Behavior Research Methods
2008, 40 (1), 154-163
doi: 10.3758/BRM.40.1.154

Corpora of Vietnamese Texts: Lexical effects of intended audience and publication place

Giang Pham, Kathryn Kohnert, and Edward Carney
University of Minnesota, Minneapolis, Minnesota

This article has two primary aims. The first is to introduce a new Vietnamese text-based corpus. The Corpora of Vietnamese Texts (CVT; Tang, 2006a) consists of approximately 1 million words drawn from newspapers and children's literature, and is available online at www.vnspeechtherapy.com/vi/CVT. The second aim is to investigate potential differences in lexical frequency and distributional characteristics in the CVT on the basis of place of publication (Vietnam or Western countries) and intended audience: adult-directed texts (newspapers) or child-directed texts (children's literature). We found clear differences between adult- and child-directed texts, particularly in the distributional frequencies of pronouns or kinship terms, which were more frequent in children's literature. Within child- and adult-directed texts, lexical characteristics did not differ on the basis of place of publication. Implications of these findings for future research are discussed.

Vietnamese is an Asian tonal language with approximately 80 million speakers globally (D. H. Nguyen, 2001). Although speakers of this language are primarily located in Vietnam (70–73 million speakers), there are also large numbers of Vietnamese speakers in Western countries, including Australia, Germany, France, and the Netherlands. There are an estimated 1.12 million Vietnamese in the United States, making this group the fourth largest Asian American population, following Chinese, Filipinos, and Asian Indians (Reeves & Bennett, 2004). Although useful information is available describing sounds, tones, lexical categories, and grammatical aspects of Vietnamese (e.g., D. H. Nguyen, 1997), only very limited information is available regarding frequency or distributional characteristics of these linguistic units. Large corpora have been collected on English (e.g., Kučera & Francis, 1967), as well as many other languages (for a review, see Wilson, Archer, & Rayson, 2006). When they are large enough in number and have an adequate variety of samples (according to one's purpose), language corpora may reveal much information about the linguistic patterns that are exemplars of "real life" language use (McEnery & Wilson, 2001).

In this article, we introduce the Corpora of Vietnamese Texts (CVT; Tang, 2006a) and compare it with the single existing corpus in Vietnamese (D. D. Nguyen, 1980). We then use the new data source to examine potential influences of publication place as well as intended audience on lexical measures. Because the CVT is composed of data published both inside and outside of Vietnam, and from adult- and child-directed texts, this type of analysis is seen as an important first step to qualify its practical utility. Preliminary to coding words into lexical classes, it is important to determine whether overall frequency counts are distributed equivalently across different source data included in the text database. This is true in any language, but takes on additional importance when dealing both with text that can be considered to be in a majority language (originally written in Vietnamese and published in Vietnam) as well as text in which the language of interest has minority language status (written or translated into Vietnamese and published in a Western country). In these situations, some of the available text may be translated from English into Vietnamese, as is often the case with children's literature. In other cases, geographic- and usage-based differences in Vietnamese across countries may result in quantitative as well as qualitative differences in language. We used the CVT to investigate potential differences and similarities in lexical frequency and distributional characteristics on the basis of place of publication (Vietnam or Western countries) and text genre (newspapers or children's books). We begin with an overview of the Vietnamese language, focusing on those aspects most relevant to corpora data collection and lexical analysis.

Characteristics of Vietnamese

Vietnamese is an isolating language, in that it does not use bound morphemes to express grammatical features such as number (singular/plural) and tense. Instead of bound morphemes, Vietnamese grammar relies on word order and function words (K. L. Nguyen, 2004). For comprehensive descriptions of Vietnamese across language domains, see D. H. Nguyen (1997) and Tang (2006b).

Modern Vietnamese script uses the Vietnamese alphabet *quốc ngữ*, or "national script," based on a Romanized script expanded with diacritics to mark certain vowels

G. Pham, tangx098@umn.edu

and tones. Vietnamese orthography is transparent, with a nearly one-to-one grapheme-to-phoneme correspondence. For the analysis of text corpora, particularly at the phonological level, this consistent sound–symbol correspondence represents a significant advantage over other languages that have a more opaque correspondence between sounds and written symbols. For instance, sound frequency counts may be conducted on the basis of written texts rather than transcriptions of spoken language.

