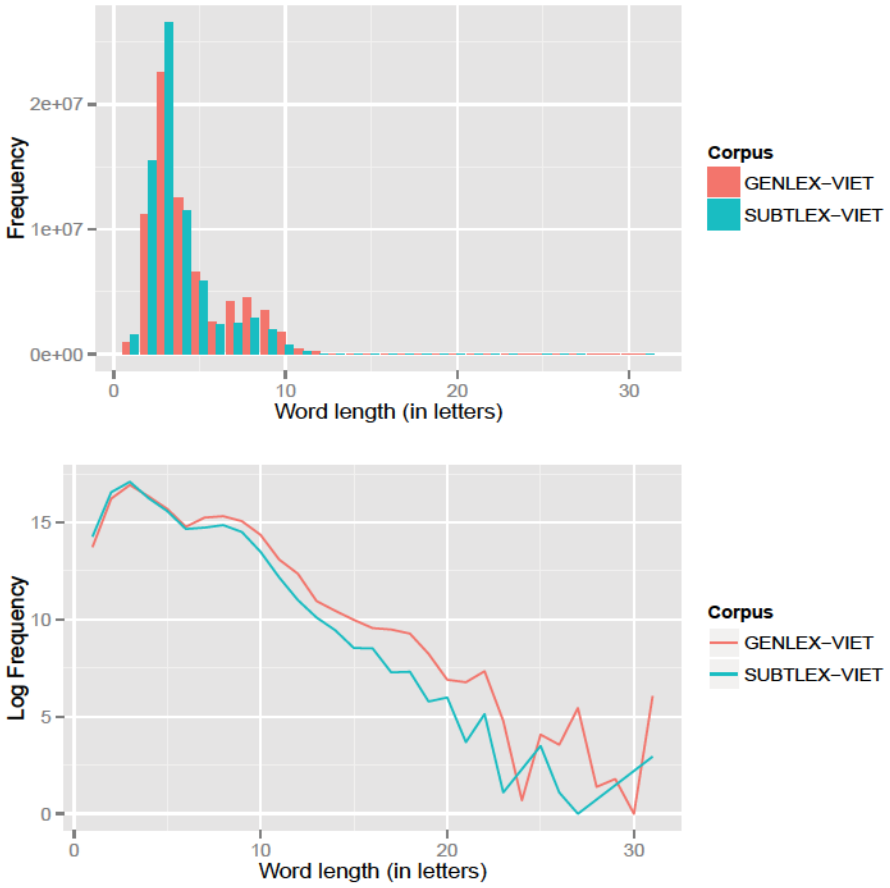

Figure 2.2: Distribution of word length in letters for monosyllabic and polysyllabic words. The upper plot shows the distribution of words' raw frequency as a function of word length. The lower plot shows the distribution of words' log frequency as a function of word length.

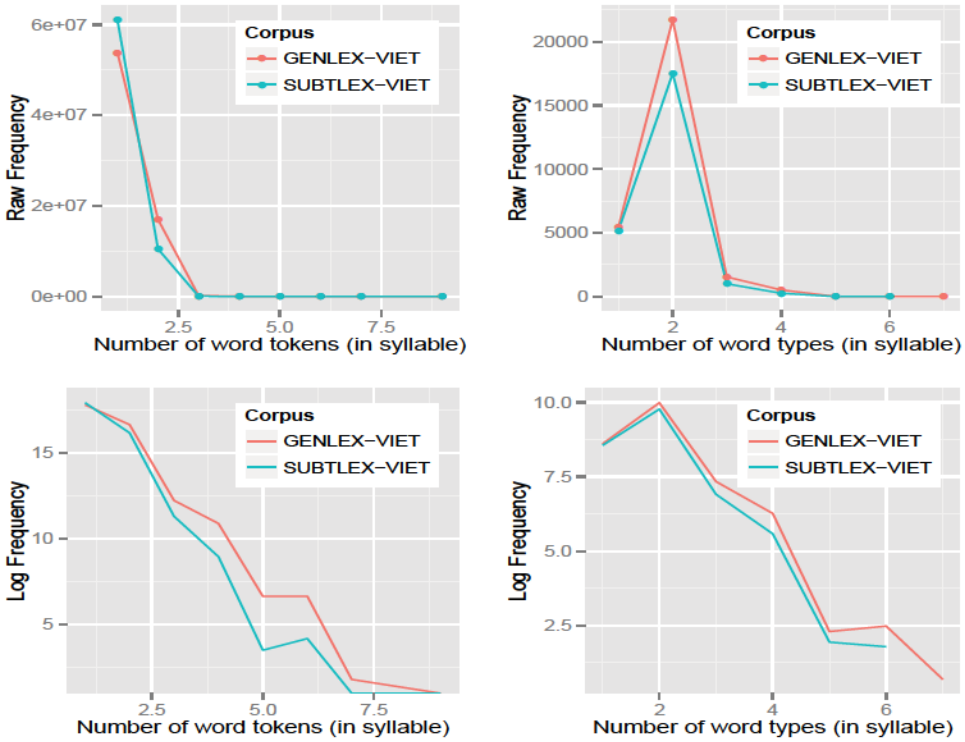

Figure 2.3: Distribution of summed word, monosyllablic and polysyllabic words, frequencies as a function of word length (calculated in number of syllables, including spaces between syllabemes, if any) for the GENLEX-VIET and the SUBTLEX-VIET. The upper plots show the raw frequency as a function of word length; the lower plots show the frequency in log-scale as a function of word length. The plots for the token frequency are on the right. The plots for the type frequency are on the left.

tigation of frequency and distributional characteristics of the Vietnamese language at the phoneme, word, and sentence levels. Although our primary purpose for constructing this database was to contribute a frequency database for psycholinguistic analysis, we expect it and the associated LSA and HAL matrixes to be used for other kinds of analysis such as linguistic analyses on general word classes including nouns, verbs, and adjectives or other linguistic analysis for the analysis of word classes or discourse analysis, and textual analysis. Frequency and distributional information at the syllabeme, word, tone, and grammatical levels is needed for a variety of pedagogical, theoretical, and experimental reasons. For example, knowledge of such frequencies will allow researchers to profile or test selected aspects of language in individuals who learn Vietnamese as a first or primary language, whereas information regarding frequency and the distributional characteristics of linguistic features is needed to develop stimuli for empirical validation and elaboration.

Table 2.8: The intercorrelations between eight available frequency measures (four word frequency measures and four syllabeme frequency measures) for 21498 words in the single-subject lexical decision experiment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Log10GenFreq							
(2) Log10GenDisp	0.99						
(3) Log10SubtFreq	0.79	0.79					
(4) Log10SubtDisp	0.79	0.80	0.99				
(5) LogFreqSyll1	0.19	0.20	0.25	0.24			
(6) LogFreqSyll2	0.24	0.23	0.26	0.26	0.20		
(7) LogDispSyll1	0.17	0.18	0.21	0.21	0.97	0.20	
(8) LogDispSyll2	0.25	0.25	0.26	0.26	0.19	0.96	0.20

Table 2.8 shows the intercorrelations between the different measures. From this table, it can be seen that the correlations between the GENLEX-VIET corpus word frequencies and the SUBTLEX-VIET word frequencies are .79 for Log10SubtFreq and

.80 for Log10SubtDisp. These are in line with the correlations observed between subtitles and written frequencies in the other languages tested, such as English, French, and Dutch. The correlation table also shows the significant negative correlations between the frequency measures, with higher correlations for the word frequencies than for the syllabeme frequencies, and not much difference between frequency and dispersion measures.

Comparing corpora using frequency profiling

In this section, we use a keyword comparison as another means to investigate differences between the two corpora (see Rayson and Garside, 2000, for details of the methodology). For this comparison we have selected the GENLEX-VIET corpus as the reference corpus (because it is larger) and the SUBTLEX-VIET corpus as the comparator. In a keyword comparison, if the word occurs 4% of the time in the comparator and 5% in the reference corpus it will not turn out to be "key" but if the scores are 24% and 5%, the first would be "key" (see e.g., Cantos Gómez, 2013, for further discussion). In other words, key words are words that appear with statistically unusual frequency in a text or a corpus in comparison with a larger reference corpus. They are words that are either more frequent in, or unique to, a target corpus compared with a general reference corpus. This phenomenon has been called *keyness* (see e.g., Scott and Tribble, 2006, for further discussion).

The resulting output consists of five columns, as shown in Table 2.9. The first column is the word column, the next two columns are its frequency in the compared corpus (the SUBTLEX-VIET corpus), and its percentage of all words in the corpus. "Keyness" and "P" together show how distinctive the word is: when "keyness" is very high, and "P" (the probability of the keyness being accidental) is very low, the word can safely be called a keyword. The word *anh* 'you (man)' appears to be the most key item.

Table 2.9: Top twenty most frequent words in two corpora. Table 2.9a is "key" words of the subtitle corpus. Table 2.9b is "key" words of the general corpus. Freq is the number of occurences in the corpus. Percent is the relative frequency within the corpus. RC_Freq is the frequency index in the reference corpus (the general corpus). RC_Percent is the relative frequency within the reference corpus. Keyness is the measure comparing the relative frequencies of a word in the subtitle corpus versus the reference corpus.

(a) Subtitle corpus

Word	Freq	Rank	Percent	Keyness	P
không	1613316	1	2.03	422529.22	0.00
là	1430765	2	1.80	196632.88	0.00
có	960343	3	1.21	19096.22	0.00
của	877458	4	1.10	-21273.89	0.00
anh	865681	5	1.09	462346.47	0.00
đã	842271	6	1.06	30982.55	0.00
sẽ	797497	7	1.00	181083.78	0.00
đi	770351	8	0.97	308145.56	0.00
đó	683711	9	0.86	161759.94	0.00
gì	641880	10	0.81	397154.09	0.00
được	624201	11	0.78	-633.58	0.00
rồi	619070	12	0.78	372238.16	0.00
phải	587736	13	0.74	94760.02	0.00
này	578759	14	0.73	27606.25	0.00
cho	548137	15	0.69	-400.61	0.00
một	517868	16	0.65	-25444.66	0.00
đây	511701	17	0.64	291359.44	0.00
biết	491875	18	0.62	238986.78	0.00
cô	456357	19	0.57	238094.97	0.00
lại	409422	20	0.51	15199.17	0.00

(b) General corpus

Word	RC_Freq	Rank	RC_Percent	Keyness	P
của	1104630	1	1.35	-21273.89	0.00
là	802993	2	0.98	196632.88	0.00
có	798898	3	0.98	19096.22	0.00
một	707834	4	0.87	-25444.66	0.00
không	676941	5	0.83	422529.22	0.00
được	667926	6	0.82	-633.58	0.00
đã	646318	7	0.79	30982.55	0.00
các	636108	8	0.78	-220904.25	0.00
cho	582674	9	0.71	-400.61	0.00
những	503852	10	0.62	-37453.10	0.00
này	424934	11	0.52	27606.25	0.00
để	369818	12	0.45	-2050.76	0.00
sẽ	360211	13	0.44	181083.78	0.00
khi	352153	14	0.43	-18252.85	0.00
cũng	344340	15	0.42	-13143.36	0.00
đến	324826	16	0.40	857.03	0.00
lại	313488	17	0.38	15199.17	0.00
phải	309717	18	0.38	94760.02	0.00
đó	302094	19	0.37	161759.94	0.00
năm	255429	20	0.31	-98542.49	0.00

Table 2.9 shows the top twenty most frequent words in the two corpora ordered by their ranks from highest to lowest. We can see that 65% of the words in these two top-twenty lists are shared.

Table 2.10: Top and tail twenty keywords in the SUBTLEX-VIET corpus (reference corpus: GENLEX-VIET corpus); 2.10a is the top-twenty keywords; 2.10b is the tail-twenty keywords.

(a) Top-20 keywords

Rank	Frequency	Keyness	Word
1	865681	462346.47	anh
2	1613316	422529.22	không
3	641880	397154.09	gì
4	619070	372238.16	rồi
5	316400	324641.50	cậu
6	770351	308145.56	đi
7	360552	300441.03	ta
8	511701	291359.44	đây
9	290960	279234.62	chúng ta
10	276530	265102.12	đấy
11	491875	238986.78	biết
12	456357	238094.97	cô
13	311360	213076.09	đâu
14	1430765	196632.88	là
15	351258	184709.30	muốn
16	797497	181083.78	sẽ
17	370565	177210.28	em
18	238821	173212.00	chứ
19	395090	165157.16	nó
20	683711	161759.94	đó

(b) Tail-20 keywords

Rank	Frequency	Keyness	Word
9339	206401	-220904.25	các
9338	31452	-152828.91	tại
9337	76008	-98542.49	năm
9336	7364	-97219.80	đồng
9335	40855	-76717.44	nước
9334	45649	-72017.98	hai
9333	232	-65161.03	doanh nghiệp
9332	20509	-47188.52	cao
9331	26180	-46351.16	do
9330	1897	-39223.54	đầu tư
9329	10021	-38720.84	bóng
9328	319776	-37453.10	những
9327	6264	-36636.09	giải
9326	5157	-36588.65	giảm
9325	4237	-35994.01	hiện
9324	1136	-35167.36	quy định
9323	1266	-34799.50	cầu thủ
9322	19757	-34425.28	đội
9321	9117	-33439.82	một số
9320	128980	-33035.99	nhiều

The words in Table 2.10a have the highest keyness scores. These words seem to highlight the genre difference between the two corpora showing many spoken words as being "key": e.g., pronouns: *anh* 'you', *ta* 'we', *chúng ta* '(inclusive) we', *em* 'you', *cô* 'aunt', *nó* 'it, he', etc., interjections: *nhé*, *ừ*, etc., and adverb: *chứ*, *đấy*, etc. In contrast, Table 2.10b shows the 'negative' keywords, i.e., words that occur more often in the reference corpus (the GENLEX-VIET corpus) than in the target corpus and are negatively key in the compared corpus. Negative keywords in the SUBTLEX-VIET corpus are, for instance, *doanh nghiệp* 'business', *đầu tư* 'invest', *quy định* 'to

stipulate', etc. - words that are rare in the SUBTLEX-VIET corpus but more common in the GENLEX-VIET corpus.

Validation Results

As mentioned in the Introduction, visual lexical decision response times are robustly sensitive to word frequencies (Balota et al., 2004; Brysbaert and New, 2009; Pham and Baayen, 2014; Yap and Balota, 2009). As a consequence, validating word frequency measures with behavioural data is an informative process for validating corpus data (see Keuleers et al., 2010, for further discussion). In the previous sections, we described the creation of two Vietnamese corpora and a wide range of lexical predictors derived from these corpora. We also explore the relationships between the two corpora using many of the calculated measures. This section presents the subsequent experimental validation of the word frequency and dispersion measures. The methods and design of the visual lexical decision data is reported in detail in Pham and Baayen (2014). In what follows, we present a brief description of the methods and the analyses comparing frequency and dispersion measures in the two corpora, following previous work on Dutch by Keuleers et al. (2010).

All mono-syllabic and disyllabic words from the Vietnamese Dictionary (Hoàng, 2000) were selected, with the exception of reduplicated words and one character words, resulting in a list of words comprising 20,051 words. The Wuggy pseudoword generator Keuleers and Brysbaert (2010) was used to construct a corresponding pseudoword for each word in the experiment. Each pseudoword differed from the reference word by one subsyllabic segment (i.e., the onset, nucleus, or coda) per syllable. This meant that a one-syllable nonword differed in one position from its reference word and that a two-syllable nonword differed in two positions from its reference word. This study is a single-subject study, with the

first author as sole participant. He responded to all 40,102 trials (including nonword foils), requiring 46 hours over a four week period to complete. The participant performed a visual lexical decision for all word and pseudoword forms. The participant randomly chose a file with stimuli to run. The participant was not allowed to run more than two blocks per day (each run equal to 1 hour, including breaks).

The two dependent variables were accuracy (percentage of correct response) and reaction time (RT) of the correct trials. Mean accuracy of the participant was 95% for the words and 96% for the nonwords. Table 2.11 shows mean RTs were 637 ms ($SD = 129$) for di-syllabic words and 688 ms ($SD = 183$) for mono-syllabic words. Mean RTs were 649 ms ($SD = 145$) for the words (both mono- and disyllabic) and 750 ms ($SD = 166$) for the nonwords (both mono- and disyllabic). The reaction time analysis reveals a main effect of word frequency measured by both SUBTLEX-VIET and GENLEX-VIET corpora frequency norms. The main effect of the GENLEX-VIET frequency count was significant, showing that high frequency words according to this frequency count were responded to 48 ms faster than the low frequency words.

Table 2.11: Mean response times grouped by word type and wordedness of the single-subject experiment. RT: Reaction times; SD: Standard deviation; CI: Confidence interval; NW: Nonword; WRD: Word.

	Compound	Word	RT	SD	CI
1	no	NW	749	201	5.95
2	no	WRD	688	183	5.54
3	yes	NW	750	154	2.48
4	yes	WRD	637	129	2.08

The first line of Table 2.12 shows that the results of the initial linear regression model in which the log frequency measure (base 10) for the GENLEX-VIET corpus frequency (Log10Gen-Freq), the GENLEX-VIET corpus dispersion (Log10GenDisp), the SUBTLEX-VIET corpus (Log10-SubtFreq), and the SUBTITLE-VIET dispersion (Log10-

Table 2.12: The percentage of reaction-time (RT) variance explained by the different frequency measures, for disyllabic words in the single-subject visual lexical decision experiment. N is the number of unique compound words use as stimuli.

N=15913	Word frequency measures			
	GENLEX-VIET		SUBTLEX-VIET	
	Frequency	Dispersion	Frequency	Dispersion
Log	12.2	12.1	11.1	11.1
Log + $\log^2$	12.7	12.5	11.5	11.5

SubtDisp) are used as predictors of RTs. As can be seen, the log frequencies of the GENLEX-VIET and SUBTLEX-VIET explained 12.2% and 11.1% respectively of the variance in RTs. In the second model, where both log(freq) and $\log^2$(freq) are used as predictors of the RTs, the percentage of variance explained by each of the measures of frequency increased in both corpora (GENLEX-VIET: 12.7% and SUBTLEX-VIET: 11.5%). The amount explained by each of the dispersion variables increased 0.4% in the model with the squared term. Because of the large number of observations, differences in explained variance as small as .1 are statistically significant.

In order to check whether our megastudy data can be generalized to the larger population, we ran a small set of stimuli on a larger number of participants. We then compare response times with the frequency measures used in the megastudy (for detailed discussion see Pham and Baayen, 2014). The subsection that follows describes this second data set, which has the same design and method as the single-subject megastudy.

The materials were 550 disyllabic compounds randomly selected from the materials in the single-subject experiment presented above. The stimulus words were selected pseudo-randomly such that high- and low-frequency compounds had an equal chance of selection. Thirty three participants were recruited at the Vietnam National University to take part in the visual lexical decision experiment (mean age 21.9, range 20 – 22 years, 12 males). All

participants were native Vietnamese speakers and had at least 14 years of education (all of them were students). We employed the same apparatus that was used in Experiment 1, but with fewer items spread across more participants.

The mean accuracy of all participants was 79% for the words and 73% for the nonwords. Table 2.13 shows that the mean RTs of the participants were 677 ms ($SD = 272$) for disyllabic words and 899 ms ($SD = 315$) for the disyllabic pseudowords. Note that there is a big difference between the accuracy rate in the single subject design vs. the multi-subject design. This may be due to participants' general unfamiliarity with the task since they only participate in the experiment once; while the subject in the single subject design participated in many sessions. Then, in a way, the single subject is more familiar with the task. He was probably also better motivated.

Table 2.13: Mean response times grouped by wordhood of the supplementary experiment. RT: Reaction times; SD: Standard deviation; CI: Confidence interval; NW: Nonword; WRD: Word.

	Word	RT	SD	CI
1	NW	899	315	7.94
2	WRD	677	272	3.82

Table 2.14: The percentage of reaction-time (RT) variance explained by the different frequency measures, for disyllabic words in the supplementary lexical decision experiment. N is the number of unique words use as stimuli.

N=542	Word frequency measures			
	General corpus		Subtlex-Viet corpus	
	Frequency	Dispersion	Frequency	Dispersion
Log	45.8	45.4	40.7	41.1
Log + $\log^2$	45.8	45.3	40.7	41.0

In order to validate the RTs with frequency and dispersion for each word, we calculated mean RTs for each word across all 33 participants. Table 2.14 displays the

percentage of RT variance accounted for by the different frequency measures. Frequency measures were log10 transformed. Because the relationship between log frequency and word processing performance is not completely linear (in particular, a floor effect seems to be reached for words with a frequency above 100 per million) (Balota et al., 2004; Baayen et al., 2006; Keuleers et al., 2010). Following Keuleers et al. (2010), we report a linear regression analysis both for log(frequency) and log(frequency) + $\log^2$(frequency).

The first lines of Table 2.14 show the results of the first linear regression model in which the log frequency measure (base 10) for the GENLEX-VIET corpus frequency (Log10GenFreq), the GENLEX-VIET corpus dispersion (Log10GenDisp), the SUBTLEX-VIET corpus (Log10SubtFreq), and the SUBTLEX-VIET dispersion (Log10SubtDisp) are used as predictors of RTs. As can be seen, the log frequencies of the GENLEX-VIET corpus and SUBTLEX-VIET corpus explained 45.8% and 40.7% respectively of the variance in RTs. In the second model, where both log(freq) and $\log^2$(freq) are used as predictors of the RTs, the percentage of variance explained by each of the measures of frequency remained the same in both models, but the amount explained by each of the dispersion variables decreased by 0.1% in the model with the squared term. This might be because of the genre characteristics of the two corpora. While the number of times a word occurs (frequency) in the newspaper corpus is important to the participants' response times, the number of films that contains a word (dispersion) is important to the participants' response times. The explained variance in this experiment is larger than the single-subject experiment might be due to the fact that the number of observations in this experiment is much smaller than that of single-subject experiment.

General Discussion

The present study describes the construction of two Vietnamese corpora and the derivation of a large set of lexical predictors from these corpora. The corpora were further shown to follow the classic quantitative pattern known as Zipf's law. Pearson correlations tests showed that the collocational measures were highly correlated in expected ways. We present these corpora in the hope they will make a useful contribution to the quantitative analysis of Vietnamese. The corpora were further validated by comparison of the lexical predictors and analysis using psycholinguistic data.

Corpus Creation

Subtitles are freely and readily available on various Internet sites and in many languages. In the development of our word frequency database, we automatically processed thousands of these subtitles with little effort.

Although using these subtitles for our research is convenient and inexpensive, there are some legal and ethical issues to consider. However, according to the Fair Use Act, researchers can use and make the frequencies available because no single full subtitle is available to the user. They can only access part of or a small part of subtitles, e.g., the frequency list. Therefore, to the best of our understanding, our use of the subtitles as described in this research is not a violation of copyright because the word frequency database is only a statistical description of the subtitles. This is considered "fair use" of copyrighted material (see e.g., Keuleers et al., 2010).

Corpus Comparison Results

Given previous research, one would expect dispersion measures from the SUBTLEX-VIET corpus to be better predict RTs in visual lexical decision task than frequency measures. There are two ways in which this prediction is not born out for our data: (1) The amount of variance explained are statistically identical; (2) In the GENLEX-VIET, frequency is better than dispersion but in the SUBTLEX-VIET corpus, dispersion is better than frequency. The frequency measures from the GENLEX-VIET corpus better predict RTs of the same experiments indicating that frequency measures from a complex and representative corpus, i.e., the general corpus in the current study, outperform the dispersion measures from the same corpus. As shown in the results of the multi-subject study, the dispersion measures from a genre-consistent corpus, i.e., the subtitle corpus in the current study, is a slightly better predictor for RTs than the frequency measure of the same corpus. In contrast to previous findings in French, English, and Dutch, the SUBTLEX-VIET frequency and dispersion measures explain some $1\% - 5\%$ less of the variance in RTs than the GENLEX-VIET corpus.

We discuss two possible explanations for the divergence from the previous literature: (1) the reference corpora (GENLEX-VIET) and the SUBTLEX-VIET corpus sample language materials from the same time period, which is not the case for the comparisions of the subtitle corpora in other research; and/or (2) the sizes of the reference corpora and the comparator corpus in the current study are equal in size, whereas the comparisons in the previous research are largely unequal in size. To illustrate these differences, we summarize the comparisons found in the previous literature. For French, New et al. (2007) compared the subtitle frequencies (from the 52-million-word corpus collected in 2006) with the frequencies from a classical French spoken corpus, the Corpus de reférence du Français parlé (Delic, 2004). The reference corpus contained 440,000 words and consisting of 134 recordings from 1998. For English, Brysbaert and New (2009) compared the English subtitle frequencies

(from 51-million-word film subtitle corpus from 1990–2007) with the frequencies in the work of Kučera and Francis (1967), which is a small 1960s corpus with only one million words. For Dutch, Keuleers et al. (2010) compared the Dutch subtitle frequencies (from 44-million-word corpus collected in 2009) with the frequencies of Dutch in Celex (Baayen et al., 1995), which consists of 930 entire fiction and non-fiction books (approx. 30% fiction, 70% non-fiction) published between 1970 and 1988. The current study differs from the above-mentioned studies in that: (1) this study compares two contemporary corpora (items were collected from the years 2006-2010); and (2) the two corpora used for the comparison are of comparable size.

There may be other reasons why the frequency effect between the subtitle and reference corpora in English, French, and Dutch did not generalize to Vietnamese. First, a large proportion of the movies and TV shows popular in Vietnam are either American or European, possibly making film subtitles less representative of daily language in Vietnam. However, as can be seen from the results presented above, the SUBTLEX-VIET corpus is nearly as predictive as the GENLEX-VIET corpus. Furthermore, since this corpus is relatively easy to create and can also quite easily be made into a monitor corpus[5] we believe that this makes it a valuable linguistic resource, even with the limitations associated with subtitles.

The comparison between the two corpora shows that the frequency measures from the SUBTLEX-VIET and the GENLEX-VIET are highly correlated. Interestingly, however, the distributions of frequency of the two corpora in terms of word length are of the same form. Differences are found only with either very short words or very

[5] A monitor corpus is a type of corpus which is a growing, non-finite collection of texts, of primary use in lexicography. A monitor corpus reflects language changes in a constant growth rate of corpora, leaving untouched the relative weight of its components (i.e., balance) as defined by the parameters. The same composition schema should be followed year by year, the basis being a reference corpus with texts spoken or written in one single year. An example of an English monitor corpus is the COCA corpus (Davies, 2010), which can be accessed at http://corpus.byu.edu/coca/.

long words. The SUBTLEX-VIET corpus contains more short words than the GENLEX-VIET, whereas the long words occur more often in the GENLEX-VIET corpus than in the SUBTLEX-VIET corpus. This might be due to the different characteristics of written language and spoken language. The correlation test between the frequency and dispersion measures of the two corpora suggests to us that we want to be aware of the (multi)collinearity between these predictors when including them into statistical analysis. The keyness comparison between the two corpora provides us with one way to examine the characteristic traits of each corpus with respect to genre, culture, or historical period.

Validation Results

We also tested the frequency measures calculated from the two corpora against a pair of lexical-decision studies involving monosyllabic (excluding one-letter words) and disyllabic Vietnamese words. The predictivity of frequency and dispersion measures in this study is somewhat different from English (Brysbaert and New, 2009) and Dutch (Keuleers et al., 2010). The dispersion measure of the subtitle corpus slightly better predicts lexical response times than the word frequency measure in the second experiment (41.1% vs. 40.7%), whereas they are the same in the first experiment (11.1%). Dispersion explained 30% – 40% of the variance of the RTs, as can be seen in the second experiment presented above. However, regarding the general corpus, the frequency measure outperforms the dispersion measure. Compared with the general corpus frequencies, the SUBTLEX-VIET frequencies explained 4% – 5% less variance in RTs. Therefore, the SUBTLEX-VIET word frequency measures might be a secondary resource for language research, especially in psycholinguistic studies, such as word recognition research. In contrast to previous findings in French, English, and Dutch, the SUBTLEX-VIET frequency and dispersion measures explain some 1% – 5% and

less off the variance in RTs than that of the General corpus. As we mentioned in the Introduction, frequency is the number of times a word appears in the whole corpus, whereas dispersion is the number of documents that contain a specific word. In other words, frequency focuses on how many times a word is repeated, while dispersion focuses on how many contexts a word appears. The reason that this study is not in line with previous studies might be due to one of two reasons: (1) the usage of the reference corpora and the SUBTLEX-VIET corpora are too distant from each other in terms of the eras in which the two corpora were built; and/or (2) the sizes of the reference corpora in previous studies are too small in comparision with the French, English, and Dutch SUBTLEX corpora. It is also conceivable that the SUBTLEX-VIET measures may outperform the GENLEX-VIET corpus in tasks such as auditory lexical decision or picture naming.

Based on the comparision of CELEX frequencies and the SUBTLEX-NL frequencies by Keuleers et al. (2010) and other corpus comparisons it might be the case that the more current the corpus, the larger the percentage of variance explained. It it still unclear whether we should compare two chronologically different corpora. This comparison may contain confounding results, as words are used based on social contexts. The social context fifty years ago may be different from the current context. The frequency of certain words, therefore, may not be the same over time. In other words, the frequencies of many of the words in older corpora, e.g., the fifty-year-old Brown corpus still commonly used in psycholinguistics, may no longer be valid for contemporary lexical decision tasks. The comparison of corpora from very different eras may be problematic. In the current study, we compared two contemporary corpora to avoid the problem of chronologically different corpora.

Conclusions

Our experience is that corpora and lexical databases can be created effectively for less commonly studied languages with a large population base and an electronic literary presence. We built two corpora and a database of statistical measures, semantic space, and dispersion measures. We follow Keuleers et al. (2010) suggestion that the obtained frequencies should be validated with lexical decision times. Nevertheless, others (e.g., Burgess and Livesay, 1998; New et al., 2007) have suggested that differences in quality between various frequency counts can already be detected with samples of only a few hundred words spread over the entire frequency range. Based on this claim, it may not be necessary to collect data for thousands of words. Keuleers et al. (2010) proposed that a typical one-hour experiment with some 1,000 words and nonwords may be sufficient. However, as I will show in the next chapter, this conclusion may underestimate of sampling widely items, see also Pham and Baayen (2014).

We end this study with a note. It might be time to shift to (or build) up-to-date corpora which reflect the contemporary language characteristics of the time at which a psycholinguistic study is carried out. With the tradition of using old and small "black lagoon"[6] corpora to select stimuli, we are pitting the young participants' RTs to old-usage words. When individual differences in terms of age are recorded, older participants show higher correlations with 'old' corpus frequencies than with the young SUBTLEX frequencies in the studies of Brysbaert and New (2009) and Cai and Brysbaert (2010). Keuleers et al. (2010) (and other previous studies found that frequency measures from subtitle corpora better predicts RTs than frequency

[6]This term is borrowed from Seidenberg's discussion at http://www.talkingbrains.org/2009/07/irvine-phonotactic-online-dictionary.html, Accessed on February 6, 2014.

measures from general corpus) might be confounded in one or more of the following ways, such as: (1) They compared the frequency values from the subtitle corpus against those of a much smaller corpus, so the latter values were less precise; (2) They compared frequency values derived from a modern corpus against those of a much older corpus. Inaccuracies might also arise because subtitles are twice removed from reality since (a) Movie scripts are not authentic. They are a screenwriter's attempt to imagine what people would say in various situations. Some screenwriters will be good at this, but some will be bad; (b) Subtitles are translations that may reflect cross-cultural differences. That is, how things are said or even what is said may differ across languages. If the translation is a good one, it may reduce or remove such differences, but it seems more likely than not that this is a difficult task. Surely, many subtitles reflect inappropriate cross-cultural and cross-linguistic artifacts; (c) Subtitles are constrained to a small part of the screen, and must reflect the spoken language in a reasonable amount of time to be read by the viewer. Constraints like this might affect which translation a translator decides to use for the subtitles, rendering a distorted translation.

Due to the copyright issues, only certain portions of the texts in our corpora are freely available to the public in the form of concordance lines per request. The lexical database for which there are no copyright issues have been made available at http://hdl.handle.net/10402/era.38532 for use by the research community. While we believe that the current corpora will serve many uses in the domains of corpus linguistics and psycholinguistics, we hope to keep expanding and improving the corpora. We hope to add metadata to the corpora classifying the genres and time periods of the texts. We also hope to be able to create a relational database from the corpora to be used to create an online user-friendly version of the corpus that can be accessed and used by non-specialists. We also would like to see the current

corpora become monitor corpora with new data being added on a regular basis.

References

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H., Feldman, L., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133:283–316.

Berry-Rogghe, G. L. M. (1973). The computation of collocations and their relevance in lexical studies. In Aitken, A. J., Bailey, R. W., and Hamilton-Smith, N., editors, *The Computer and Literary Studies*, pages 103–112. Edinburgh University Press, Edinburgh.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Burgess, C. and Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods*, 30:272–277.

Burgess, C. and Lund, K. (1998). The dynamics of meaning in memory. In Dietrich, E. and Markman, A., editors, *Cognitive dynamics: Conceptual change in humans and machines*. Lawrence Erlbaum Associates, Mahwah, NJ.

Cai, Q. and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS One*, 5(6).

Cantos Gómez, P. (2013). *Statistical methods in language and linguistic research*. Equinox Publishing Limited, Sheffield & Briston.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Cuetos, F., Glez-Nosti, M., Barbon, A., and Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica*, 32:133–143.

Davies, M. (2010). Corpus of contemporary American English (COCA). http://www.americancorpus.org/.

Delic, E. (2004). Présentation du *Corpus de référence du Français parlé. Recherches sur le Français parlé*, 18:11–42.

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., and Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behaviour: The case of Greek. *Frontiers in Psychology*, 1:1–12.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

Đỗ, V. X. (2012). *Cuộc hành trình đi tìm chữ Việt cổ [The journey to seek for the Vietnamese old scripts]*. Hồng Đức, Hà Nội.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press, London.

Gagné, C. L. and Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–87.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437.

Gries, S. T. (2009). Dispersions and adjusted frequencies in corpora: Further explorations. *Language and Computers*, 71(1):197–212.

Hasher, L. and Zacks, R. T. (1984). Automatic processing of fundamental information. The case of frequency of occurrence. *American Psychologist*, 39:1372–1388.

Hoàng, P., editor (2000). *Từ điển tiếng Việt [Vietnamese Dictionary]*. Khoa học Xã hội, Hà Nội. Viện Ngôn ngữ học.

Juilland, A., Brodin, D., and Davidovitch, C. (1970). *Frequency dictionary of French words*. Romance languages and their structures. Hague, Paris.

Keuleers, E. and Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behaviour Research Methods*, 42(3):627–633.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behaviour Research Methods*, 42(3):643–650.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lê, H. P., Nguyen, T. M. H., Roussanaly, A., and Ho, V. (2008). A hybrid approach to word segmentation of Vietnamese texts. In Martin-Vide, C., Otto, F., and Fernau, H., editors, *Language and automata theory and applications*, volume 5196 of *Lecture Notes in Computer Science*, pages 240–249. Springer Berlin, Heidelberg.

Lê, H. P., Roussanaly, A., Nguyen, T. M. H., and Rossignol, M. (2010). An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In *Traitement Automatique des Langues Naturelles - TALN 2010*, page 12, Montréal Canada. ATALA (Association pour le Traitement Automatique des Langues).

Libben, G., Gibson, M., Yoon, Y. B., and Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84:50–64.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments and Computers*, 28(2):203–208.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, pages 660–665, Hillsdale, NJ. Erlbaum.

McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review*, 88:375–407.

McDonald, S. A. and Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3):295–323.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677.

New, B., Pallier, C., Brysbaert, M., Ferr, L., Holloway, R., Service, U., and Joliot, H. F. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, and Computers*, 36:516–524.

Nguyễn, D. D. and Lê, Q. T. (1980). *Dictionnaire de fréquence du Vietnamien.* Université de Paris VII, Paris.

Nguyễn, D. H. (1997). *Vietnamese: Tiếng Việt không son phấn.* John Benjamins, Amsterdam.

Oakes, M. (1998). *Statistics for corpus linguistics.* Edinburgh University Press, Edinburgh.

Ooi, V. (1998). *Computer corpus lexicography.* Edinburgh University Press, Edinburgh.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, 331(6157):585–589.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1989). Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience*, 1(2):153–170.

Pham, G., Kohnert, K., and Carney, E. (2008). Corpora of Vietnamese texts: Lexical effects of intended audience and publication place. *Behavior Research Methods*, 40(1):154–163.

Pham, H. and Baayen, H. R. (2013). Semantic relations and compound transparency: A regression study in CARIN theory. *Psihologija*, 46(4):455–478.

Pham, H. and Baayen, H. R. (2014). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Submitted for publication.*

Pinker, S. (1999). *Words and rules: The ingredients of language.* Basic Books, New York.

R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling.

In *Proceedings of the workshop on Comparing Corpora, held in conjunction ACL 2000. October 2000, Hong Kong*, pages 1–6.

Read, T. and Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer series in statistics. Springer-Verlag, New York.

Scott, M. and Tribble, C. (2006). *Textual patterns : Key words and corpus analysis in language education*. John Benjamins, Amsterdam.

Shaoul, C. and Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, 38:190–195.

Southeast Asian Languages Library, S. (2009). Vietnamese text corpus. http://sealang.net/vietnamese/corpus.htm. Accessed: 2014-02-16.

Trung tâm từ điển học, V. (1998). Vietnamese corpus. http://vietlex.com/kho-ngu-lieu. Accessed: 2014-02-16.

Wild, F. (2011). LSA: Latent semantic analysis. *R package version 0.63-3*.

Yap, M. J. and Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60(4):502–529.

Zipf, G. K. (1935). *The psycho-biology of language*. Houghton Mifflin, Boston.

Appendix

Appendix A:

POS tags used in the corpora

ID	POS-tags	POS in English	POS in Vietnamese
1	Np	Proper noun	danh từ riêng
2	Nc	Classifier noun	danh từ chỉ loại
3	Nu	Unit noun	danh từ đơn vị
4	N	Common noun	danh từ chung
5	V	Verb	động từ
6	A	Adjective	tính từ
7	P	Pronoun	đại từ
8	R	Adverb	phụ từ
9	L	Determiner	định từ
10	M	Numeral	số từ
11	E	Preposition	giới từ
12	C	Subordinating conjunction	liên từ phụ
13	CC	Coordinating conjunction	liên từ kết hợp
14	I	Interjection	từ cảm thán
15	T	Auxiliary word, modal words	trợ từ
16	Y	Abbreviation	từ viết tắt
17	Z	Bound morphemes	yếu tố cấu tạo từ (bất, vô...)
18	X	Undetermined	không (hoặc chưa) xác định

Appendix B:

POS-tagged XML sample

```
<doc>
...
  <s>
            <w pos="P">Đây</w>
            <w pos="V">là</w>
            <w pos="N">câu_chuyện</w>
            <w pos="E">về</w>
            <w pos="M">một</w>
            <w pos="N">chàng_trai</w>
            <w pos="V">gặp</w>
            <w pos="M">một</w>
            <w pos="N">cô_gái</w>
            <w pos=".">.</w>
      </s>
      <s>
            <w pos="N">Ngày</w>
            <w pos="N">thứ</w>
            <w pos="M">nhất</w>
      </s>
      <s>
            <w pos="N">Chàng_trai</w>
            <w pos=",">,</w>
            <w pos="Np">Tom_Hansen</w>
            <w pos=",">,</w>
            <w pos="V">sinh_ra</w>
            <w pos="E">ở</w>
            <w pos="Np">Margate</w>
            <w pos=",">,</w>
            <w pos="Np">New_Jersey</w>
            <w pos=",">,</w>
      </s>
      <s>
            <w pos="A">lớn</w>
            <w pos="R">lên</w>
            <w pos="CC">và</w>
            <w pos="V">tin</w>
            <w pos="C">rằng</w>
            <w pos="M">...</w>
      </s>
      <s>
            <w pos="P">mình</w>
            <w pos="R">sẽ</w>
            <w pos="V">không_bao_giờ</w>
            <w pos="V">có</w>
            <w pos="R">được</w>
            <w pos="N">hạnh_phúc</w>
            <w pos="A">thực_sự</w>
            <w pos="...">...</w>
      </s>
      <s>
            <w pos="E">cho_đến</w>
            <w pos="N">ngày</w>
```

```
<w pos="V">gặp</w>                    <w pos="""">"</w>
<w pos="R">được</w>                   <w pos=".">.</w>
<w pos="""">"</w>              </s>
<w pos="N">người</w>     ...
<w pos="P">ấy</w>      </doc>
```

Appendix C:

Dispersion measures

Word	FREQ	RANGE	MAXMIN	SD	VARCOEFF	CHISQUARE	D_EQ	D_UNEQ
cảnh gần	8.00	8.00	1.00	0.01	147.34	220704.86	0.65	0.53
cánh phấn	2.00	2.00	1.00	0.00	294.69	15572.46	0.29	0.23
cảnh sát	25695.00	12434.00	46.00	0.76	5.16	869451.64	0.99	0.99
cao lương	100.00	49.00	23.00	0.08	135.55	575687.79	0.67	0.75
cặp lồng	30.00	23.00	3.00	0.02	94.21	191176.18	0.77	0.64
cạp nia	9.00	7.00	3.00	0.01	179.34	137331.11	0.57	0.51

D2	S_EQ	S_UNEQ	D3	DC	IDF	ENGVALL	U_EQ	U_UNEQ	UM_CARR
0.17	0.00	0.00	-5426.44	0.00	14.41	0.00	5.17	4.28	1.38
0.06	0.00	0.00	-21709.50	0.00	16.41	0.00	0.59	0.46	0.11
0.76	0.06	0.05	-5.66	0.06	3.80	1839.48	25376.75	25394.28	19446.87
0.26	0.00	0.00	-4592.74	0.00	11.79	0.03	67.47	74.61	25.61
0.25	0.00	0.00	-2218.07	0.00	12.88	0.00	23.22	19.08	7.61
0.15	0.00	0.00	-8039.77	0.00	14.60	0.00	5.13	4.63	1.37