Vietnamese was once erroneously considered to be a monosyllabic language, with each word equal to one syllable (e.g., Thompson, 1965). It is now recognized that Vietnamese words may consist of one, two, three, or even four syllables (D. H. Nguyen, 1997). Although a Vietnamese word may contain more than one syllable, single syllables continue to be separated in the writing system. That is, the spacing between each syllable creates the illusion that each syllable is one word. For instance, the single word "clock" is made up of two syllables separated by a space: *đồng hồ*. With regard to meaning, it is often difficult to define what constitutes a word in Vietnamese. For instance, although *mẹ con* may be translated into two English words ("mother" "child"), most Vietnamese linguists consider it one compound word (e.g., Do, 1981), because it signifies a single concept of mother–child relations. The ongoing debate about the definition of a "word," combined with the orthographically separated syllables in Vietnamese, poses a significant challenge for the creation of language corpora. Currently available corpora software programs are able to calculate frequency counts based on lexical form, but are not able to parse forms into word units based on meaning.

At the lexical–semantic level, words in Vietnamese as well as English can be divided into content and function words. Content words carry semantic meaning, whereas function words relate content words to each other (Stubbs, 2001). Content words for both English and Vietnamese may be further divided into word classes, such as nouns, verbs, and adjectives. In Vietnamese as well as English, lexical forms may have more than one meaning or belong to more than one word class, with meaning and grammatical class disambiguated by sentence context. In English, words may keep the same form (e.g., tree *bark* vs. dogs *bark*) or change in form (e.g., sit in the *chair* vs. he *chaired* the meeting) when changing word class (see Bauer, 1983). Vietnamese words change in word class without altering form (Tang, 2006b), which poses a challenge for corpora analyses. Word forms that may serve as nouns as well as verbs, for instance, can only be distinguished within the context of each sentence. No software programs are available to parse lexical items into separate word classes in Vietnamese. Needless to say, manual calculations of this type would be quite onerous for corpora containing millions of words.

Both Vietnamese and English have pronouns to substitute for nouns or noun phrases. An important language characteristic of Vietnamese that is not found in English is the use of kinship terms. Most Vietnamese kinship terms may be used as pronouns to reflect age, status, and gender of both speaker and listener (Tang, 2006b). Kinship terms that serve as pronouns are used with persons within and outside of one's family (Luong, 1990). There are only a few pronouns that are not kinship terms that can be used in a general sense, such as *tôi* ("I"). Within the family pronominal, kinship terms distinguish between paternal and maternal sides of the family, age, gender, and blood relations as opposed to in-law status (K. L. Nguyen, 2004). Unfamiliar speakers and listeners also refer to each other and themselves differently depending on social factors, including age and status. For example, a person who is approximately the age of one's uncle or aunt could be addressed as *chú* or *cô*, respectively, while referring to oneself as *cháu* ("niece/nephew") in the northern dialect or *con* ("son/daughter") in the southern dialect. When meeting someone approximately the age of one's older sister, one may refer to himself or herself as *em* ("younger sibling") and address the speaker as *chị* ("older sister"). When the relative ages of the speaker and listener are not known, it is common to address the listener with pronouns that indicate older age, as a sign of respect, because older age is associated with higher status (Luong, 1990).

Unlike English pronouns, Vietnamese pronouns do not indicate number. In order to indicate plurality in Vietnamese, a quantifier is added before the pronoun. For example, *các* ("some") is added before *chú* ("uncle") to indicate more than one male who is approximately the age of one's uncle: *các chú*. Vietnamese pronouns do not indicate person (speaker, listener, or third party), which poses another challenge for analyzing corpora data. Although frequency counts can be conducted at the form level, the meaning of the person reference can only be interpreted within the sentence or paragraph context. In English, there are different pronouns that indicate sentential subject and predicate positions (e.g., "she" vs. "her"). Vietnamese pronouns do not change form and therefore do not indicate subject and predicate position.

Vietnamese uses affixation, compounding, and reduplication to create new meanings from existing lexical forms. Affixation is the process by which a language attaches meaningful linguistic units (bound morphemes) to a word to change its meaning. Examples of affixation in English are *un-* in *unreal* or *-ful* in *wonderful*. Vietnamese uses prefixes and suffixes as well, although they are used differently. Rather than attaching to the word itself, affixes appear separate from the word. For instance, the prefix *bán* ("half, semi") appears before *cầu* ("sphere") to create the word *bán cầu* ("hemisphere") The suffix *hóa* ("-ize, -fy") appears after *Việt Nam* ("Vietnam") to create the word *Việt (Nam) hóa* ("to Vietnamize"; D. H. Nguyen, 1997). Since affixes are not attached to the word in Vietnamese, this may affect word-frequency counts in Vietnamese corpora data.