AF_EQ	AF_UNEQ	Ur_KROM	F_ARF	AWT	F_AWT	ALD	F_ALD	DP	DPnorm
0.00	0.00	8.00	5.63	7427879.64	5.72	7.13	6.27	1.00	1.00
0.00	0.00	2.00	1.20	34989851.57	1.21	7.79	1.38	1.00	1.00
1606.13	1393.76	17028.34	9384.85	20117.13	2111.81	4.15	5980.62	0.93	0.93
0.02	0.08	58.00	34.54	2075863.59	20.46	6.49	27.38	1.00	1.00
0.00	0.02	26.33	15.24	3737474.52	11.37	6.77	14.35	1.00	1.00
0.00	0.00	7.83	4.32	11980483.74	3.55	7.30	4.29	1.00	1.00

Abbreviation	Measure
FREQ	observed frequency of word w
RANGE	number of parts with word w
MAXMIN	max. freq. of w/part—min. freq. of w/part
SD	standard deviation of frequencies
VARCOEFF	variation coefficient of frequencies
CHISQUARE	chi-square value of the frequency distribution
D_EQ	Juilland et al.'s D (assuming equal parts)
D_UNEQ	Juilland et al.'s D (not assuming equal parts)
D2	Carroll's D_2
S_EQ	Rosengren's S (assuming equal parts)
S_UNEQ	Rosengren's S (not assuming equal parts)
D3	Lyne's D_3
DC	Distributional Consistency
IDF	Inverse Document Frequency
ENGVALL	Engvall's measure
U_EQ	Juilland et al.'s usage coefficient U (assuming equal parts)
U_UNEQ	Juilland et al.'s usage coefficient U (not assuming equal parts)
UM_CARR	Carroll's U_m
AF_EQ	Rosengren's Adjusted Frequency AF (assuming equal parts)
AF_UNEQ	Rosengren's Adjusted Frequency AF (not assuming equal parts) Ur_KROM
	Kromer's U_R
F_ARF	Savický and Hlaváčová's f_{ARF}
AWT	Savický and Hlaváčová's AWT
F_AWT	Savický and Hlaváčová's f_{AWT}
ALD	Savický and Hlaváčová's ALD
F_ALD	Savický and Hlaváčová's f_{ALD}
SELF_DISP	Washtell's self-dispersion
DP	Gries's Deviation of Proportions
DP_norm	Gries's Deviation of Proportions (normalized)

CHAPTER 3

Pros and cons of single versus multiple-subject experiments: lexical processing in Vietnamese

A revised version of this chapter has been submitted for publication as Hien Pham and Harald Baayen. (2014) Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Submitted for publication.*

Abstract

Although Vietnamese has a long history of linguistic research (see e.g., Trần et al., 1941; Nguyễn, 1963; Thompson, 1965; Nguyễn, 1997), as yet no psycholinguistic studies addressing lexical processing in this language have been carried out. This paper addresses the reading of Vietnamese bi-syllabic compound words, with a methodological focus on the pros and cons of large regression designs with a single subject and small regression designs with many subjects. We show for Vietnamese that a single-subject comprehensive experiment

(with 15,000 bi-syllabic compounds) provides more precise insight into the consequences of language structure for language processing than a multi-subject small-scale experiment. We report the novel finding of an inhibitory instead of a facilitatory (cf. Moscoso del Prado Martín et al., 2004) family size effect, as well as effects relating to the density of the compound graph of Vietnamese, replicating earlier findings reported for English (Baayen, 2010).

Keywords: compounds, family size, Vietnamese, mega-study, generalized additive modeling, shortest path lengths

Introduction

Many, if not most, of the world's languages have insufficient educational and/or scientific infrastructure to afford experimental studies of language processing. Under such circumstances, researchers seeking to initiate an experimental research tradition on language processing face the choice between two opposite experimental designs, a single-subject design with many words (15,000) and a multiple-subject regression design with few words (500).

Single-subject research designs have been used in applied fields of education and psychology. Traditionally, single-subject designs, with a single participant responding to all stimuli across different sessions, have been favored when the goal is to clarify how specific individuals perform (see McReynolds and Thompson, 1986; Horner et al., 2005, for details).

Multiple-subject research designs have been the bread and butter of all psycholinguistic studies of lexical processing that we are aware of. These experiments make use of small numbers of items (typically less than 100, often less than 50) and multiple subjects (typically 20, sometimes fewer, sometimes many more). More recently, so-called mega-studies (Seidenberg and Waters, 1989) have examined large numbers of words (see Spieler and Balota, 1997; Balota et al., 2004, 2007; Ferrand et al., 2010; Keuleers et al., 2010; Yap et al., 2010). Early mega-studies collected response latencies from many different subjects, with any given subject responding to only a small subset of the total item set. More recently, mega-studies with some 20 subjects in which each subject responds to all items have been carried out. Such multi-subject mega-studies require substantial resources, and will often be beyond the means of researchers studying languages with little or no educational infrastructure and no prior history of experimental psycholinguistics.

The present study addresses the question whether a single-subject mega-study might be preferable to a multi-subject study with a much smaller set of items. The language for which we evaluate these two designs is Vietnamese, a language for which, to our knowledge, no previous experimental work on language processing exists. However, Vietnamese is a language with a long tradition of philology, with a research tradition in linguistics (see e.g., Trần et al., 1941; Nguyễn, 1963; Thompson, 1965; Nguyễn, 1997) that has informed the

experiments reported below. Importantly, for a 'new' language, because Vietnamese has a strong internet presence, the first author was able to compile comprehensive lexical resources for Vietnamese (see Pham et al., 2014, for details) that we will make use of in the statistical analyses reported below. These lexical resources were derived from two corpora compiled by the first author, a newspaper corpus and a film subtitle corpus, both comprising 160 million words which have been tokenized and segmentized at the word (compound) level, as well as tagged for part of speech. A large number of lexical-distributional measures were calculated on the basis of these corpora, which made it possible for us to implement regression designs for the single-subject and multiple-subject experiments.

In summary, the present study primarily addresses the question of how to optimize limited financial resources for the initial exploration of a language for which no prior research on language processing is available, but that does have a written tradition (allowing reading research) and an internet presence (allowing the construction of lexical resources). Before discussing our experiments, we first provide a brief introduction to Vietnamese.

Vietnamese

Vietnamese (tiếng Việt), the official language of Vietnam, is spoken by approximately 86 million people in Vietnam (based on figures of the General Statistics Office of Vietnam) and about 4 million speakers living abroad (such as in the U.S., France, Australia, Canada, Germany, the Netherlands, etc.). It belongs to the Việt-Mường sub-branch of the Vietic branch of the Mon-Khmer family, which is itself a part of the Austro-Asiatic family. In this isolating tone language, all syllables are single morphemes and all morphemes are monosyllabic. Vietnamese linguists have introduced the term *syllabeme* to refer to the syllable-morpheme identity (see e.g., Ngô, 1984, for further information on syllabeme). Vietnamese words may consist of one syllabeme (e.g., *cây* 'tree', *gạo* 'rice', *mắt* 'eye') or multiple syllabemes, e.g., *hoa hồng* 'rose' (lit. flower pink), *tàu hoả* 'train' (lit. vehicle fire).

In the present-day writing system of Vietnamese, (*chữ Quốc ngữ*, an alphabetic script introduced by Catholic missionaries from Portugal, France, Spain and Italy), syllabemes are

separated by spaces. The spacing conventions of Vietnamese follow its neighbor China, albeit without using the characters familiar from this country's orthography. The result is a straightforward writing system that enables Vietnamese speakers to learn how to read and write within a few months. It serves as the official orthography nation-wide (Nguyễn, 1997). A syllabeme is written as a sequence of Roman letters, with additional diacritics for distinguishing phonemes that are not properly distinguished by the Roman alphabet, and with additional diacritics for the tones of Vietnamese (*ngang* mid level, *huyền* low falling (breathy), *hỏi* mid falling (-rising), harsh, *ngã* mid rising, glottalized, *sắc* mid rising, tense, and *nặng* mid falling, glottalized, short).

Vietnamese syllables are severely restricted phonotactically, and consist of an optional onset consonant, followed optionally by a bilabial consonant glide, followed by an obligatory vowel (with one of six tones), followed optionally by a single coda consonant. Table 3.1 presents a partition of the most common syllabemes in contemporary Vietnamese. The total number of attested syllabemes in actual use is 6,651. By comparison, the total number of English syllables as attested in the CELEX lexical database for English wordforms (Baayen et al., 1995), differentiated for stress (no stress, primary stress, secondary stress) is 17,918. Without differentiating between stress, the number of different syllables remains substantially larger than in Vietnamese (11,492). This difference between Vietnamese and English will become important below in the discussion of constituent effects in Vietnamese.

Table 3.1: Vietnamese syllable type frequency

Type	Frequency	Example	English gloss
CwV	141	*hoa, quê*	flower, countryside
CwVC	436	*hoang, xoay*	uncultivated, revolve
wV	11	*oà, uỷ*	burst out crying, commissioner
CV	1106	*ngủ, xu*	sleep, coin
wVC	27	*oách, oằn*	dapper, to curve
CVC	4681	*bên, xương*	side, bone
V	50	*ả, ý*	lass, idea
VC	188	*ác, ai*	fierce, anybody

Although almost all syllabemes are independent words, the majority of words in Vietnamese

comprise more than one syllabeme. The focus of the present study is on words consisting of two-syllabemes, which show a similar range of semantic transparency with respect to their constituent syllabemes as do English compounds. In Vietnamese, however, the problem of — often a lack of — semantic transparency is especially acute as any syllable is also in use as a word with its own meaning.

Experiment 1 is an exhaustive experimental survey of all two-syllabeme compounds of Vietnamese listed in a major dictionary (Hoàng, 2000), using the visual lexical decision task.

Experiment 1: Single-subject large-scale lexical decision experiment

Methods

Materials All disyllabic words from the Vietnamese Dictionary (Hoàng, 2000) were selected, with the exception of those words involving reduplication, resulting in a list of target words comprising 15021 words. In addition, nearly 5000 single syllabeme (monomorphemic) words were included, resulting in a total of 20,000 Vietnamese words.

For each word, 18 (highly correlated) corpus-based counts were compiled. These counts included the frequency of occurrence of the two-syllable words in the newspaper corpus and in the subtitle corpus, as well as measures of dispersion in these corpora. Furthermore, corresponding counts were collected for the first and second syllabemes. In addition, the primary (Moscoso del Prado Martín et al., 2004) and secondary (Baayen, 2010) family size counts for the syllabemes were obtained, as well as their dispersion. Finally, additional family size counts were compiled for the constituents, once disregarding only diacritics for tone, and once disregarding all diacritics. As the collinearity of this set of predictors was very high ($\kappa = 610.58$), we orthogonalized them using principal components analysis. A scree plot revealed three primary principal components. The first principal component, henceforth `PC freq`, had large negative loadings for the compound frequency and dispersion measures. Family

size measures, with or without diacritics, had small negative values on this component. The second principal component contrasted family size measures (large negative loadings) with compound frequency and dispersion (large positive loadings). This component is henceforth referred to as `PC freq-fam`. The third principal component, `PC left-right`, contrasted family size measures for the second syllabic constituent (large negative loadings) with family size measures for the first syllabic constituent (large positive loadings). The proportion of the variance captured by the three principal components were 0.37, 0.23, and 0.19.

In addition to the three principal components, the length of the compound (in letters), session number (1–16), and the time of day the block was run (in minutes from midnight) were included as predictors. Furthermore, the lexical tone of the first syllable (1–6) as well as that of the second syllable (1–6), and the word category of the compound were entered into the model as random-effect factors. Table 3.2 presents the distribution of tones. As fixed-effect factors we included whether the first/second syllable constituents are also used as classifiers, and whether the compound is part of a strongly connected component of the Vietnamese directed compound graph.

Table 3.2: Distribution of tones in Vietnamese single-syllabeme and two-syllabeme words.

Tone	Single Syllabeme		First Compound Syllabeme		Second Compound Syllabeme	
	types	tokens	types	tokens	types	tokens
ngang	984	14,130,780	6641	5,059,200	4693	3,443,209
huyền	802	11,543,156	3840	2,586,797	3360	2,295,111
ngã	313	3,314,686	858	386,988	1054	547,700
hỏi	514	5,075,897	2145	1,884,127	2277	1,868,108
sắc	1365	11,823,632	5507	4,128,831	5918	4,015,755
nặng	976	7,218,239	3361	2,784,402	4995	4,560,463

A strongly connected component of a directed graph is a subgraph with the property that each vertex (node) in the graph can be reached from any other vertex by following the directed edges (links). Baayen (2010) studied the directed compound graph of English (restricted to bi-morphemic compounds), i.e., a graph in which compound constituents are the vertices, and in which directed edges connect first constituents to second constituents. The English compound graph has one (large) strongly connected component. The Vietnamese

compound graph is characterized by two (also large) strongly connected components. Within these strongly connected components, cyclic chains exist, as illustrated in Figure 3.1. Each pair of nodes linked by a directed edge represents an existing compound. Compounds in a strongly connected component are part of a particularly dense area of the lexicon. Just as neighborhood density at the segment level may affect lexical processing, neighborhood density at the syllable/constituent level may help explain response latencies. A final related numeric predictor that comes into play only for words in the strongly connected component is the length of the shortest path from second syllabeme to the first. In Figure 3.1, these shortest path lengths are 2, 4, 8, and 10 respectively.

For each of the 20,000 words in the experiment, a pseudoword was generated using the Wuggy pseudoword generator (Keuleers and Brysbaert, 2010). Each pseudoword differed from its reference word by one subsyllabic segment (i.e., the onset, nucleus, or coda) per syllable. As a consequence, a two-syllable nonword differed in two positions from its reference word. A further constraint on pseudoword generation was that the position selected for change was chosen such that it resulted in the smallest possible overall change in syllable frequency, transitional frequency between syllables, and subsyllabic frequency. As a result, the pseudo-morphological structure of the nonwords resembled the morphological structure of the words as closely as possible, as can be seen in Table 3.3.

Table 3.3: Examples of compound words and their equivalent pseudowords. None of psedowords are existing word in Vietnamese.

Word	Pseudoword
ác cảm	ác bạm
á hậu	á đấu
ẩn nấp	ẩm bấp
âm hưởng	âm bượng
áp thấp	áp cháp
nghị sĩ	nghì sự
thể nghiệm	thử nghiêm
vị thế	vù thị
xoắn ốc	xoán óc
xuất viện	xuất tiên

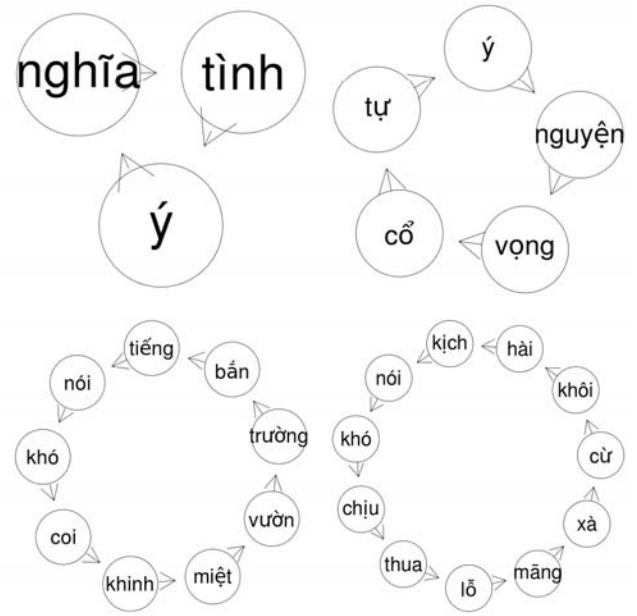

Figure 3.1: Examples of cycles in the compound directed graph: shortest head-to-modifier paths for *ý* → *nghĩa*, *ý* → *nguyện*, *miệt* → *vườn*, and *xà* → *cừ*. English glosses of the compounds for the upper left panel: *nghĩa tình* 'sentimetal attachment', *tình ý* 'intention', *ý nghĩa* 'mean, sense'; for the upper right panel: *ý nguyện* 'wishes', *nguyện vọng* 'aspiration', *vọng cổ* 'name of a traditional tune', *cổ tự* 'ancient writing', *tự ý* 'willingly'; for the lower right panel: *kịch nói* 'play', *nói khó* 'beg', *khó chịu* 'uncomfortable', *chịu thua* 'yield', *thua lỗ* 'lose', *lỗ mãng* 'coarse', *mãng xà* 'python', *xà cừ* 'conch, nacre', *cừ khôi* 'splendid', *khôi hài* 'funny, humorous', *hài kịch* 'comedy'; for the lower left panel: *tiếng nói* 'voice', *nói khó* 'beg', *khó coi* 'unsightly, unaesthetic', *coi khinh* 'despise', *khinh miệt* 'despise, think little and scorn', *miệt vườn* 'hick', *vườn trường* 'school garden', *trường bắn* 'rifle range', *bắn tiếng* 'spread word'.

Subject The first author, a native speaker of Vietnamese, served as the single participant of this experiment. Responding to all forty thousand trials required 46 hours, over a 4-week period.

Procedure All the stimuli, including both words and nonwords, were merged into one list. A script was written to randomly select equal numbers of word and pseudoword stimuli from the list, which were then merged into a template script for DMDX. Thanks to this automated procedure, the participant (who also implemented the experiment) remained completely uninformed about the words to appear in a given experimental session. The total experiment comprised 80 blocks of 500 stimuli. Each block took about 60 minutes to finish (including breaks) and was subdivided into five sub-blocks of 100 stimuli each. Between each sub-block, the participant was asked to press the space bar to continue. The participant felt that the interruptions increased his control and provided him with information about his progress through the block. The participant completed a maximum of two blocks per day.

Stimuli were presented on a 17-in. Acer laptop with a refresh rate of 85 Hz and a resolution of 1,600 x 900 pixels, which was controlled by an Intel Core i7 1.6GHz processor. Stimuli were presented in lowercase 26-point Courier New font, and they appeared as black characters on a grey background. Stimuli were presented and responses collected with the DMDX software (Forster and Forster, 2003).

The participant indicated as quickly and as accurately as possible whether a presented letter string formed a word or not in Vietnamese by pressing a button on a Microsoft USB wired Xbox 360 game controller for Windows with his left (No) and right (Yes) index fingers. Each trial started with a centered fixation point '+' that was presented for 500 msec, followed by the target letter string, which stayed on the screen until the participant responded or until 2 seconds had elapsed. The lexical decision experiment started with 12 practice trials in each session, followed by 500 experimental trials, separated by four breaks.

Results

Response latencies were inverse transformed (-1000/RT) to reduce the skew in their distribution. In order to properly model nonlinear functional relations in two or more dimensions, we made use of generalized additive mixed-effects regression models GAMMs, (see, e.g., Hastie and Tibshirani, 1990; Wood, 2006) as implemented in the `mgcv` package (Wood, 2006, 2011) of the `R` statistical computing software (R Core Team, 2013). For detailed description and worked examples of the use of generalized mixed-effects additive models in psycholinguistics, see Baayen (2014); Baayen et al. (2010); Tremblay and Baayen (2010); Matuschek et al. (2012); Kryuchkova et al. (2012) and Balling and Baayen (2012), and for applications in linguistic studies, Wieling et al. (2011); Kösling et al. (2013) and Tomaschek et al. (2013). For the modeling of wiggly curves, thin plate regression splines were used, and for the modeling of wiggly surfaces, we made use of tensor products.

Table 3.4: Generalized Additive Model fitted to the inverse transformed lexical decision latencies of the large single-subject study (edf: estimated degrees of freedom).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.5829	0.0477	-33.1898	< 0.0001
Word Length	0.0160	0.0014	11.0923	< 0.0001
ICC: True	-0.0651	0.0352	-1.8486	0.0645
B. smooth terms	edf	Ref.df	F-value	p-value
smooth PC frequency	4.0473	5.0885	277.9798	< 0.0001
smooth PC freq-fam: ICC False	1.0000	1.0000	207.1894	< 0.0001
smooth PC freq-fam: ICC True	3.8749	4.8666	160.2241	< 0.0001
smooth Minutes	4.3712	5.0122	38.3893	< 0.0001
random effect tone of first syllable	4.0966	5.0000	7.2894	< 0.0001
random effect tone of second syllable	4.1705	5.0000	4.9090	0.0001
random effect word category	7.7133	10.0000	11.7848	< 0.0001
smooth Session number	8.4037	8.7715	41.3732	< 0.0001
smooth PC left-right syllable	3.5894	4.5488	3.6806	0.0038

Tables 3.4 and 3.5, and Figures 3.4 and 3.2 (upper panels) summarize the generalized additive mixed model fitted to the inverse-transformed response latencies. Longer words tended to elicit longer latencies. A marginal facilitatory main effect of membership in the strongly connected component (SCC) went hand in hand with strong evidence for modulation of the

effect of PC freq-fam by membership in the SCC. As shown in the second and third upper panels of Figure 3.4, the effect of the frequency-family contrast was stronger for words belonging to the SCC. When the syllabemes of a compound have larger families, and when these families belong to highly interconnected sections of the compound graph, response latencies apparently become progressively longer. When separate predictors for constituent frequencies are considered, they give rise to inhibitory effects (models not shown). In English, one typically finds facilitation, or facilitation that is modulated by compound frequency (Kuperman et al., 2008, 2009). Likewise, family size effects tend to be facilitatory and not inhibitory (see, e.g. Moscoso del Prado Martín et al., 2004). This raises the question why in Vietnamese, family size (and constituent frequency) effects become inhibitory.

We suspect that the strong phonotactic restrictions on syllabemes are at issue here, resulting in a relatively small set of individually meaningful constituents that are 'recycled' in compounds of varying degrees of transparency, and that are written with intervening spaces. From a discrimination learning perspective (Ramscar and Baayen, 2013; Baayen et al., 2011), discriminating between the meanings of the constituent syllabemes and the meanings of the compounds is harder compared to English, because there is more overloading of the constituents.

The upper left panel of Figure 3.4 visualizes the frequency effect. As expected, higher-frequency words (more negative values on PC freq) elicited shorter latencies, with a slight leveling off for the highest-frequency words. The upper right panel shows the non-linear, U-shaped effect of PC left-right. Recall that large negative values on this principal component reflect large families for the second syllable, whereas large positive values reflect large families for the first syllable. Apparently, when the families are out of balance, i.e, when the one family is large at the expense of the other, then responses are delayed. Processing appears to be optimal when both families are in balance (i.e., when PC left-right assumes values around zero).

Table 3.4 indicates that all three random-effect factors (the tone on the first syllable, the tone on the second syllable, and word category) contribute significantly to the model fit (all $p < 0.0001$). Inspection of the posterior models for the tone of the first syllabeme shows that

the *huyền* low falling (breathy) and *sắc* mid rising, tense tones elicited longer latencies than the other four tones. With respect to the second syllabeme, the *ngã* mid rising, glottalized tone elicited the shortest latencies, and the *huyền* low falling (breathy) and *ngang* mid level tone the longest. The major word categories (noun, verb, adjective) were responded to more quickly than the minor word categories.

Finally, smooths for the time of day at which the experiment was run (`Minutes`) and session number (`Session`) were well supported. Their partial effects are shown in Figure 3.2. The upper left plot shows that responses were faster in the afternoon than in the morning. The upper right plot shows that in the course of this month-long experiment, responses were elongated at the beginning and halfway through the experiment, and that towards the end of the experiment, responses were shorter. We were not able to find any interactions involving these two predictors that would improve the model fit. We also could not detect any effects of `Trial` (the rank of an item in its experimental list).

Table 3.5: Reduction in AIC as predictors are added to an intercept only baseline model for the single-subject dataset

models	AIC
+ Minutes and Session	586.15
+ Tone1 and Tone2	91.24
+ Word Category	21.27
+ Word Length	43.92
+ PC frequency	1817.84
+ PC freq-fam:ICC	947.78
+ PC first-sec syll	11.67

Table 3.5 lists the decrease in AIC when, starting with an intercept-only model, predictors or groups of predictors, are added to the model formula. The most important predictor is `PC freq`, unsurprisingly, as it captures the word frequency effect. The second most important predictor is `PC freq-fam`, which contrasts words with large families and low frequency with high-frequency words with small families. Next in importance are the experimental variables MINUTES and SESSION. As expected for a language rich in tones, the two tone random effect factors also contribute substantially to the goodness of fit. Contributions of the remaining

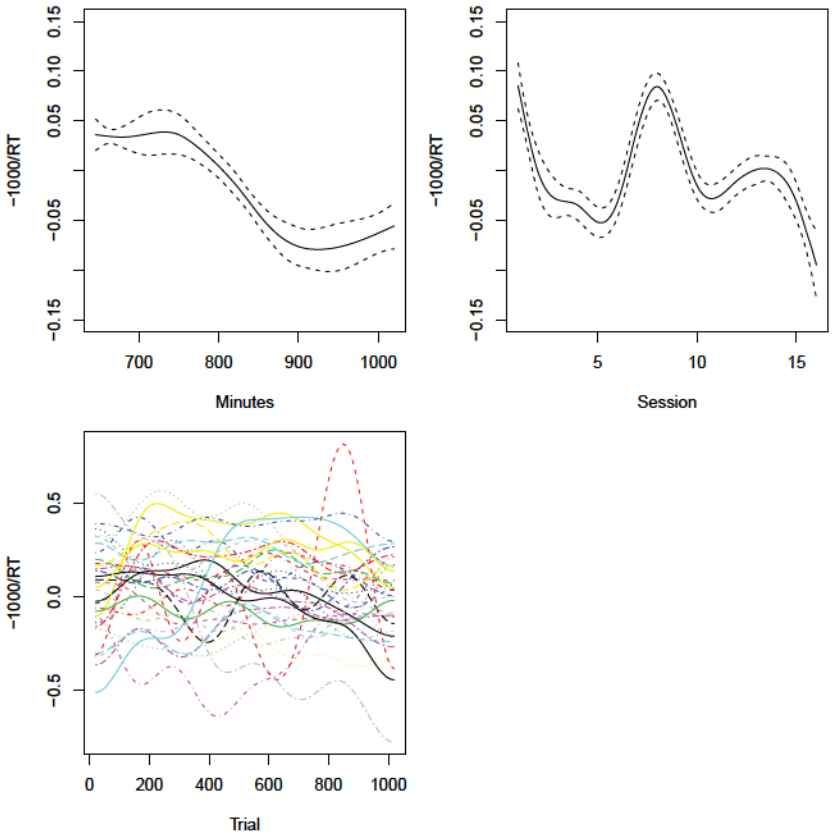

Figure 3.2: Partial effects of time of day (in minutes past midnight) and session number for the single-subject experiment, and by-participant random smooths for Trial for the multi-subject experiment.

predictors were modest.

Table 3.6 presents a generalized additive mixed model fitted to the subset of compounds that are part of the strongly connected component of the compound graph (11392 of the 15021 observations). For these compounds, the length of the shortest path from head to modifier is of potential relevance. When the shortest path length is included as predictor, `PC left-right` loses significance, and interactions emerge with whether the second syllable-constituent is also in use as a classifier. For those compounds with a second constituent that is not also a classifier, and only for these compounds, an interaction of `PC-freq` by shortest path length was present, as revealed by the tensor product smooth shown in Figure 3.3.

Table 3.6: Generalized Additive Model fitted to the inverse transformed lexical decision latencies of the large single-subject study, restricted to the words in the strongly connected component of the compound graph (edf: estimated degrees of freedom).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.6509	0.0393	-41.9976	< 0.0001
Word Length	0.0161	0.0017	9.6701	< 0.0001
Second Syl. is Classifier: TRUE	-0.0115	0.0179	-0.6403	0.5220
B. smooth terms	edf	Ref.df	F-value	p-value
smooth PC frequency	3.3490	4.2718	167.2776	< 0.0001
smooth PC freq-fam	3.8493	4.8373	152.8841	< 0.0001
smooth Minutes	3.8974	4.5590	29.4380	< 0.0001
random intercepts tone of first syllable	3.9878	5.0000	4.8267	0.0020
random intercepts tone of second syllable	4.3135	5.0000	5.5958	< 0.0001
random intercepts word category	7.4343	10.0000	7.3111	< 0.0001
smooth Session	8.1698	8.6756	31.4299	< 0.0001
smooth shortest path length	1.0000	1.0000	39.6244	< 0.0001
tensor smooth Sh. Path by PCfreq: 2nd is Class: FALSE	2.8869	3.5853	3.0730	0.0199
tensor smooth Sh. Path by PCfreq: 2nd is Class: TRUE	1.0000	1.0000	1.1487	0.2838

Figure 3.3 presents the fitted surface as a function of `Shortest Path Length` and `PC freq`. Greener colors denote short latencies, pink to white colors denote longer latencies. As on a terrain map, contour lines connect points that have the same vertical height. Contour lines are 0.05 units apart on the -1000/RT scale.

For this GAMM model, we adopted a decompositional approach with separate smooths for `Shortest Path Length` and `PC freq`, combined with a tensor smooth for the partial effect of the interaction of these two predictors. (Inclusion of the interaction smooths for compounds with second constituents differentiated by their classifier status reduced the AIC by 4.3.) Figure 3.3 shows that for high-frequency words (large negative values of `PC freq`), the effect of path length is small, with an optimum of shortest responses around paths of length 2–4. As frequency decreases (larger, positive values of `PC freq`), the effect of path length reverses, such that for the lowest frequency words, lengths 4–6 are least optimal, with the longest response latencies. In other words, the word frequency effect is strongest for compounds with a shortest path length of 4–5 — for these two path lengths, the greatest number of contour lines is crossed in Figure 3.3 when moving horizontally along the Y-axis.

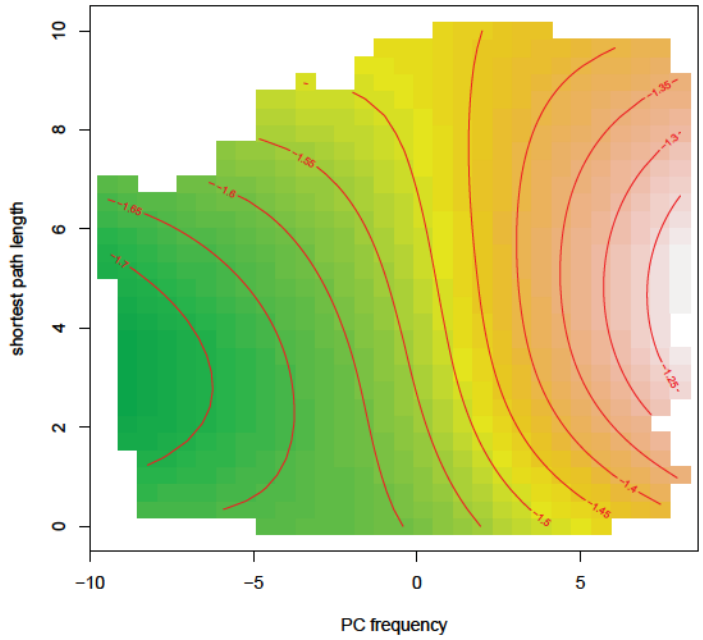

Figure 3.3: Tensor product surface for the interaction of Shortest Path Length and PC freq for compounds the second constituent of which is not in use as a classifier, in the single subject experiment.

The modulation of shortest path length by frequency is very similar to the interaction of shortest path length by first constituent family size reported in Baayen (2010) for word naming in English. Interactive activation theories might explain the observed pattern as resulting from activation spreading from the second constituent through the compound graph and ultimately returning to the first constituent, resulting in confusion about the functional status of the first constituent (e.g., modifier in the target compound, but head of the previous compound in the compound chain). This confusion would then arise primarily for low-frequency compounds and intermediate path lengths. For short paths, activation would arrive back too early to interfere, at a time when there still is strong bottom-up support. For long paths, activation would have decayed too much to cause strong interference (see Baayen, 2010, for further discussion).

Whereas the graph-theoretical effects observed for Vietnamese converge with similar effects observed for English, the sign of the effect of `PC freq-fam` is specific to Vietnamese also in this subanalysis.

Experiment 2: Multiple-subject small lexical decision experiment

Experiment 1 made use of a design with a comprehensive coverage of two-syllabeme compounds, but only a single subject. Running a multi-subject experiment of the same scope is impossible without either a pre-existing multi-university research tradition (as for the English Lexicon Project), or extensive funding. What is more feasible in many situations is a multi-subject experiment with a small set of items. Experiment 2 was run in Vietnam with 33 participants, and 550 words (and 550 nonwords). The number of items was chosen to provide as extensive coverage as possible within a single experimental session of approximately one hour.

Methos

Materials 550 disyllabic compounds were randomly selected from the 15,000 items in the single-subject experiment, such that high- and low-frequency compounds had an equal chance of being selected.

Subjects Thirty three students at the Vietnam National University were recruited to take part in the lexical decision experiment (mean age 21.9, range 20 – 22 years, 12 males, 21 females). All participants were native Vietnamese speakers and had at least 14 years of education.

Procedure The same experimental equipment was used as in Experiment 1. Eight lists, each with the items in a different random order, were constructed for counterbalancing; subjects were randomly assigned to these lists. The experiment was administered in the same way as a block in Experiment 1.

Results

Table 3.7 summarizes the generalized additive mixed model fitted to the inverse-transformed response latencies. In addition to the random effect factors for tone and word category, item and subject were included as random-effect factors. For subjects, random slopes for `PC frequency` and `PC freq-fam` were found to be justified as well. The most important random-effect component turned out to be by-subject random smooths for `Trial`, shown in the lower panel of Figure 3.2. As the experiment proceeded, subjects' performance fluctuated substantially, and non-linearly. Although for some subjects, these fluctuations were mild, other subjects showed performance that changed substantially. One subject started out as the slowest subject, but by the end of the experiment responded fastest, possibly indicating an effect of habituation to the task. Conversely, the subject starting out as the fastest responder became one of the slowest responders in the second half of the experiment. One subject revealed a highly oscillatory pattern, with tremendous slowing down, followed by speeding, up, in the last quarter of the experiment. The factor smooth for `Trial` by `Subject` is by far the most important predictor for the response latencies in this experiment (see Table 3.8 below).

Generally, the effects observed in the multi-subject experiment mirror those for the single-subject experiment. The effects of the tone of the second syllable, and of word category, are lost, probably due to a lack of power, and the same holds for the main effect of membership of the strongly connected component. An effect of `PC left-right` is also captured, but only for words in the strongly connected component. The lower panels of Figure 3.4 present the partial effects of the principal components. The main differences with the single-subject experiment (upper panels) concern the reduced effect of `PC freq-fam` for words with a second constituent in the strongly connected component (third column), and the qualitative change in the effect of `PC left-right` from U-shaped to linear.

As for the single-subject experiment, we investigated the contributions of the predictors (or groups of predictors) in terms of the extent to which they contributed to reducing the AIC of the model. Table 3.8 indicates that subject and item variability dwarfs the linguistic

Table 3.7: Generalized Additive Model fitted to the inverse transformed lexical decision latencies of the smaller-scale multiple-subject study

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.7505	0.0728	-24.0342	< 0.0001
Word Length	0.0067	0.0040	1.6998	0.0892
In SCC: TRUE	-0.0251	0.0442	-0.5672	0.5706
B. smooth terms	edf	Ref.df	F-value	p-value
smooth PC frequency	2.3021	2.4940	46.2649	< 0.0001
smooth PC freq-fam: ICC FALSE	1.4554	1.5540	5.7468	0.0076
smooth PC freq-fam: ICC TRUE	1.0004	1.0005	17.4312	< 0.0001
random intercepts tone of first syllable	3.5901	5.0000	42.8616	< 0.0001
random intercepts tone of second syllable	0.1776	5.0000	0.0572	0.3677
random intercepts word category	0.8549	3.0000	4.6191	0.0821
smooth PC left-right syll: ICC FALSE	1.0001	1.0001	0.5445	0.4606
smooth PC left-right syll: ICC TRUE	1.0006	1.0007	9.6210	0.0019
random intercepts Item	367.1179	534.0000	2.3184	< 0.0001
random intercepts Subject	15.7220	32.0000	0.9690	< 0.0001
random by-subject slopes for PC freq-fam	25.6535	32.0000	11.4338	< 0.0001
random by-subject slopes for PC frequency	26.9802	32.0000	13.8741	< 0.0001
by-subject random smooths for Trial	232.2945	296.0000	2779.4477	0.0008

Table 3.8: Reduction in AIC as predictors are added to an intercept only baseline model, for the multiple-subject data

models	AIC
+ Trial by Subject	7416.76
+ subject random intercepts and slopes	2177.01
+ item random intercepts	955.81
+ tones	3.35
+ Word Category	0.37
+ Word Length	0.18
+ PC frequency	5.02
+ PC freq-fam:ICC	2.40
+ PC first-sec syll	2.01

predictors. This pattern is strikingly different from that observed for the single-subject experiment, for which the first two principal components (`PC frequency` and `PC freq-fam`, in interaction with membership in the strongly connected component) effected the greatest changes in AIC.

An analysis of the subset of words with a second constituent in the strongly connected component was carried out to inspect whether the interaction of `Shortest Path Length` by `PC freq` by the second constituent being in use as a classifier would persist (model not shown). This interaction was again present, and as before, it was restricted to those compounds with a second constituent that is not in use as a classifier. The fitted tensor surface is shown in the second panel of Figure 3.5, next to the corresponding tensor surface for the single-subject experiment. Due to data sparsity, the tensor for the multi-subject experiment captures only the bottom half of the effect that emerges from the single-subject experiment (which has 30 times as many items).

Finally, we note that the general inhibitory effect of family size in Vietnamese replicated well in Experiment 2.

General Discussion

This study reports what is — to our knowledge — the first experimental study of Vietnamese. When embarking on the experimental study of a new language, a practical question that arises when resources are limited is whether to conduct a large study with one, or only a few, participants, or whether to conduct a study with more participants and fewer items.

A single-subject large-scale experiment offers the advantage of statistical power with respect to lexical properties, but it raises the question of whether the single subject is a typical, representative subject, and of whether variation in when the experiment was administered may have skewed results. With respect to the latter question, we note that although we did observe strong effects of time of day, and of experimental session, we were unable to detect interactions of the lexical variables with these predictors. For instance, we have no evidence

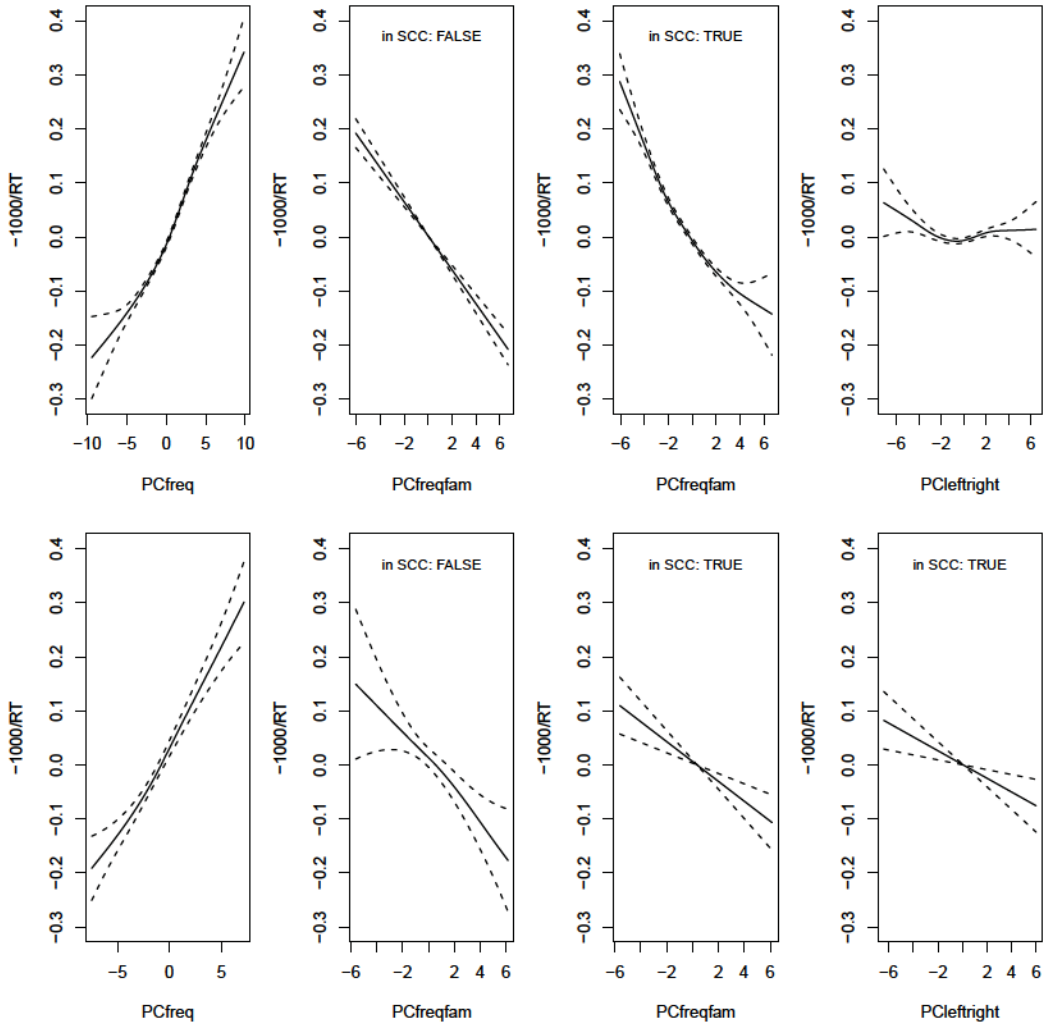

Figure 3.4: Smooths for lexical predictors for the single-study data (top) and the multiple-subject data (bottom).