Compounding, the process of combining two or more words to create a new word, occurs in both Vietnamese and English. English examples include "armchair" and "beehive." Vietnamese examples include *hải quân* [(ocean armed-force) "(the) navy"] and *bàn ghế* [(table chair) "furniture"]. Traditionally, Vietnamese compound words appear as two or more separate syllables in the writing system, which, as mentioned earlier, poses a challenge for word-frequency counts based on large corpora.

In addition to compounding by combining two different words, compounding can also be achieved by repeating or reduplicating lexical forms. Compounding by reduplication rarely occurs in English and is primarily used in words that reflect sounds, or noises, such as "click clack" (Thompson, 1965). Vietnamese frequently uses reduplication in content words, such as verbs, adjectives, and nouns. Reduplications may consist of the replication of an entire syllable or of its individual components such as the rime, initial consonant sound, or principal vowel, and serve various semantic functions (G. T. Nguyen, 2003). Reduplication of a verb typically indicates movement. For instance, *gật* [*đầu*] ["to nod (one's head)"] can be reduplicated to indicate a continuous nodding motion: *gật gật đầu*. In the case of adjectives, reduplication can imply a lesser degree of a quality. For example, color terms such as "green" (*xanh*), can indicate a lighter shade when the word is reduplicated, *xanh xanh*. Certain nouns can be reduplicated to indicate reoccurrence or multiple instances, such as *ngày ngày* ("day day"), which implies many days or all days (C. T. Nguyen, 1999; D. H. Nguyen, 1997; G. T. Nguyen, 2003; K. L. Nguyen, 2004). Reduplications may affect the accuracy of lexical counts since they are typically thought of as one word but would be counted twice. (For additional information on characteristics of Vietnamese, see Tang, 2006b.)

CVT Collection and Characteristics

The CVT is composed of two different text genres, one typically directed toward adults (newspaper articles) and the other typically directed toward children (children's books). Because a general purpose of the CVT is to investigate language use in Vietnamese Americans as well as Vietnamese nationals, texts published in Vietnam as well as in Western countries were collected. Sources and word counts for these different text genres (adult directed or child directed) and publication places (Vietnam or other) are summarized in Table 1. A complete list of all sources is available online at vnspeechtherapy.com/vi/CVT/3_CVT_The%20Basics.htm.

The first text genre is made up of online Vietnamese newspaper articles from a total of four sources: two sources published in Vietnam and two sources published in the United States. Articles were collected from April to July of 2006. Article topics included world and national news, politics, health and medicine, education, current events, sports, editorials, economics, science and technology, relaxation, love, and daily life. Advertisements and comics were excluded from the corpus. Adult-directed texts were in electronic format and were collected from online newspaper sources; full articles were selected and pasted into a word processing program. As shown in Table 1, the total word count for newspaper articles is 851,174, making up 80% of the CVT. Of this total, 265,282 words (31%) come from articles published in Vietnam and 585,892 words (69%) were from articles published in the U.S.

The second genre consists of over 350 children's books, varying in reading level from preschool through fifth grade, including what are typically referred to as picture books, repetitive books, and folklore stories. Chapter books and comics were excluded from the corpus. Children's books were collected from elementary schools, libraries, and bookstores in the United States and Vietnam. Access to children's books was more limited, because they were not available in electronic format. The vast majority of books were published in Vietnam, because of the relatively limited availability of children's books in Vietnamese from other countries. Picture books that were published outside of Vietnam were primarily from the United States and England, with a few books published in Australia and New Zealand. Child-directed texts made up 20% of the CVT (see Table 1). In the child-directed texts, there were four times as many words from books published in Vietnam (163,543 words, or 79%) as there were words from books published in Western countries (43,845 words, or 21%), because of the limited amount of children's literature in Vietnamese available in English-speaking countries. In order to obtain relatively similar numbers of words across place of publication, we used almost twice as many words from adult-directed texts published in Western countries as we did words from adult-directed texts from Vietnam. Child-directed texts were manually typed into a word processor, since access to text-scanning software for Vietnamese was not available at that time (but see VnDOCR, 2006).

Table 1
CVT Composition and Word Counts

Publication Place	Newspaper Articles	Children's Literature	Total Words
Vietnam published	267,905	163,543	431,448
Other published	588,619	43,845	632,464
Total words	856,524	207,388	1,063,912

Note—Newspaper articles were collected from several sections of two newspapers published in Vietnam (*Thanh Niên, Tuổi Trẻ*) and two newspapers published in the United States (*VOA, VNN*) in the year 2006. Children's literature consisted of 279 picture books published in Vietnam and 78 picture books published in Western countries.