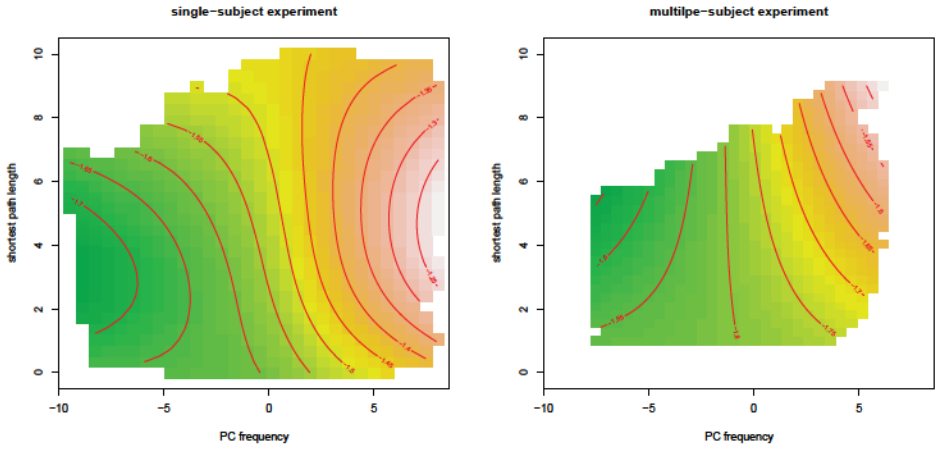

Figure 3.5: Tensor product surface for the interaction of Shortest Path Length and PC freq for compounds the second constituent of which is not in use as a classifier, in the single-subject (left) and multi-subject (right) experiment.

that the frequency effect would vary with time of day. From this, we conclude that although time of day, and session number, will co-determine response latencies, their effect is benign, in the sense of independence vis-a-vis the effects of linguistic predictors.

As to the question of the representativeness of a single participant, no general conclusion can be drawn. All we can say is that for our study, the single-participant performed the experiment much more consistently than the subjects in the multi-participant study. This is apparent from Figure 3.2, when the top panels (on a scale of -0.15 to 0.15) are compared with the bottom panel (on a scale of -1 to +1). The between-subject variation in the multi-participant experiment (bottom panel) within a single (and only) session is much larger than the between-session variation observed for the single participant study. This dovetails well with the observation that in the single-subject experiment, linguistic predictors were most successful in reducing AIC, whereas in the multi-subject experiment, the subject (and also item) random effects accounted for AIC values that were two or three orders of magnitude greater than those tied to the linguistic predictors. A comparison of adjusted R-squared values for the single subject experiment without the experimental predictors (`Minutes, Session`), 0.17, and the multiple-subject experiment, 0.13, obtained by taking the adjusted

R-squared of the full model and subtracting the adjusted R-squared for a model with random by-subject smooths for Trial only, indicates that the single-subject experiment performs better in capturing variance due to linguistic variation.

An important caveat here is that for the multi-subject experiment, participants were recruited with a high level of education. As shown by (Kuperman and Van Dyke, 2011, 2013), substantial differences exist in reading skills (and reading habits) as a function of education and vocation. These differences, although of great importance for education, are, to the extent of our knowledge, completely ignored in current mega-studies of reading, which thus far have all addressed highly-skilled readers. The only way in which these important individual differences can be assessed is through multi-subject experiments. A key issue that needs more careful thought than has been usual in psychology is what the subject population is from which subjects are sampled. The results of the present study suggest that the data obtained from a single, dedicated subject in a mega-study are of at least the same, if not higher quality compared to the data offered by a multi-participant study with informants with roughly the same general level of education. We think that for languages with few speakers, or languages with few speakers with the necessary metalinguistic skills required for most psycholinguistic behavioral paradigms, a comprehensive single-subject may therefore be an excellent solution. A multi-subject study with the convenience subject samples traditionally used in psycholinguistics do not provide better insight in variation in language processing across full populations of speakers in a language (across age, education, the sexes, occupation, etc.), and only seem to come with more noise and less power.

In the present study, the interaction of frequency by shortest path length in the compound graph illustrates well the advantage of having more items. In our single-subject mega-study, the large number of items (15,000) provides the analyst with the full range of variation on the two dimensions of frequency and shortest path length, enabling the generalized additive model to provide an optimal estimate of the true form of the interaction. In the multi-subject study, the reduction in the number of items, due to random sampling of only 3.3% of the items in the mega-study, led to sparse sampling on both dimensions, with a loss of information about especially high-frequency compounds with longer shortest path lengths.

As a consequence, the GAMM for the multi-subject experiment captures only the 'bottom half' of the true effect, and misses completely that the pattern reverses for the 'top half' (Figure 3.5). Fortunately, the consequences of incomplete sampling do not appear to marr the other predictors in the present study. For instance, the surprising *inhibitory* effect of family size in Vietnamese replicated, with only minor variations, in Experiment 2. Unfortunately, when running experiments with small numbers of items, one will never know whether the patterns observed for such small samples, especially when non-linearities are involved, will replicate in larger samples. If resources are limited, and a single-subject mega-study can be run, it offers substantial advantages for understanding the interplay of language structure and language processing.

References

Baayen, R. H. (2010). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In Olsen, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H. (2014). Multivariate statistics. In Podesva, R. and Sharma, D., editors, *Research Methods in Linguistics*, pages 337–372. Cambridge University Press, Cambridge.

Baayen, R. H., Kuperman, V., and Bertram, R. (2010). Frequency effects in compound processing. In Scalise, S. and Vogel, I., editors, *Cross-Disciplinary Issues in Compounding*, pages 257–270. Benjamins, Amsterdam/Philadelphia.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balling, L. and Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133:283–316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496.

Forster, K. I. and Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers*, 35(1):116–124.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, CRC Press.

Horner, R. H., Carr, E. G., Halle, J., Mcgee, G., Odom, S., and Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71.

Hoàng, P., editor (2000). *Từ điển tiếng Việt [Vietnamese Dictionary]*. Khoa học Xã hội, Hà Nội. Viện Ngôn ngữ học.

Keuleers, E. and Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behaviour Research Methods*, 42(3):627–633.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behaviour Research Methods*, 42(3):643–650.

Kryuchkova, T., Tucker, B. V., Wurm, L. H., and Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: evidence from electroencephalography. *Brain and language*, 122(2):81–91.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23:1089–1132.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of polymorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35:876–895.

Kuperman, V. and Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, 65:42–73.

Kuperman, V. and Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):802.

Kösling, K., Kunter, G., Baayen, H., and Plag, I. (2013). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech*.

Matuschek, H., Kliegl, R., and Holschneider, M. (2012). An Explicit ANOVA-like decomposition of Thin Plate Splines.

McReynolds, L. V. and Thompson, C. K. (1986). Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. *The Journal of speech and hearing disorders*, 51(3):194–203.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

Ngô, T. N. (1984). The syllabeme and patterns of word formation in Vietnamese. New York University dissertation.

Nguyễn, D. H. (1997). *Vietnamese: Tiếng Việt không son phấn*. John Benjamins, Amsterdam.

Nguyễn, K. T. (1963). *Nghiên cứu về ngữ pháp tiếng Việt [Studies in Vietnamese grammar]*. Nhà xuất bản Khoa học, Hà Nội.

Pham, H., Tucker, B., and Baayen, H. (2014). Constructing two Vietnamese corpora and building a lexical database. *Manuscript*.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramscar, M. and Baayen, R. H. (2013). Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Language Sciences*, page doi: 10.3389/fpsyg.2013.00233.

Seidenberg, M. S. and Waters, G. S. (1989). Word recognition and naming: A mega study. *Bulletin of the Psychonomic Society*, 27:489.

Spieler, D. H. and Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, 6:411–416.

Thompson, L. (1965). *A Vietnamese grammar*. The University of Washington Press, Seattle.

Tomaschek, F., Wieling, M., Arnold, D., and Baayen, H. (2013). Word frequency, vowel length and vowel quality in speech production: An ema study of the importance of experience. *Language and Speech*. To appear in *Interspeech 2013*.

Tremblay, A. and Baayen, R. H. (2010). Holistic Processing of Regular Four-word Sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In Wood, D., editor, *Perspectives on formulaic language: Acquisition and communication*, pages 151–173. The Continuum International Publishing Group, London.

Trần, T. K., Phạm, D. K., and Bùi, K. (1941). *Việt-Nam văn-phạm [Vietnamese grammar]*. Lê Thăng.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613.

Wood, S. N. (2006). *Generalized additive models*. Chapman & Hall/CRC, New York.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Yap, M. J., Liow, S. J. R., Jalil, S. B., and Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42(4):992–1003.

CHAPTER 4

Morphological effects in reading aloud Vietnamese compounds

This chapter has been submitted for publication as Hien Pham, Benjamin Tucker, and Harald Baayen. (2014) Morphological effects in reading aloud Vietnamese compounds. *Submitted for publication.*

Abstract

This mega-study reports a word naming experiment addressing the production of Vietnamese compounds. Instead of considering as response variable only the naming latency, we also investigate as response variable the acoustic duration of the speech produced. Effects of compound frequency, word length, and constituent family size were present in both latencies and durations, but effects of constituent frequency were absent. This sets Vietnamese apart from languages such as English and Dutch, for which constituent frequency effects are well attested. We attribute the absence of constituent frequency effects to bisyllabic structures constituting the basic, unmarked phonological form of the Vietnamese word. Our data also challenge models of speech production holding that the onset of speech production would

not be affected by the properties of non-initial syllables, as we observed naming latencies to be co-determined by the family size, tone, and syllable type of the second syllable. A remarkable convergence of the effects of frequency and family size for response latencies and acoustic durations validates the word naming task combining these two response variables as an excellent experimental paradigm for the study of phonological encoding in speech production.

Keywords: lexical naming, frequency, lexical density, acoustic duration, Vietnamese, mega-study, speech production

Introduction

Vietnamese is a syllable-timed tone language of the Mon-Khmer family, well known in the linguistic typology literature for being a morphologically isolating language. Vietnamese has no inflection nor derivation, but compounding is ubiquitous. Most Vietnamese words consist of two so-called syllabemes, phonotactically highly restricted syllables that simultaneously function as the basic phonological and the basic morphological units. Because many syllabemes can also be used as independent onomasiological units, albeit typically infrequently,[1] two-syllabeme words pose similar questions about morphological processing as do compounds in Germanic languages such as English, Dutch, and German. Is lexical processing in Vietnamese decompositional, in the sense that access in comprehension and production is, in some sense, mediated by the constituent syllabemes? Or is lexical processing in this language more holistic, and is the pivotal role in lexical processing reserved for the unmarked form of a Vietnamese word, the two-syllabeme compound?

Although Vietnamese has a rich history of linguistic research (e.g., Trần et al., 1941; Nguyễn, 1963; Thompson, 1965; Nguyễn, 1976, 1981, 1985, 1997, 2011), there are no experimental studies of lexical processing in this language, with the exception of Pham and Baayen (2014). These authors report a large-scale visual lexical decision experiment. This experiment revealed opposite effects for compound frequency and constituent frequency and family size (Schreuder and Baayen, 1997; Moscoso del Prado Martín et al., 2004) counts. Higher frequency compounds elicited shorter responses, but compounds with a high constituent frequency and/or family size elicited elongated responses. This pattern of results is very different from that typically reported in regression studies for languages such as English and Dutch (Kuperman et al., 2008, 2009; Baayen et al., 2010), where especially the left constituent's frequency typically predicts shorter response latencies. The results obtained for Vietnamese are consistent with an interpretation of the bi-syllabeme word being the

[1] Calculations on the Vietnamese contemporary dictionary (Hoàng, 2000) indicate that although 81.5% of syllabemes can be used as monosyllabic words, only 15% of all words consist of one syllabeme only. The remainder of the lexicon is comprised of 71% disyllabic words, 13% three-to-four syllabic words, and 1% five-syllabic words.

unmarked, phonologically preferred form of the Vietnamese word.

The present study aims to provide further clarification of the role of the compound and its constituents in lexical processing, but targeting speech production rather than language comprehension. To test this, we opted for a word naming task. Word naming requires an initial comprehension process, the reading of the visually presented input, which is followed by processes involved in preparing for and executing articulation. Word naming is vulnerable to the criticism that as a hybrid of comprehension and production, it provides little insight into the details of speech production. To assess this criticism, we collected not only naming latencies, but also the acoustic durations of the words read aloud. A central question, then, is whether naming latencies and acoustic durations reveal similar or dissimilar functional dependencies on lexical-distributional properties such as frequency and morphological family size on the one hand, and on intrinsic properties such as lexical tone and syllable structure on the other hand. The more the functional dependencies for latencies and durations diverge, the more likely it is that the naming latencies reflect comprehension, and the acoustic durations reflect speech production. The more similar the two are, the more the naming task is validated as a production task. To see this, note that the acoustic durations reflect the speed with which articulation took place. This speed is likely to be determined primarily by the processes preparing for and executing articulation. Naming latencies provide information only on the processes preparing for articulation. The more these latencies functionally approximate the durations, the more likely it is that the naming latencies are not dominated by initial comprehension processes.

We opted for a single-subject mega-study (for multiple-subject mega-studies, see, e.g., Balota et al., 2007; Keuleers et al., 2010; Yap et al., 2010) of Vietnamese, given logistic constraints on running Vietnamese subjects in Canada on the one hand, and good experiences with a single subject mega-study vis-à-vis a smaller-scale multiple-subject design (Pham and Baayen, 2014). In our experience, a broad coverage of the lexicon is essential for a proper assessment of the interplay of the different factors interacting during lexical processing.

For the analysis of the data harvested with this single-subject regression design, we make use of generalized additive mixed models (GAMMs). Our regression design allows us to address a

wide range of issues with comprehensive statistical modeling. By investigating whole-word versus constituent effects, the role of decompositional processes in speech production can be addressed. The model of Levelt et al. (1999) predicts frequency effects of a complex word's constituents, and the absence of whole-word frequency effects. For Dutch, these predictions appear to be correct, although there are hints of whole-word effects in some experiments (e.g., Bien et al., 2005, 2011). However, whole-word frequency effects in speech production using the picture naming task have also been reported (Janssen et al., 2008). Given the results previously obtained for Vietnamese using the lexical decision task, it seems unlikely that we will find strong constituent frequency effects for Vietnamese in speech production.

With the present mega-study, we targeted two further issues in the speech production literature. The first issue concerns the role of phonological neighbors in speech production. The WEAVER^{++} model of Levelt et al. (1999) predicts the absence of neighborhood effects in speech production. However, neighborhood effects in speech production have been reported, although results have been somewhat inconsistent (Vitevitch, 2002; Vitevitch and Stamer, 2006; Gahl et al., 2012; Sadat et al., 2014). Given the highly restricted phonotactics of Vietnamese syllabemes, neighborhoods are expected to be dense, which leads us to expect delayed selection of the proper syllabemes in speech production. In the present study, we operationalized neighborhood density by means of the Levenshtein distance calculated over the IPA transcriptions of the compounds.

		T o n e		
		R i m e		
Onset		Glide	Nucleus	Coda

Figure 4.1: Vietnamese syllable structure

The second issue concerns the degree to which speech production is modularized. According to the model of Levelt et al. (1999), the phonological encoding of morphemes proceeds sequentially, beginning with the first morpheme. The model makes the strong prediction that the planning of the initial morpheme is not influenced by subsequent morphemes. The implicit priming task shows, for instance, that knowledge of non-initial segments or mor-

phemes does not facilitate production. Co-articulation processes across syllable boundaries (Bell-Berti and Harris, 1979; Mok, 2010), however, challenge such highly modularized procedures. In the case of Vietnamese, speakers have to realize complex F_0 trajectories for the tones of the syllabemes, and it is conceivable that the planning of these trajectories (and possibly tone sandhi processes cannot be postponed, see e.g., Nguyễn and Ingram, 2007), and instead require advanced planning. Given the simple phonological structure of Vietnamese syllabemes, as shown in Figure 4.1, it might even be the case that the syllable structure of the second syllabeme is already being planned during the planning of the first syllabeme. Specifically, the Vietnamese syllable structure follows the following scheme:

$$(\mathrm{C1})(\mathrm{w})\mathrm{V}(\mathrm{C2})+\mathrm{T} \tag{4.1}$$

where *C1* is the initial onset consonant; *w* is the bilabial approximant glide /w/; *V* is the vowel nucleus; *C2* is the final coda consonant; *T* is the superimposed tone.

Experiment

Method

Materials. This study examined Vietnamese disyllabic words. Almost all stimuli were taken from the *Từ điển tiếng Việt* 'Vietnamese Dictionary' (Hoàng, 2000). We first included all the disyllabic words in this dictionary, resulting in a total of 16,883 compound words. From this list, we removed all reduplicative words. The resulting set of target words, which comprises of 13,999 compound words, covered all parts of speech. For each word in the list, we obtained a wide range of lexical-distributional variables, including word length, the frequency of occurence and dispersion of the compounds in a newspaper corpus, the number of phonological neighbors, and the frequency, dispersion, and family size counts (Baayen, 2010; Moscoso del Prado Martín et al., 2004; Pham et al., 2014) for the first and the second syllabemes. In addition, the tone realized on the first and second syllabeme, as well as the

syllable types of the syllabemes were added as factorial predictors. As controls for voicekey artifacts, we included manner and place of articulation of the first segment, as well a factor indicating whether the first segment was voiced or voiceless. Predictors pertaining to the compound graph such as shortest path lengths, which Pham and Baayen (2014) found to be predictive for visual lexical decision in Vietnamese, did not reach significance in the present study, and will therefore not be discussed further. The analyses reported below are based on the 8875 words for which all lexical distributional statistics were available to us, for which the voicekey was triggered properly, and for which the speech was succesfully recorded.

Participants

The first author was the only participant in this experiment. He required 48 hours to complete the experiment, at his own pace, over a 6-week period.

Apparatus

All the stimuli were merged into one list. A script was written to randomly select 400 stimuli from the list, then merge those stimuli into a template script for DMDX in which 10 practice trials were filled at the beginning of the experiment for accomodating the participant and also for checking that the recording system works properly. By doing so, the participant, who was also the experiment designer, remained completely uninformed about the actual contents of each subexperiment. The list for a given subexperiment was also selected randomly from the set of lists.

Stimuli were presented on a 21-in. Dell computer screen with a refresh rate of 60 Hz and a resolution of 1440 x 900 pixels; the computer was controlled by an Intel 3.6GHz processor. Stimuli were presented in lowercase 26-point Courier New font, and they appeared as black characters on a grey background. Stimuli were presented and responses collected with the DMDX software (Forster and Forster, 2003).

The experiment was run in a double-walled sound attenuated booth. Speech was recorded by DMDX via an Alesis Multimix8 USB 2.0 system, which acted as an external sound card and

provided phantom power for a Countryman E6 Earset microphone. During the experiment, the microphone was worn on the participant's ear and a boom extending the capsule to the participant's mouth. The microphone's capsule was placed at approximately 2 cm from the corner of the participant's mouth. The participant sat at an approximate distance of 60 cm from the monitor.

Procedure

In an experimental session, the first author read aloud 400 words presented in random order. A session took about 60 minutes (including breaks) and was subdivided into four blocks of 100 stimuli each. Between each block, the first author pressed the space bar to continue. These interruptions provided the participant with information about his progress through the session. The participant completed a maximum of two sessions per day.

The first author read as quickly and as accurately as possible the word presented on the screen. A trial started with a centered fixation point '+' that was presented for 500 msec, followed by the target letter string, which remained on the screen until the participant responded or until two seconds had elapsed. A voice trigger was used to obtain naming latencies. Furthermore, all responses were recorded and processed with Praat (Boersma and Weenink, 2012) using a script implementing an intensity analysis to extract onset latency and the acoustic duration of the word read out loud.

Results

Mispronunciations, responses contaminated with coughs, and responses lost due to equipment failure, as well as words that are missing lexical distributional statistics are removed from the dataset. The resulting dataset contained 8875 compound words. We analyzed both the naming latencies and the acoustic durations of the words read out loud by the participant.