From a word processing program, all of the texts were then formatted for MonoConc Professional 2.2 (Barlow,

Table 2
Comparison of Vietnamese Corpora

Characteristic	D. D. Nguyen (1980)	Tang (2006a)
Type	Text	Text
Size	524,500 words	1,063,912 words
Format	Paper	Electronic
Description	Consists of newspaper articles, poetry, theatrical works, children's literature, and Ho Chi Minh's writings from 1956 to 1972	Consists of newspaper articles from 2006 and children's picture books from 1976 to 2006
Coding level	Separates lexical frequency by categories including nouns, verbs, adjectives, numbers, connecting words, proper nouns, and so on	Vietnamese-specific vowels and tones coded to be read by MonoConc Professional 2.2 concordance program
Overlap of top 100	—	67
Rank correlation[a]	—	.660*

[a]Based on common words of the 100 most frequent words of each corpus ($n = 67$). *$p < .0005$ in a one-tailed analysis.

2003), a concordance software program. Although MonoConc Professional 2.2 had the capability to read a variety of languages, the software was not able to read Vietnamese. Therefore, certain tones and vowels specific to Vietnamese were numerically coded using the find and replace function of the word processing program (for a complete list of codes, see Tang, 2006a). It should be noted that the word count electronically calculated by the word processor was 1,055,617, whereas MonoConc Professional 2.2 calculated a total of 1,063,912. This minor discrepancy (0.78%) may be due to the fact that neither the word processor nor the concordance program was programmed to count words in Vietnamese. Since we used the concordance program throughout the analyses, we used the word total of 1,063,912, calculated by the same program, for consistency.

There were notable differences in sample size across the four corpora. Sample sizes for newspapers were larger than were sample sizes for children's literature because newspapers were available electronically; access to children's books was limited to those available in libraries, bookstores, and elementary schools. Also, a text-scanning program for Vietnamese was not available at the time. The time needed to manually type children's books into a word processor was another practical limitation for the children's literature sample. Children's books that were available were primarily published in Vietnam; the sample size of children's books published in other countries was much smaller, by comparison. Tang (2006a) collected a larger sample size of newspapers published in other countries in order to counterbalance unequal sample sizes in children's literature. The following is a comparison of the CVT with an older Vietnamese corpus.

Existing Corpora Data in Vietnamese

Existing corpora data in Vietnamese are sparse. Prior to the CVT (Tang, 2006a), there was one published text-based corpus, by D. D. Nguyen (1980). There are no available corpora on spoken Vietnamese. The primary purpose of the D. D. Nguyen corpus was to identify fundamental Vietnamese vocabulary to contribute to the field of lexicology. Words were manually parsed, and frequency counts were divided into word classes on the basis of sentence meaning. The result of corpus analysis was a summary of basic Vietnamese vocabulary, with French translations. Table 2 summarizes general characteristics of the D. D. Nguyen corpus as compared with the CVT. Differences between the two corpora include size, format, and composition. The D. D. Nguyen corpus consists of 524,500 words from a variety of text genres, including novels, poetry, theatrical works, children's literature, newspaper articles, and Ho Chi Minh's writings. The D. D. Nguyen corpus was made up of texts published between 1956 and 1972. Over 66% (350,400/524,500 words) of the corpus by D. D. Nguyen consists of literary works, such as novels, poetry, theatrical works, and children's literature. Children's literature made up close to 14% (48,500/350,400) of the literary texts and 9% (48,500/524,500) of the entire corpus. Apart from the sample of children's literature, all text genres were for an adult audience. The CVT (Tang, 2006a) consists of 1,063,912 words from children's literature and newspaper articles. The children's literature was published between 1976 and 2006, and the newspaper articles were all published in 2006. The D. D. Nguyen corpus is available in paper format, whereas the CVT is in electronic format (vnspeechtherapy.com/vi/CVT/ResearchChude.htm).

Table 3
Overlap From 100 Most Frequent Words of the CVT

Comparison	Shared Words
Adult VN–Adult Other	78
Child VN–Child Other	80
Adult VN–Child VN	57
Adult Other–Child Other	53
Adult VN–Child Other	56
Adult Other–Child VN	56

Note—Displays the number of words shared across subcorpora.