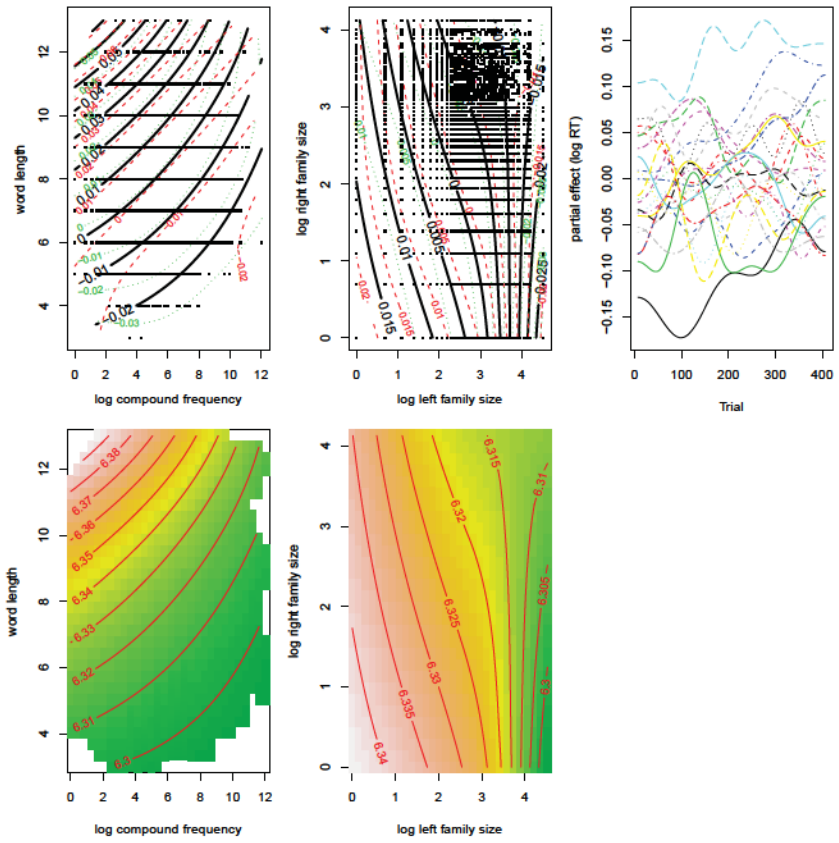

Figure 4.2: The GAMM for the naming latencies. Left panels: the tensor smooth for compound frequency by word length; central panels: the tensor smooth for left family size by right family size; right upper panel: the by-session random curves for trial; The upper left and central panels depict the partial effects with 1 standard error confidence regions around the contour lines. The corresponding lower panels show the corresponding fitted surface (predicted from all regressors in the model at their most typical values).

Table 4.1: Generalized additive mixed model fitted to the naming latencies

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	6.3301	0.0212	298.1473	< 0.0001
Phonological Neighborhood Size	0.0033	0.0014	2.3448	0.0191
B. smooth terms	edf	Ref.df	F-value	p-value
tensor Compound Frequency by Word Length	4.4913	5.2448	34.3124	< 0.0001
tensor Left Family Size by Right Family Size	5.3862	6.1132	15.0522	< 0.0001
by-session random smooths for Trial	159.2157	215.0000	12.6853	< 0.0001
random intercepts tone first syllabeme	4.7843	5.0000	21.9638	< 0.0001
random intercepts tone second syllabeme	3.4996	5.0000	1.8328	0.0449
random intercepts second syllable type	2.3402	7.0000	1.5703	0.0037
random intercepts first place	4.6478	5.0000	23.5155	0.0259
random intercepts first manner	2.9848	3.0000	718.2719	< 0.0001

Naming latencies

We included three predictors characterizing the acoustic properties of the first segment to account for variation caused by these properties in the voice key measures (see e.g., Treiman et al., 1995): `Phonation` (with levels *voiced* and *voiceless*), `Place` of articulation (levels: *alveolar, bilabial, glottal, labiodental, palatal, velar*), and `Manner` of articulation (levels: *approximant, fricative, nasal, plosive*). The latter two factors were entered as random-effect factors into the model specification, in order to obtain a model that not only accounts for potential differences across factor levels, but that is also parsimonious in the number of parameters. The same considerations motivated bringing the tones (six in all) for the first syllabeme, and the tones for the second syllabeme (again six levels) into the model specification as random-effect factors. Because the way tone is written indeed adds orthographic complexity to the written word (tone *ngang* does not have a tone mark in the orthography, tone *nặng* has a visually smaller diacritic than tone *ngã* or tone *sắc*) we include tone as random-effect factors. Details about tone can be seen in the Discussion section. Naming latencies were log-transformed in order to remove most of the rightward skew in their distribution, thereby avoiding potential adverse effects of overly influential outliers.

We fitted a generalized additive mixed model (GAMM) to the data, using the `mgcv` package (Wood, 2006, 2011) in the `R` programming environment (R Core Team, 2013). Generalized

additive mixed models offer the analyst better possibilities for capturing autocorrelational structures in the data that arise in response latencies (and, as we shall see, in acoustic durations) across the sequences of trials in an experiment. Standard linear mixed-effects models (e.g., Pinheiro and Bates, 2000; Bates et al., 2013) allow the analyst to model subject-specific trends in the time series of responses by means of subject-specific intercepts and slopes. However, the actual trends tend not to be linear, but curvilinear. GAMMs provide factor smooths for curvilinear trends that are shrunk towards zero, effectively creating 'wiggly' trends over experimental time (`Trial`) that are truly random effects. GAMMs offer the researcher the additional possibility of bringing into the model specification that the errors are autocorrelated following a simple ARIMA (AR) process:

$$e_{t+1} = \rho e_t + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma), \quad -1 \leq \rho \leq 1. \tag{4.2}$$

A further advantage of generalized additive mixed models is that they offer improved means for modeling interactions between numeric predictors. The multiplicative interaction of the standard linear model assumes the fitted surface is part of a hyperbolic plane, which for many actual data is unrealistic. We therefore opted for tensor products to model the interplay of lexical predictors.

Table 4.1 summarizes the GAMM fitted to the naming latencies. The upper part of this table presents the parametric part of the model, with an intercept and a positive slope for phonological neighborhood size. Naming latencies increased for words with more phonological neighbors. There was no significant effect of `FirstPhonation`, thus voicing did not affect latencies measured.

The second half of the table presents the smooth terms in the model, including the effects of random-effect factors. An interaction of compound frequency by word length (in letters) was modeled with 4.6 effective degrees of freedom (*edf*) using a tensor product smooth. (Greater values for *edf* indicate greater degrees of wiggliness.) The upper left panel of Figure 4.2 shows the partial effect of these two predictors, with 1 SE confidence intervals around the contour lines. The corresponding lower panel presents the fitted surface with all other regressors in

the model taken into account. As word length and compound frequency do not interact with other predictors, this surface is the same as the partial surface, except for an upward shift along the Z-axis that is mostly due to the intercept having been added to the predicted values. The tensor surface shows that word length has an inhibitory effect that is strongest for low-frequency compounds, and substantially reduced for high-frequency compounds. Conversely, (log-transformed) compound frequency is in general facilitatory, but its effect is very weak for short words, whereas its effect is most pronounced for long words.

The central panels of Figure 4.2 present the tensor smooth for the interaction of the family size of the left and right syllabemes. Across the full range of right syllabeme family sizes, we observe an effect of left syllabeme family size. However, although there is a facilitatory effect of right syllabeme family size for lower values of left syllabeme family size, this facilitation disappears for words with large left syllabeme family sizes.

The upper right panel of Figure 4.2 illustrates the intertrial dependencies at the different sessions during which data were collected, modeled with shrunk factor smooths. These factor smooths also take into account the modulations of the intercept across the different sessions — note that the session curves reach the vertical axis at different heights.

The last five rows of Table 4.1 clarify that the tones of the first and second syllabemes contribute to the fit of the model, and that the same holds for the controls for the voicekey, manner and place of articulation. An alternative to entering the tones of the two syllabemes as separate predictors is to include as random effect the pairs of tones realized on the compound's syllabemes. The resulting model is very similar both with respect to overall goodness of fit as with respect to the effects of the other predictors. Finally, the syllable type of the second (but not the first) syllable (CV, CVC, CwV, CwVC) reached significance. With a more fine-grained partition of syllable types that distinguishes between syllables with and without the first consonant (CV, V, CVC, VC, CwV, wV, CwVC, wVC), the random effect for the type of the first syllable also reaches significance. The token frequencies of the first and second syllabeme did not reach significance.

Inspection of the autocorrelation function of the errors revealed the errors were still autocorrelated even though the effect of trial was explicitly modeled. We therefore used an

AR1 model for the residuals, with $\rho = 0.3$, thereby avoiding anti-conservative p-values. We checked for the presence of overly influential outliers by removing data points with absolute scaled residuals smaller than 2.5 SE. As results remained highly similar, we reported the untrimmed full model.

Acoustic durations

Trials eliciting extreme word durations (less than 200 ms or exceeding 800 ms) were removed from the dataset (a loss of about 100 data points). The total number of words analysed is 8795.

Table 4.2 and Figure 4.3 summarize the GAMM fitted to the acoustic durations of the words read aloud. The upper part of Table 4.2 indicates that words beginning with a voiceless segment had a shorter duration by approximately 14 milliseconds. This is likely due to the difficulty in identifying the onset of the stop closure.

A second predictor with a linear effect is the frequency of the second syllabeme. Durations increased with second syllabeme frequency. The frequency of the first syllabeme did not reach significance, nor did the two frequency measures enter into an interaction.

Table 4.2: Generalized additive mixed model fitted to the acoustic durations

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	564.9062	28.3273	19.9421	< 0.0001
Left Syllabeme: Voiceless Initial Segment	-13.8162	1.8336	-7.5348	< 0.0001
Frequency Second Syllabeme	2.9582	1.0859	2.7243	0.0065
B. smooth terms	edf	Ref.df	F-value	p-value
tensor Comp. Freq by Word Length	7.2248	9.1284	25.3215	< 0.0001
smooth Phonological Neighborhood Size	3.6078	4.2956	9.5478	< 0.0001
tensor left Fam by right Fam Size	5.8543	7.3901	6.4407	< 0.0001
by-session random smooths for Trial	75.6504	215.0000	4.2266	< 0.0001
random intercepts tone left syllabeme	4.7723	5.0000	35.6649	< 0.0001
random intercepts tone right syllabeme	4.9941	5.0000	1778.4508	< 0.0001
random intercepts syllable type left syllb.	2.1840	3.0000	4.4002	0.0526
random intercepts syllable type right syllb.	2.9633	3.0000	871.7621	< 0.0001
random intercepts left place	4.8771	5.0000	410.7771	< 0.0001
random intercepts left manner	2.9414	3.0000	1373.8096	< 0.0001

The first tensor product in the subtable of smooth terms concerns the interaction of compound frequency and word length. The partial effect and the corresponding fitted surface are shown in the left panels of Figure 4.3. For low-frequency compounds, the effect of word length is strong, but for high-frequency words, its effect is minimal. The effect of word frequency is facilitatory across the board, but most clearly so for the longer words, and least for short words.

The family size measures for the left and right syllabeme entered into a non-linear interaction visualized in the second column of Figure 4.3. Durations are longest when both family sizes are small, and shortest when both are large. The lower right panel of Figure 4.3 illustrates that words with more phonological neighbors are realized with longer acoustic durations. The effect is virtually non-existent for the lower counts, but manifests itself clearly for higher neighborhood sizes. The upper right panel shows how acoustic durations change as the speaker proceeded through a session. As for the response latencies, we observe substantial local consistency, and global variation, as a session unfolds.

Inspection of the autocorrelation function of the errors revealed no further autocorrelation in the residuals. A model with potentially overly influential outliers (with absolute scaled residuals smaller than 2.5 SE) was inspected. As results were virtually identical, the original untrimmed model is reported here.

Discussion

For both onset latencies and acoustic durations, we observed an effect of compound frequency in interaction with word length. The interaction of the two predictors was somewhat stronger for the acoustic durations. Nevertheless, for both response variables, long low-frequency words emerged with the longest naming latencies as well as the longest acoustic durations. The facilitatory effect of compound frequency for the naming latencies fits well with other studies on compound processing (e.g., Baayen, 2010; De Jong et al., 2002; Juhasz et al., 2003; Kuperman et al., 2008, 2009), and has been interpreted as evidence for whole-word lexical representations.

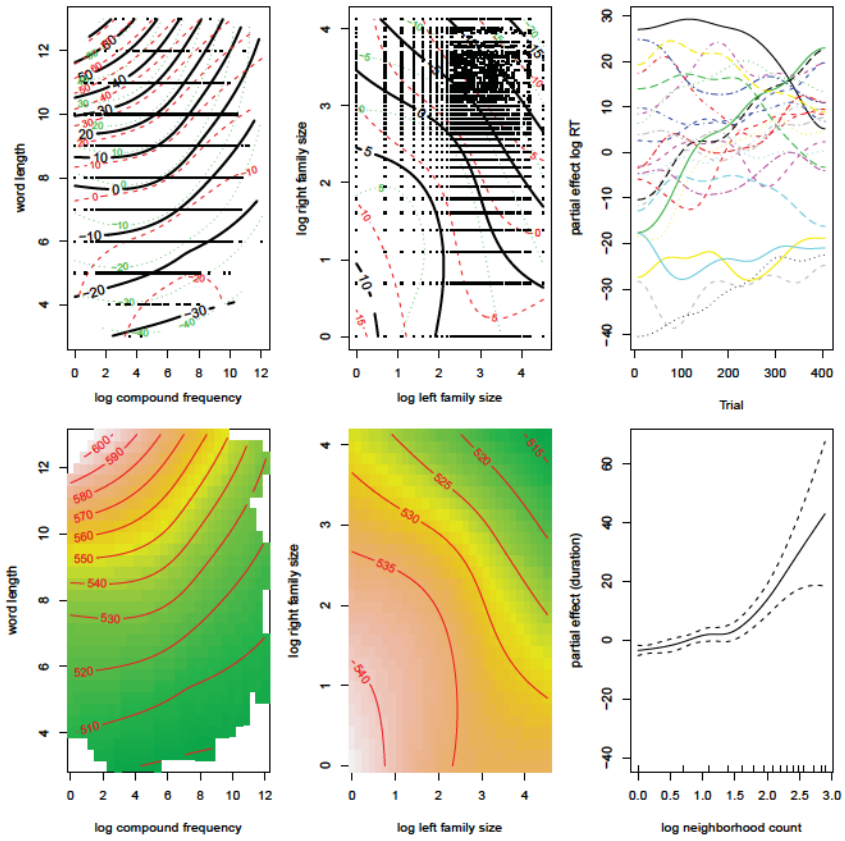

Figure 4.3: Non-linear effects in the GAMM fitted to the acoustic durations. Left panels: the interaction of frequency by word length; center panels: the interaction of left and right family size; upper right panel: the by-session random curves for trial, lower right panel: the effect of phonological neighborhood size.

Some recent studies (see e.g., Adelman et al., 2006; Brysbaert and New, 2009; Plummer et al., 2013) reported words' contextual diversity (known as dispersion in lexical statistics) to be an important factor for gauging how well words are entrenched in memory. Adelman et al. (2006) observed for English that when frequency is residualized on contextual diversity, frequency is no longer a significant predictor. We failed to replicate this effect for Vietnamese. Adding dispersion as a predictor does not improve the fit of the model to the data. Frequency can be replaced by dispersion, but this does not lead to a better fit either — in fact, the quality of the fit decreases slightly. It is conceivable that measures more sensitive to contextual co-occurrence (MacDonald and Shillcock, 2002; Baayen, 2010) may provide superior predictivity.

Surprisingly, we were not able to document any reliable frequency effects tied to the constituent syllabemes for the response latencies. Replacing syllabeme frequency by syllabeme dispersion led to inferior model fits, with no support whatsoever for a frequency effect for the second syllabeme, and a non-significant trend for facilitation from the first syllabeme's dispersion ($p = 0.0885$).

A syllabeme frequency effect is present for the acoustic durations, but only for the second syllabeme, such that as the frequency of the second syllabeme increased so did the duration. For English, a decrease in acoustic duration has been reported to go hand in hand with increasing lexical frequency (Aylett and Turk, 2004). This has led to the hypothesis that informationally redundant words (i.e., high-frequency words) would be shortened to maintain a uniform flow of information per time unit (the smooth signal hypothesis). Counterevidence for this hypothesis has been reported for Dutch interfixes by Kuperman et al. (2006) and for the German high front vowel by Tomaschek et al. (2013). The present study adds further evidence that the relationship between frequency and acoustic duration may not be straightforward.

However, it is possible that second syllabeme frequency is a sub-optimal measure. One might think a dispersion based variant would perform better. Unfortunately, the correlation of second syllabeme frequency and second syllabeme dispersion is extremely high ($r = 0.96$). When added to the model reported above, it helps increase the goodness of fit, but with a sign

opposite to that of second syllabeme frequency, indicating suppression (Friedman and Wall, 2005). Furthermore, second syllabeme frequency remains significant with a substantially smaller standard error compared to the dispersion measure. Thus, it is unlikely that changing from frequency to dispersion is at all helpful.

The frequency of the second syllabeme is also well correlated with second syllabeme family size ($r = 0.69$). Importantly, when the family size effects are witheld from the model specification, the second syllabeme frequency is no longer significant. When second syllabeme frequency is witheld, most of its effect is absorbed by the right syllabeme family size, especially for low first syllabeme family size values, with a minimal decrease in goodness of fit. We therefore refrain from interpreting the second syllabeme frequency effect.

For both naming latencies and acoustic durations, solid syllabeme effects are present in the form of family size effects, thereby extending the evidence for the relevance of this lexical distributional measure from Indo-European, Semitic, and Finno-Ugric (Moscoso del Prado Martín et al., 2004) to the Austro-Asiatic language family. In terms of network science (Jungnickel, 2007; Baayen, 2010), the family size of the first syllabeme is its out-degree, and the family size of the second syllabeme is its in-degree in the Vietnamese directed compound graph. The main pattern is that a greater degree (evidence for a more important status as a 'hub' in the network) affords both shorter naming latencies and shorter acoustic durations. For the present data, the centrality of a syllabeme as a hub in the lexical network appears to be a much more solid predictor than its bare frequency of occurrence.

However, it is conceivable that a cumulative frequency measure for syllabemes, counting not how often a syllabeme occurs as an independent word, but how often it is used as part of a compound, would capture at least part of the variance currently explained by the family size measures. Earlier work on Dutch (Schreuder and Baayen, 1997) suggested the cumulative frequency of a constituent across compounds to have little independent predictivity, but we need to establish that this holds as well for Vietnamese. We calculated the left and right cumulative syllabeme frequency for all of the compounds in our data set, aggregating for each syllabeme the frequency with which it occurs in a compound. When we added the log-transformed cumulative syllabeme frequency counts as predictors to the model, they

failed to reach significance, whereas the family size measures remained fully significant. This supports the conclusions reached by Schreuder and Baayen (1997) that for morphologically related words, it is type counts that are predictive, and not token counts.

What the present data indicate is that for word naming latencies, and for acoustic durations as well, the frequencies of a compound's constituent syllabemes are at best weakly predictive. The absence of solid support for syllabeme frequency effects in naming latencies contrasts markedly with the strong constituent frequency effects in word naming reported for the first constituent of compounds in English (Baayen et al., 2010). Thus, the role of the syllabeme as a constituent appears to be substantially reduced in Vietnamese compared to English. Further evidence for such a difference can be gleaned from a prior study on lexical processing in Vietnamese, using the visual lexical decision task. Pham and Baayen (2014) reported a principal components regression analysis indicating that a latent variable representing both syllabeme frequency and family size predicted longer response latencies. To disentangle potentially distinct effect of syllabeme family size and frequency, we conducted a supplementary analysis of their data, which suggested strong inhibition for syllabeme frequency and mild facilitation for syllabeme family size. This inhibitory constituent frequency effect in visual lexical decision also stands in marked contrast to the facilitation typically observed for English (Baayen et al., 2010) and Dutch (Kuperman et al., 2008, 2009). This result plus the whole compound frequency effect then supports a processing model where these compounds are being processed as wholes.

Why is constituent frequency inhibitory in Vietnamese lexical decision and non-predictive in Vietnamese word naming, whereas in English and Dutch, facilitatory effects of constituent frequency are well attested (see, e.g., Baayen et al., 2010)? We think that a marked cross-linguistic difference in the relative frequency of compound and constituent frequencies is at issue here. For the current set of 8875 Vietnamese compounds, 72% have a compound frequency exceeding the frequency of the left syllabeme. The corresponding percentage for the right syllabeme is 73%. By contrast, using the data set of 1252 English compounds studied by Baayen (2010), the corresponding percentages are an order of magnitude smaller at a mere 0.9% and 0.8%. What we observe here is a fundamental difference in the markedness

of compounding. In English, simple words are the central, unmarked, onomasiological units, whereas compounds present the marked case. In Vietnamese, two-syllabeme compounds constitute the unmarked word formation pattern, whereas single syllabeme words instantiate the marked, less frequently used pattern. Given that syllabemes are both syllables and lexemes (in the sense of Aronoff, 1994), we can equivalently describe the preferential phonological form of a Vietnamese word as bi-syllabic.

In the visual lexical decision task, single syllabeme words are low-frequency competitors of the compounds containing them, and as such slow lexical decisions. The reason why constituent frequency effects disappear in word naming is likely to be due to the nature of the word naming task, which is less sensitive to meaning compared to the lexical decision task (see, e.g., Baayen et al., 2006). This explanation implies that the family size effect for compounds' constituents in Vietnamese, which has also been observed for English word naming (Baayen et al., 2010), must reflect familiarity with the articulation of these constituents in speech production, rather than semantic co-activation of family members. This interpretation of the family size effect fits well with the analysis of the acoustic durations. Words with larger morphological families were pronounced shorter, whereas words with many phonological neighbors were realized with longer acoustic durations.

Table 4.3: Posterior modes for tone by syllabeme and task

	1st Syllb: Duration	1st Syllb: Latency	2nd Syllb: Duration	2nd Syllb: Latency
ngang	-1.2073	0.0105	47.6179	-0.0013
huyền	9.4485	-0.0027	61.8613	0.0008
hỏi	-5.4276	-0.0159	13.1098	-0.0040
ngã	12.5321	-0.0085	-23.9459	0.0041
sắc	-8.3506	0.0096	-7.3964	-0.0006
nặng	-6.9951	0.0070	-91.2466	0.0009

Table 4.3 summarizes the effects of the six tones of northern Vietnamese (see e.g., Đoàn, 1977; Nguyễn and Edmondson, 1998, for further details): (1) *ngang* mid level, (2) *huyền* low falling (breathy), (3) *hỏi* mid falling(-rising), harsh, and (4) *ngã* mid rising, glottalized, (5) *sắc* mid rising, tense, and (6) *nặng* mid falling, glottalized, short. The table presents a cross-classification by response variable (latency versus duration) and position (first versus

second syllabeme). There are substantial changes in duration of the same tone depending on whether it is realized on the first or on the second syllabeme. For the second syllabeme, we find more extensive lengthening (see, e.g., tone 2) as well as more extensive shortening (see, e.g., tone 6). This suggests a tight relation between vowel duration and tone (cf. Cao, 1962; Pham, 2001). Thus, tone *nặng* has a steep fall that is executed quickly and is often glottalized or followed by a glottal stop, resulting in a short vowel duration. The greater effect on duration for the second syllabeme is supported by a paired t-test on the absolute magnitudes of the effects ($p = 0.055$).

Interestingly, the effects of the tones on the latencies present a mirror image. Here, effects are large for the initial syllabeme, and an order of magnitude smaller for the second syllabeme (p =0.005, paired t-test). These results provide an important constraint on models of speech production. According to the WEAVER^{++} model of Levelt et al. (1999), the production of a word would proceed morpheme by morpheme, and within a word, segment by segment. Various experiments (e.g., Cholin et al., 2004, 2006) using the implicit priming paradigm carried out on Dutch suggested that knowledge of units later on in a sequence does not afford shorter production onsets. The present experiment on a non Indo-European language provides unambiguous evidence that the onset of articulation is co-determined by the tone to be realized on the second syllabeme. The effects are small, but sufficiently well-supported by the GAMM. This indicates that these compounds are processed as a whole.

This effect fits well with the facilitating effect of the family size of the second syllabeme on the naming latency. As can be seen in Figure 4.2, the magnitute of the family size of the left constituent is larger than that of the right family size (which is even entirely absent for words with large left syllabeme families). As the initial syllabeme has to be articulated first, a strong effect of its family size effect is unsurprising. When the initial syllabeme is used less often in compounds, we do find a small effect of the family size of the second syllabeme, consistent with the small effect of the tone of the second syllabeme.

The importance of properties of the second syllabeme for the onset of articulation of the compound receives further support from the effect of the syllable type of the second syllabeme. As can be seen in Table 4.4, second syllabemes ending in a consonant give rise to

elongated naming latencies. For word durations, the magnitude of the effects of the syllable type are much larger for the second syllabeme than for the first. Table 4.4 illustrates that acoustic durations are longer for vowel-final second syllabemes, and reduced for second syllabemes ending in a consonant. It is likely that part of this effect is due to the fact that in Vietnamese final stops are unreleased. As a consequence, there is little information in the acoustic signal to be used to determine the end of the closure gesture, thus the end of the word would be marked at the onset of the closure. The segmenter will take the beginning of the closure as the end of articulation, whereas in reality, the gesture for the closure may extend considerably in time before a (burstless) release takes place. Therefore, the durational differences observed for the second syllable are confounded with inevitable inaccuracies of automatic word segmentation of speech files. Note that for the initial syllabeme, syllable final consonants do not present a problem for the segmenter, as the initial consonant in the onset of the second syllable provides sufficient information to properly delimit the end of the preceding consonant.

Table 4.4: Posterior modes for syllable type by syllabeme and task

	2nd Syllabeme: Latency	1st Syllabeme: Duration	2nd Syllabeme: Duration
CV	-0.0000	2.7996	14.6455
CVC	0.0009	3.0161	-26.1235
CwV	0.0001	-4.4389	27.3200
CwVC	-0.0043	-1.3768	-15.8421

Both the analysis of the response latencies and the analysis of the acoustic durations provide support for an effect of the count of phonological neighbors of the compound. The effect of neighborhood density on the durations is substantial (on the order of 40 ms, see Figure 4.3), whereas the effect on the naming latencies is tiny (on the order of 5 ms). The inhibitory effect of neighborhood density in the response latencies is consistent with the inhibitory effect reported for this measure for picture naming by Sadat et al. (2014). The effect of neighborhood density on the acoustic durations, however, is different from the effect of neighborhood density reported by Gahl et al. (2012) for English. These authors show for English that words with many neighbors are easier to produce, and hence afford phonetic reduction. It is unclear whether similar considerations would pertain to Vietnamese,

which, unlike English, is not a stress-timed language but a syllable-timed language (see e.g., Roach, 1982; Cummins et al., 1999; Romano et al., 2011) in which reduction phenomena linked to morphologically determined syllable structures in stress-timed languages such as Dutch or English (Kemps et al., 2005a,b) are unlikely to occur. Our results indicate that in Vietnamese, a large neighborhood density gives rise to substantial durational enhancement instead of durational shortening.

In a lexical decision task, with a superset of the present words and the same participant (Pham and Baayen, 2014), the neighborhood count had a facilitatory effect ($\hat{\beta} = -0.0192, p < 0.0001$). The joint evidence of the two experiments indicates that facilitation arises in lexical decision due to neighbors providing evidence for lexicality, whereas in tasks in which a given target has to be selected for articulation, the same neighbors become competitors, delaying onset of articulation, and leading to prolonged acoustic durations. One reason why in speech production large neighborhoods give rise to a processing disadvantage could be that there are strong phonotactic restrictions on syllable structure. The four syllable types discussed above (CV, CVC, CwV, CwVC), in combination with a phoneme inventory with twenty-two consonant onsets, thirteen vowels, three diphthongs, one glide, and a restricted set of eight consonants for the coda position, allow for only a relatively limited number of syllabemes, and hence a dense lexical space. Preparing an articulatory pathway through this dense space, and maintaining that pathway against the alternative pathways of lexical neighbors, is likely to be time-costly, leading to both delayed onset of production, and to longer acoustic durations.

Given the analyses of the acoustic durations and the response latencies, the question remains to what extent durations and latencies are correlated. It turns out that there is a small but significant negative correlation between the two ($r = -0.05, t(8784) = -5.13, p < 0.0001$). The more time spent on preparation for articulation, the faster the execution of the articulatory programs can take place.

Concluding remarks

For the study of speech production, the picture naming task has the advantage of using a stimulus (a picture) that is non-linguistic in nature. By contrast, the word naming task combines a linguistic comprehension task (reading a word) with a speech production task (subsequently saying the word out loud). A disadvantage of the picture naming task is that many words are not, or not easily, depictable. The word naming task does not suffer from this disadvantage. For gauging speech production processes, however, the naming latency captures only the time required for preparing articulation. In addition to the naming latency response variable, this study considers an important factor as a second response variable, which has often been disregarded in naming studies, namely the acoustic duration of the speech produced.

With a coverage of nearly 9,000 words, we observed some remarkable similarities, as well as some subtle differences with findings in studies on English and Dutch. Words with larger phonological neighborhoods required longer response preparation times, and were articulated with longer acoustic durations. Lower-frequency longer words also demand longer preparation, and come with elongated durations. Furthermore, the greater the family sizes of the left and right syllabemes are, the shorter the time required for preparing for articulation, and the shorter a compound's acoustic duration can be. The convergence of the GAMMs for latency and duration indicates that the word naming task captures important aspects of the later stages of the speech production process with remarkable fidelity.

Yet, some subtle differences are worth pointing out. The effect of phonological neighbors appears to be driven more by higher numbers of neighbors when it comes to acoustic durations, as evidenced by the non-linear positive accellerating functional form of its partial effect. The effect of the second syllabeme's family size is much stronger for the acoustic durations than for the response latencies, indicating that the response latencies are dominated by processes preparing the articulation of the initial syllabeme. In addition, the effects of length and compound frequency (and their interaction) appear to be somewhat stronger for the acoustic durations. It is conceivable that the importance of these effects for the production process is

underestimated when querying response latencies, due to confounding with the preceeding comprehension process. Finally, for reasons that are as yet unclear to us, the effect of the tone realized on the first syllabeme is large for the response latency and small for acoustic duration, whereas for the tone realized on the second syllabeme, the reverse holds. Further research will have to clarify whether there are asymmetries in these effects with respect to the acoustic durations of the constituent syllabemes. Irrespective of how the effects of tone are to be understood, it is clear that even during the planning of the articulation of the first syllabeme, the properties of the second syllabeme are being taken into consideration, challenging the speech production model of Levelt et al. (1999), but consistent with the literature on co-articulatory effects in speech (Nittrouer and Studdert-Kennedy, 1987; Fowler and Saltzman, 1993).

References

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes.* The MIT Press, Cambridge, Mass.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47:31–56.

Baayen, R. H. (2010). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In Olsen, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H., Feldman, L., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

Baayen, R. H., Kuperman, V., and Bertram, R. (2010). Frequency effects in compound processing. In Scalise, S. and Vogel, I., editors, *Cross-Disciplinary Issues in Compounding*, pages 257–270. Benjamins, Amsterdam/Philadelphia.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely,

J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Bates, D., Bolker, B., Maechler, M., and Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*.

Bell-Berti, F. and Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *The Journal of the Acoustical Society of America*, 65(5):1268–1270.

Bien, H., Baayen, R. H., and Levelt, W. M. J. (2011). Frequency effects in the production of Dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes*, 27:683–715.

Bien, H., Levelt, W. M. J., and Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences of the USA*, 102:17876–17881.

Boersma, P. and Weenink, D. (2012). *Praat: Doing phonetics by computer [Computer program]*. Version 5.3.23.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Cao, X. H. (1962). Bàn về giải thuyết âm vị học một số vận mẫu có nguyên âm ngắn trong tiếng Việt [Discussion on the phonological hypothesis on some rimes containing short vowels in Vietnamese]. *Thông báo khoa học Đại học tổng hợp Hà Nội*, 2:146 – 154.

Cholin, J., Levelt, W. J., and Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, 99(2):205 – 235.

Cholin, J., Schiller, N. O., and Levelt, W. J. M. (2004). The preparation of syllables in speech production. *Journal of Memory and Language*, 20(50):47–61.

Cummins, F., Gers, F., and Schmidhuber, J. (1999). Comparing prosody across many languages. Technical report, Corso Elvezia 36, CH 6900 Lugano, Switzerland. Technical Report IDSIA-07-99, Istituto.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., and Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language*, 81:555–567.

Đoàn, T. T. (1977). *Ngữ âm tiếng Việt [Vietnamese phonetics].* Đại học và Trung học chuyên nghiệp, Hà Nội.

Forster, K. I. and Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers*, 35(1):116–124.

Fowler, C. A. and Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and speech*, 36 (Pt 2-3):171–195.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59:127–136.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction inspontaneous speech. *Journal of Memory and Language*, 66(4):789 – 806.

Hoàng, P., editor (2000). *Từ điển tiếng Việt [Vietnamese Dictionary].* Khoa học Xã hội, Hà Nội. Viện Ngôn ngữ học.

Janssen, N., Bi, Y., and Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language Cognitive Processes*, 23(7):1191–1223.

Juhasz, B. J., Starr, M. S., Inhoff, A. W., and Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from lexical decision, naming, and eye fixations. *British Journal of Psychology*, 94:223–244.

Jungnickel, D. (2007). *Graphs, networks and algorithms.* Springer, Berlin.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. H. (2005a). Prosodic cues for morphological complexity: The case of Dutch noun plurals. *Memory and Cognition*, 33:430–446.

Kemps, R., Wurm, L. H., Ernestus, M., Schreuder, R., and Baayen, R. H. (2005b). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, 20:43–73.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behaviour Research Methods*, 42(3):643–650.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23:1089–1132.

Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2006). Morphological predictability and acoustic salience of interfixes in Dutch compounds. *JASA*, 122:2018–2024.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of polymorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35:876–895.

Levelt, W. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38; discussion 38–75.

MacDonald, S. A. and Shillcock, R. C. (2002). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44:295–323.

Mok, P. K. (2010). Language-specific realizations of syllable structure and vowel-to-vowel coarticulation. *The Journal of the Acoustical Society of America*, 128(3):1346–1356.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish

compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

Nguyễn, D. H. (1997). *Vietnamese: Tiếng Việt không son phấn*. John Benjamins, Amsterdam.

Nguyễn, K. T. (1963). *Nghiên cứu về ngữ pháp tiếng Việt [Studies in Vietnamese grammar]*. Nhà xuất bản Khoa học, Hà Nội.

Nguyễn, T. A. T. and Ingram, J. C. L. (2007). Stress and tone sandhi in Vietnamese reduplications. *Mon-Khmer Studies*, 37:15–39.

Nguyễn, T. C. (1981). *Ngữ pháp tiếng Việt: Tiếng – Từ ghép – Đoản ngữ [A Vietnamese grammar: Syllable – Compound – Phrase]*. Nhà xuất bản Đại học và Trung học chuyên nghiệp, Hà Nội.

Nguyễn, T. G. (1985). *Từ vựng học tiếng Việt [Vietnamese lexicology]*. Đại học và Trung học chuyên nghiệp, Hà Nội.

Nguyễn, T. G. (2011). *Vấn đề "từ" trong tiếng Việt [The issues of "word" in Vietnamese]*. Nhà xuất bản Giáo Dục, Hà Nội.

Nguyễn, V. L. and Edmondson, J. A. (1998). Tones and voice quality in modern northern Vietnamese: Instrumental case studies. *Mon-Khmer Studies*, 28:1–18.

Nguyễn, V. T. (1976). *Từ và vốn từ trong tiếng Việt hiện đại [Words and word stocks in modern Vietnamese]*. Nhà xuất bản Đại học và Trung học chuyên nghiệp, Hà Nội.

Nittrouer, S. and Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of speech and hearing research*, 30(3):319–329.

Pham, H. and Baayen, H. R. (2014). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Submitted for publication*.

Pham, H., Tucker, B., and Baayen, H. (2014). Constructing two Vietnamese corpora and building a lexical database. *Manuscript*.

Pham, H. T. (2001). *Vietnamese tone: Tone is not pitch*. University of Toronto, (Doctoral dissertation).

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Plummer, P., Perea, M., and Rayner, K. (2013). The influence of contextual diversity on eye movements in reading. *Journal of experimental psychology. Learning, memory, and cognition*.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In Crystal, D. and Palmer, F. R., editors, *Linguistic controversies*, pages 73–79. Arnold.

Romano, A., Mairano, P., and Calabro, L. (2011). Measures of speech rhythm in East-Asian tonal languages. In *Proceedings of ICPhS XVII 2011*, pages 1714–1717, Hong Kong, China.

Sadat, J., Martin, C. D., Costa, A., and Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, 68(0):33 – 58.

Schreuder, R. and Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, 37(1):118–139.

Thompson, L. (1965). *A Vietnamese grammar*. The University of Washington Press, Seattle.

Tomaschek, F., Wieling, M., Arnold, D., and Baayen, H. (2013). Word frequency, vowel length and vowel quality in speech production: An ema study of the importance of experience. *Language and Speech*. To appear in *Interspeech 2013*.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., and Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, 124(2):107–136.

Trần, T. K., Phạm, D. K., and Bùi, K. (1941). *Việt-Nam văn-phạm [Vietnamese grammar]*. Lê Thăng.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28(4):735–747.

Vitevitch, M. S. and Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, 21(6):760–770.

Wood, S. N. (2006). *Generalized additive models*. Chapman & Hall/CRC, New York.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Yap, M. J., Liow, S. J. R., Jalil, S. B., and Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42(4):992–1003.

CHAPTER 5

Conclusions

This dissertation investigates two seemingly contradictory perspectives of the processing of Vietnamese compound words, namely corpus linguistic and psycholinguistic paradigms. They have previously been considered contradictory perspectives because: (1) Looking back to the Chomskyan era, the field of psycholinguistics was highly influenced by Chomsky, who helped establish a new relationship between linguistics and psychology in the 1950s. (2) Corpus linguistics (in the sense of McEnery and Wilson, 1996), on the one hand, is a system of methods and principles of which use corpora (bodies of language data) in language studies. On the other hand, psycholinguistics concerns the mental and neural processes as well as the behaviours associated with language. Chomsky has argued against the use of corpora in linguistic research, suggesting that corpora are not a useful source for linguistic data, since the linguist must seek to model language competence rather than performance and a corpus is reflective of performance. In response to Chomsky's criticism, linguists, interested in corpus data, have proposed the concepts of balanced (of text genres) and representative (of the language included in the corpus) corpora. Fillmore (1992, p. 35), who more or less implied these concepts in corpus linguistics when saying that: "I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding

out about in any other way".

Nevertheless, a connection between a set of fields, such as corpus linguistics and psycholinguistics, promise more fruitful exchanges (Gries, 2012) with many advances which have been obtained in each field. My primary goal in this dissertation has been to understand the systematic relations that exist between compound words and their constituents and how they influence the processing and storage of words in the mental lexicon. In addition to frequency, I investigated family size effects for compounds' constituents, secondary family size effects, the effects of tones, the effects of syllable types, and the phonetic characteristics of syllabemes, such as phonation types, place of articulation, and manner of articulation. This dissertation elucidates how Vietnamese compound words are represented and accessed in the mental lexicon by investigating the effects and interaction of these factors in the domain of visual word recognition. Chapter 2 addressed the construction of two Vietnamese corpora and the lexical database. Chapter 3 investigated the morphological aspects of constituents through the visual lexical decision experiment. Chapter 4 studied the lexical naming times of Vietnamese disyllabic words. In what follows, I summarize the findings observed in the main chapters of the dissertation and provide some additional discussion.

Single-subject large-scale experiment paradigm

The advances of large-scale studies and the use of large non-experimental data sources have become increasingly common in research in recent years. In the field of psycholinguistics, methods of data collection have been rapidly evolving along various dimensions. One of the trends called megastudies is to build large laboratory-based experiments without constrained[1] research questions, which differs from small-scale experiments in which using only small number of items. Megastudies collect behavioral measures for many items using tasks such as lexical decision and lexical naming, with some pioneer projects such as the English Lexicon Project (Balota et al., 2007) or the British Lexicon Project (Keuleers et al., 2012).

[1]This can be thought of as a solution to the question with regard to the researcher's intention which has been raised by Cutler (1981) and Forster (2000). Furthermore, the mega-study approach has been considered a real experiment (see e.g., Baayen, 2010a; Balota et al., 2012).

The number of datasets produced using the megastudy approach is increasing for different languages and using different experimental tasks, such as lexical decision for French: (Ferrand et al., 2010), speeded naming and lexical decision for English (Hutchison et al., 2013), lexical decision for Dutch: (Keuleers et al., 2010).

As discussed in Chapter 3, the single-subject large-scale research design has the advantage of statistical power for lexical properties, but is not necessarily representive of the full population of speakers. With regard to the question of the representativeness of a single participant, no general conclusion can be drawn for the population of language speakers. However, the single-subject large scale experiment covered a wide range of the language. We also found that the single-participant was more consistent and better in capturing variance due to lexical predictors than the subjects in the multi-subject study using a subset of items. We believe that for languages where access to speakers is limited, a comprehensive single-subject study may be an excellent solution for psycholinguistic investigation in these languages.

The corpus and lexical database

In chapter 2, I reported a two-fold study: (1) the construction of two Vietnamese corpora and (2) associated lexical predictors. The two corpora, the general corpus GENLEX-VIET and the subtitle corpus SUBTLEX-VIET are representative of written language and spoken language (based on translations of movie dialogues) respectively. This paper describes the stages of constructing the corpora from the text collection to the calculation of lexical variables. Three versions of each corpus were created, including the plain text, tokenized, and POS tagged versions. The lexical-statistical predictors aimed to serve as variables in the statistical analysis of the data described in Chapters 3 and 4. As introduced realier, there is an interdependence between psycholinguistics and corpus linguistics. There are many ways in which corpus linguistic data can play an important role in psycholinguistics, such as in the design and interpretation of psycholinguistic experiments. First, corpus data can be used as a natural/authentic source of language for selecting stimuli for psycholinguistic experiments.

Second, quantitative measures based on corpus data, such as frequency, dispersion, and collocation, are essential variables for psycholinguistic research. Corpora allow psycholinguists to research the types of statistical predictors. For that reason, I also expect the data to be useful for future investigation of Vietnamese in the areas of quantitative corpus linguistics, psycholinguistics, and computational linguistics.

To test the validity of the corpus data, we validated the frequency and dispersion measures with behavioral data using the visual lexical decision data described in Chapter 3. We found that the dispersion measures computed from the subtitle corpus better predict RTs from the two lexical decision experiments. Nevetherless, the frequency measures from the general corpus better predict RTs of the same experiments. Given the differences in complexity and representativeness of two corpora, I suspect that the word frequency in a more complex and representative corpus, i.e., word frequency of GENLEX-VIET corpus, matters the response times. Whereas, with a genre-consistent corpus, i.e., SUBTLEX-VIET corpus, the dispersion, which is the number of contexts that contain a word, matters the response times. This finding is not in line with previous studies and I provide two possible explanations: (1) the usage of the reference corpora and the film subtitle corpora in previous studies are too distant from each other in terms of the eras in which the two corpora were built; and/or (2) the sizes of the reference corpora in previous studies are too small in comparision to the French, English, and Dutch film subtitle corpora.

We conclude that constructing a film subtitle corpus as a supplementary language resource, especially for underdocumented languages, is a viable way to create corpus material for underdocumented languages. However, the creation of a large monitor and genre-rich corpus is encouraged and beneficial when resources permit.

The non-decompositional model for reading Vietnamese compound words

Linguistic units are not simply strung together like beads on a string in certain patterns. Thus, a sentence is created using a grammar which stipulates how these units can be put together. Language researchers, however, presume that words constitute the (mental) lexicon. One of the most prominent linguistic units used is the word. Even though the word would seem to fill a fundamental role in these approaches to linguistic description. There is, however, no common definition agreed between linguists with regard to the exact nature of the word. In psycholinguistics the word has also been assumed to be a fundamental and often the main unit of storage in the mental lexicon. This dissertation defines the lexicon as the inventory of constructions[2] which are used to produce and comprehend language. The communicative purpose can only be accomplished when the speaker/writer and the listener/reader are using the same, or at least similar, form-meaning pairs.

The lexicon is considered a repository of all simplex and complex constructions that are idiosyncratic and/or conventionalized. First, idiomatic features of a linguistic construct characterized by the unpredictable properties that can only be learned and acquired by the speaker. For instance, in Vietnamese, the compound word *(bệnh) càng cua* 'whitlow' is constructed by two monosyllabic words, *càng* 'pincers' and *cua* 'crab'; or *bổ túc* 'complementary education, continuation education' is made up by the combination of two monosyllabic words, *bổ* 'to add to the shortage' and *túc* 'full'. Since the meaning of these compounds do not directly relate to the meaning of their constituents, as well as the frequency effects observed in the present study, I conclude that they are constructions which have their own status in the mental lexicon along with their constituents used as monosyllabic words. Two-syllable compounds or phrases are the minimal size for lexical construction; the larger size may be idioms, proverbs, or sentences. Second, conventionalized features characterized by that a construction is stored in the mental lexicon if speakers experience the repetition in language usage to establish their functions and meaning, which is important for the establishment of

[2] A construction consists of form and meaning (Goldberg, 2005, 2006; Booij, 2010).

a construction.

The first question addressed in this dissertation is whether compound words are processed as wholes or as separate morphemes. Chapter 3 and 4 found strong experimental evidence demonstrating that known-compound words in Vietnamese are represented in the mental lexicon as wholes rather than morpheme-by-morpheme representation. Specifically, when whole-word and syllabeme frequencies varied, word frequency had an effect on the lexical decision response times, suggesting that Vietnamese compound words are represented as wholes in the mental lexicon and that lexical access occurs via those whole-word units. These results lend strong support to non-decompositional models of processing.

Chapter 3 investigated whether whole-compound frequency better predicts the response times in a visual lexical decision task. The whole-compound frequency, which was captured by `PC freq`, was found to be the most important predictor in the mixed-effects model for the single subject dataset. The second most important predictor was PC freq-fam, which differentiates words with large families and low frequency from high-frequency words with small families. We also found a marginally facilitatory effect for compound words in the strongly connected component (SCC) which was present together with strong evidence for modulation of the effect by `PC freq-fam`. The modulation of shortest path length by frequency is very similar to the interaction of shortest path length by first constituent family size for word naming in English (Baayen, 2010b). In the multi-subject experiment, we observed almost identical effects as those found in the single-subject experiment. An effect of `PC left-right` was captured, but only for words in the strongly connected component. Analyzing the subset of words with a second constituent in the strongly connected component, we also found an interaction of `Shortest Path Length` by `PC freq` when the second constituent is used as a classifier. The interaction was restricted to those compounds with a second constituent that is not used as a classifier as found in the single-subject experiment. The general inhibitory effect of family size in Vietnamese was replicated in this multi-subject experiment.

In chapter 4, we found using naming latencies that longer words were responded to more slowly. This effect was strongest for low frequency compounds and was substantially reduced for high-frequency compounds. We also found that log-transformed compound frequency is

in general facilitatory for long compounds, whereas its effect is weak for short compounds. The token frequency of the first and second syllabeme did not reach significance. I also analyzed the acoustic durations from the words in the naming experiment. I observed an interaction of compound frequency with word length. The interaction of the two predictors was relatively stronger for the acoustic durations. This might be due to the importance of these effects for the production process is underestimated when querying response latencies, due to a confound with the preceeding comprehension process. The effect of word frequency is facilitatory across the board, and is strongest for the longer words. For both naming latencies and acoustic durations, long low-frequency words emerged with the longest responses. The facilitatory effect of compound frequency for naming latencies has been interpreted as evidence for whole-word lexical representations.

Chapter 3 approached the questions from both a single-subject experiment and multi-subject experiment paradigms. We found that compound processsing appears to be optimal when the left and the right families are in balance. The length of the shortest path from head to modifier is potentially relevant, i.e, when the shortest path length is included as a predictor, `PC left-right` loses significance. From the graph/network theory perspective, the graph-theoretical effects observed for Vietnamese converge with similar effects observed in English, the sign of the effect of `PC freq-fam` is specific to Vietnamese.

In chapter 4, we considered the frequency effects linked to the constituent syllabemes for the response latencies but did not find any effects. Models fitted with syllabeme dispersion did not improve the fit either. For the acoustic durations, a syllabeme frequency effect is present but only for the second syllabeme, such that as the frequency of the second syllabeme increased so did the duration. This finding is not in line with the hypothesis that informationally redundant words, e.g., high-frequency words, would be shortened to maintain a smooth flow of information per time unit (Aylett and Turk, 2004). For both naming latencies and acoustic durations, reliable syllabeme effects are present in the form of family size effects. The pattern found is that a greater degree of the centrality of a syllabeme as a hub in the lexical network affords both shorter naming latencies and shorter acoustic durations better than its bare frequency of occurence.

Tone effects

In a tone language like Vietnamese where tone is one of the obligatory components of the orthographic syllabeme, tones play an important role in visual language processing. Furthermore, some tones are more productive than others in the language. This raises the question of how this is possible, given some tones are more productive than others, that more productive tones are recognized faster and that certain combinations of tones on compounds are easier to read than others.

In chapter 3, we found the two tone random effect factors contribute significantly to the goodness of fit of the mixed-effects model. Investigation of the posterior modes for the tone of the first syllabeme shows that tones *huyền* and *sắc* elicit longer latencies than the other four tones. With respect to the second syllabeme, the tone *hỏi* elicited the shortest latencies, and the *huyền* and *ngang* tones the longest. As expected for a language rich in tones, the two tone random effect factors also contribute substantially to the goodness of fit. In the multi-subject experiment, the effect of tone on the second syllable were lost, likely due to a lack of power.

In chapter 4, the effect of the tones of the first and second syllabemes also contribute to the fit of the model. An alternative to entering the tones of the two syllabemes as separate predictors is to include them as random effects as in chapter 3. The resulting model is very similar both with respect to overall goodness of fit as with respect to the effects of the other predictors. I also found that there are substantial changes in duration of the same tone depending on whether it is realized on the first or on the second syllabeme. For the second syllabeme, we found more extensive lengthening (e.g., tone 2) as well as more extensive shortening (e.g., tone 6). Unsurprisingly, this suggests a tight relation between vowel duration and tone (see e.g., Yu, 2010, for further discussion). The effects of the tones on the latencies show a different effect. Effects are large for the initial syllabeme, and an order of magnitude smaller for the second syllabeme. We present evidence that the onset of articulation is co-determined by the tone to be realized on the second syllabeme. This indicates that when the first constituent being processed, the second constituent is also

processed. This can be interpreted as that Vietnamese compounds are processed as a whole as discussed in the previous section.

Lexical storage and lexical processing

This dissertation presents evidence that Vietnamese readers are sensitive to compounds as a whole in visual word recognition. This raises the question of whether such whole-word form might be the stored lexical representation, together with the representation of its constituents. In Chapter 3, I approached this question by investigating whether the effect of whole-compound interacts with the effects of its constituents in the visual lexical decision task. I carried out a lexical decision task, and found that there was indeed a frequency effect of the whole-word form in processing which was stronger for compounds than for their constituents. This is consistent with the view that the semantic information is present in the lexical representations of compounds. In other words, the hypothesis of lexical storage of whole-compound representation is consistent with the pattern of frequency effects observed in the lexical decision experiments in Chapter 3.

The hypothesis of lexical storage of full-form compounds was also supported by the lexical naming data in Chapter 4. In Chapter 4, I considered the lexical processing times of word naming task by gauging speech production processes with naming latency response variable and acoustic duration of the speech produced. Analyzing these complimentary variables together with many lexical predictors, such as `compound frequency`, `family size`, `tone`, `syllable type`, etc., sufficient evidences have been observed to support a processing model where these compounds in Vietnamese are being processed as wholes. This is also consistent with the view that compound words are presented in the memory as proposed by the non-decomposition models.

Implications to the concept of word in Vietnamese

One implication of this study is to revisit the concept of word from psychological perspective. The concept of *word* seems clear in many Indo-European languages, such as English. In Vietnamese, however, it is by no means a clear and intuitive notion. In the Vietnamese literature, the salient and intuitive concept is *chữ* 'grapheme'. This term, as presented in the previous chapters, actually has three distinct meanings in general usage: it can mean a morpheme in word formation, it can mean a syllable in the spoken language, or it can mean an orthographic grapheme. The powerful influence of *chữ* has led to the belief that in Vietnamese each syllable (or grapheme) is a word (Nguyễn, 2011, p. 125).

Psycholinguistic research on compound word processing has documented that not only simplex words are stored in the brain but also the complex words leave traces in lexical memory, as shown by Taft (1979) and Sereno and Jongman (1997) for comprehension in English, Baayen et al. (1997, 2002) for comprehension in Dutch, and Bien et al. (2005) for production in Dutch. These studies found that the frequency of complex words were predictive for processing times, autonomously from the frequencies of their constituents. These robust frequency effects are regarded as evidence for the independent representation of complex words in memory. Wurm et al. (2006) observed that the presence of word frequency is in harmony with the absence of root frequency effects. The presence of word frequency and root frequency effects have been interpreted as the evidence of whole-word based processing versus decompositional processing respectively. As shown in this study, compound words in Vietnamese are real and fundamental constructs, which differs from the point of view that they are artifacts of Western linguistic analysis (see e.g., Nguyễn, 1984; Nguyễn, 2011, p. 119). Therefore, the notion 'word', including compounds, may universally exist in language comprehension and language production as a real linguistic construct.

Following Wurm and colleagues (2006), I also propose that whole-word frequency can be interpreted as a joint probability of the co-occurrence of constituents of compounds. In other words, lexical frequency effects evolve from memory for constituents and memory for sequences of these constituents (compound words). The response latencies for compound

processing indicate simultaneously both kinds of memory (see also Baayen, 2007).

Challenges and topics for further research

As some of the first psycholinguistic work on Vietnamese, this dissertation opens the door to many topics that require further investigation. In this section I discuss several topics that seem most relevant. First, the network of words in the mind might be represented in an interconnected form such that orthorgraphical, phonological, and associative semantic surfaces interact with each other and between compounds' constituents. A comprehensive model of word processing might not be complete if lacking one of these surfaces. The presence of strong frequency effects for compound words over their constituent frequency effects in both lexical decision and lexical naming should be investigated in a larger and more diverse population of speakers. Although the frequency effect on compounds as wholes is more robust than on their constituents, studies on the associative semantics of compound words as well as the semantic relations between compounds' constituents are needed in order to reveal the more insightful representation of words in the mental lexicon.

Second, as an isolating language, the Vietnamese syllabeme is thought of as the prominent unit in morphology, having been partially revealed through the experiments in this study. My proposal for the model of word processing in Vietnamese is that complex words known by the reader/speaker are stored as wholes in the mental lexicon. In other words, constructions are the basic unit in Vietnamese lexical processing. In this view, as supported with the data presented in Chapters 3 and 4, Vietnamese whole-compound recognition can be achieved directly and independently from the morphological parsing route. The non-deconpositional model offers an explanation for the results obtained in Chapters 3 and 4. On this account, for word naming latencies, and for acoustic durations as well, the frequencies of a compound's constituent syllabemes are at best weakly predictive. Thus, the role of the syllabeme as a constituent appears to be substantially reduced in Vietnamese. It is conceivable that, as a consequence, readers of Vietnamese access compound words in their mental lexicon in the nondecomposition mode. Further studies of interest investigating how Vietnamese language

users process new or unknown compounds. How do they use compound constituents to understand these compounds? What are the features of constituents needed for comprehend new compounds?

A third question concerns the semantic transparency of compound words, which is one of the key issue in the area of compound processing (Libben, 1998; Libben et al., 2003; Sandra, 1990; Zwitserlood, 1994). Semantic transparency refers to the consistency between the meaning of a compound word and it constituents, whether it is transparent (e.g., *tea table*), semi-transparent(e.g., *strawberry*), or opaque (e.g., *moonshine*). Libben (1998) argued that both transparent and opaque compounds are processed through a morphological-decomposition procedure at the lexical form level. The absence of a semantic-priming effect for opaque words was due to the lack of connections between opaque compounds and their constituents at the semantic level. For example, the opaque compound *hogwash* activates the lexical representations of *hogwash*, *hog*, and *wash*. The lexical representation of *hogwash* is connected to its semantic representation as a whole word, but there are presumably no connections between the lexical representation of *hogwash* and the semantic representations of *hog* and *wash*. Even though the activation of *hog* and *wash* at the lexical level would activate their semantic representations as well, their connections with *hogwash* would be indirect. This argument has been supported in a number of studies (e.g., Libben et al., 2003; Zwitserlood, 1994). I reason that (Vietnamese) opaque compound words may fall under the Lexical Integrity Hypothesis, in which syntax does not have access to word-internal information (Jackendoff, 1972). Given that Vietnamese has both right headed and left headed compounds, future research will need to consider the interaction between semantic transparency and headedness in modelling the processing cost for compounds. Having established that there are in fact compound words built from two stems in which the semantic relation between two constituents to the whole word may be either transparent or opaque. An implementation for English using CARIN theory (Gagné and Shoben, 1997) has been modelled by Pham and Baayen (2013). Vietnamese, with a limited number of syllabemes, has a large number of compound words. The consideration of interaction between semantic relations and synonymy and polysemy factors might be an interesting area to learn more about

processing knowledge of language in general.

Fourth, it remains an open question whether the reading models proposed for Indo-European languages are applicable to Vietnamese. The fact that some effects found in Vietnamese are not in line with other previous findings, raises the need for modelling the reading mechanism in Vietnamese. By implementing a computational model for the behavioral data we will know how accurate the model fitted to our data. In addition, the modeling results may reveal the psychological and neurobiological reality of language processing. Given that the behavioral data on lexical decision and lexical naming, modelling the isolating word processing for Vietnamese need further implementation. For simulations in the prospective study, we will train the lexicon networks built with the Rescorla-Wagner equations (see Wagner and Rescorla, 1972, for further information) on the input lexicon describe in Chapter 2 using the Näive Discriminative Learning (NDL) model described by Baayen et al. (2011).

Fifth, as previously mentioned, words are inter-connectedly represented in our mind as proven in free word association studies, such as the University of South Florida Free Association Norms (Nelson et al., 2004), the Edinburgh Associative Thesaurus (EAT)[3] and the Small World of Words project (De Deyne and Storms, 2008). This is mapping with the mental space semantic measures observed in corpus data, e.g., the LSA or HAL measures presented in Chapter 2. To better predict lexical response times and semantic relatedness in language behavioural data, such as lexical decision and lexical naming experiments, I suggest including family size of word associations into models of language comprehension and productions. In order to understand the relations among words and inspired by the work of De Deyne et al. (2013), I have initiated an experiment to measure how people's intuitive associations between words in Vietnamese by massively collecting data from the public at the website http://www.smallworldofwords.com/new/visualize/. I expect that rich data derived from word associations will contribute fine-grain size information to various phases of modelling lexical and semantic processing.

Finally, the current study focused on Vietnamese compound compreshension, e.g., compound reading and compound naming. However, actual language processing, such as auditory lan-

[3] http://www.eat.rl.ac.uk/

guage processing, requires further research to address the question. I propose that auditory lexical decision experiments need to be implemented to investigate the gap in this domain. Specifically, besides the robust effects of traditional variables for lexical experiments in general such as frequency and dispersion of both compound words and their constituents, we will evaluate the relative importance of onset characteristics, token (words and constituents) acoustic duration, neighborhood density of orthographic form and phonological form, uniqueness point, synonym family size, and age of acquisition on response times and accuracy.

To conclude, this dissertation provides evidence for non-decompositional models of lexical access and lexical naming in Vietnamese compound words. It also supports the psychological status of compound words in Vietnamese. Compounds, under the view of construction morphology, are constructions which comprise the blending of form, meaning and other inputs, resulting in word-like entities in the mind that somewhat line up with the stream of consciousness. Further research is necessary to elucidate the semantic transparency, semantic relations and headedness between compound constituents, auditory recognition as well as the word associations between compound words in Vietnamese.

References

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47:31–56.

Baayen, R. H. (2007). Storage and computation in the mental lexicon. In Jarema, G. and Libben, G., editors, *The Mental Lexicon: Core Perspectives*, pages 81–104. Elsevier.

Baayen, R. H. (2010a). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1):149–157.

Baayen, R. H. (2010b). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In Olsen, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 36:94–117.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Baayen, R. H., Schreuder, R., De Jong, N. H., and Krott, A. (2002). Dutch inflection: the rules that prove the exception. In Nooteboom, S., Weerman, F., and Wijnen, F., editors, *Storage and Computation in the Language Faculty*, pages 61–92. Kluwer Academic Publishers, Dordrecht.

Balota, D., Yap, M., and Hutchison, K. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In Adelman, J. S., editor, *Visual word recognition: Models and methods, orthography and phonology*, volume 1, pages 90–116. Psychology Press, London and New York.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Bien, H., Levelt, W. M. J., and Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences of the USA*, 102:17876–17881.

Booij, G. E. (2010). *Construction morphology*. Oxford University Press, Oxford; New York.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic at all in 1990? *Cognition*, 10:65–70.

De Deyne, S. and Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40(1):198–205.

De Deyne, S. D., Navarro, D., and Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45(2):480–498.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In Svartvik, J., editor, *Directions in corpus linguistics: Proceedings of Nobel symposium 82, Stokholm, 4-8 August 1991*, Trends in Linguistics. Studies and Monographs [TiLSM] Series. De Gruyter.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory and Cognition*, 28:1109–1115.

Gagné, C. L. and Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–87.

Goldberg, A. (2005). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Goldberg, A. (2006). *Constructions at work. The nature of generalization in language*. Oxford University Press, Oxford.

Gries, S. T. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In Mukherjee, J. and Huber, M., editors, *Corpus linguistics and variation in English: Theory and description*, pages 41–63. Rodopi, Amsterdam.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., and Buchanan, E. (2013). The semantic priming project. *Behavior research methods*, 45(4):1099–1114. 00002.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. MIT Press, Cambridge, MA.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 dutch mono- and disyllabic words and nonwords. *Frontiers in psychology*, 1:174. 00046.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, 44(1):287–304.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, 61(1):30–44.

Libben, G., Gibson, M., Yoon, Y. B., and Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84:50–64.

McEnery, T. and Wilson, A. (1996). *Corpus linguistics*. Edinburgh University Press, Edinburgh.

Nelson, D., McEvoy, C., and Schreiber, T. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Nguyễn, T. G. (1984). Về mối quan hệ giữa "từ" và "tiếng" trong Việt ngữ [On the relation between "word" and "syllable" in Vietnamese]. *Ngôn ngữ [Language]*, 3:41 – 57.

Nguyễn, T. G. (2011). *Vấn đề "từ" trong tiếng Việt [The issues of "word" in Vietnamese]*. Nhà xuất bản Giáo Dục, Hà Nội.

Pham, H. and Baayen, H. R. (2013). Semantic relations and compound transparency: A regression study in CARIN theory. *Psihologija*, 46(4):455–478.

Sandra, D. (1990). On the representation and processing of compound words: automatic access to constituent morphemes does not occur. *Quarterly Journal of Experimental Psychology*, 42A:529–567.

Sereno, J. and Jongman, A. (1997). Processing of English inflectional morphology. *Memory and Cognition*, 25:425–437.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition*, 7:263–272.

Wagner, A. R. and Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.

Wurm, L., Ernestus, M., Schreuder, R., and Baayen, R. H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and conditional root uniqueness points. *The Mental Lexicon*, 1(1):125–146.

Yu, A. C. (2010). Tonal effects on perceived vowel duration. In Fougeron, C., Kuehnert, B., Imperio, M., and Vallee, N., editors, *Laboratory Phonology 10*, pages 151–168. De Gruyter, Berlin, Boston.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language & Cognitive Processes*, 9:341–368.