Although the two corpora differ in many ways, they are comparable in general word frequencies. Appendix A lists words shared between the CVT (Tang, 2006a) and D. D. Nguyen (1980), based on the 100 most frequent words of each corpus ($n = 67$). A Spearman rank correlation was calculated as one measure of corpus similarity. There was a significant positive correlation between the two corpora ($r = .66, p < .001$), indicating that not only were the vast majority of words shared across corpora, but the frequency rankings were also similar.

In Appendix A, words are listed in descending order of log likelihood (LL) ratios with corresponding raw frequency counts and frequency rankings from each corpus. Rayson and Garside (2000) proposed using LL ratios for frequency profiling when comparing corpora, to estimate the relative frequency difference between two corpora. High LL ratios indicate great disparities in frequency rankings, whereas low LL ratios indicate high similarity in frequency ranking order across corpora. Rayson and Garside calculated LL ratios with the following equation: $2 * \{[a * \ln(a/E1)] + [b * \ln(b/E2)]\}$, where a = the frequency count of a word from Corpus 1, b = the frequency count of the same word from Corpus 2, and E is the expected value that is calculated using the following equation: $E_i = (N_i \Sigma O_i)/(\Sigma N_i)$. The combination of frequency ranking and LL ratios further informs our understanding of similarities and differences between the two corpora. For example, the kinship term *anh* ("older brother") occurs frequently in both Tang (2006a) and D. D. Nguyen (1980) but differs substantially in frequency ranks (64 and 9, respectively), yielding the highest LL ratio of 13,533.08. At the other extreme, the verb *có* ("to have") greatly differs in raw frequency across corpora but is ranked third in each corpus, with a corresponding LL ratio < 0.01. Another example is the word *và* ("and"), with the highest frequency in both corpora but also a relatively high LL ratio, indicating that its

Table 4
Spearman Rank Correlations Across Corpora

Corpus	Adult VN	Adult Other	Child VN	Child Other
Adult VN	–	.85	.40	.52
Adult O		–	.46	.52
Child VN			–	.79
Child O				–

Note—Based on the 100 most frequent words that occurred across all genres and places of publication ($n = 46$). All correlations are statistically significant at $p < .005$, on the basis of one-tailed analysis.

Table 5
Estimated Distributions of Word Classes Across Corpora

	Adult Vietnam		Adult Other		Child Vietnam		Child Other	
Word Class	Raw	%	Raw	%	Raw	%	Raw	%
Nouns	21,879	38.86	8,173	43.80	30,957	47.43	7,768	38.96
Verbs	20,766	36.89	6,408	34.34	24,285	37.21	7,228	36.26
Adjectives	14,397	25.57	4,421	23.69	14,402	22.06	4,831	24.24
Numerators	2,442	4.34	866	4.64	1,129	1.72	935	4.69
Pronouns[a]	1,697	3.01	210	1.13	18,291	28.02	3,412	17.11
Adverbs	9,948	17.67	2,369	12.70	10,743	16.46	3,315	16.63
Conjunctions	8,022	14.25	2,656	14.23	6,219	9.53	2,943	14.76
Prepositions	6,172	10.96	2,314	12.40	2,696	4.13	1,584	7.95
Total[b]	85,323	151.55	27,417	146.93	108,722	166.56	32,016	160.60
Subcorpus total[c]	56,295		18,659		65,272		19,936	

Note—Word class categorization was based on Tan (1994) and the *Vietnamese Dictionary and Translation* (2006). [a]Most pronouns are also Vietnamese kinship terms. [b]Many items may belong to more than one word class and were counted for each possible word class. [c]Based on the 100 most frequent words in each subcorpus.

use or relative "importance" may vary across corpora. The CVT by Tang (2006a) contributes to Vietnamese language corpora with the addition of current texts (1976–2006), electronic accessibility, and larger samples of daily language use (e.g., newspapers vs. literature). The composition of the CVT is further described in the following section and is the focus of all subsequent analysis.

Analyses of the 100 Most Frequent Words of the CVT

The CVT was divided into four separate corpora for comparison: newspapers published in Vietnam (Adult VN), newspapers published outside Vietnam (Adult Other), children's books published in Vietnam (Child VN), and children's books published outside Vietnam (Child Other). Given that the CVT was not parsed or tagged, we performed preliminary analyses on the 100 most frequent words of each subcorpus to investigate the potential composition of the entire corpus (see Appendix B for complete lists). Table 3 displays the number of words shared across intended audience and place of publication on the basis of the 100 most frequent words of each subcorpus. Texts directed toward adults (Adult VN, Adult Other) shared a relatively high number of words (78 of 100), and texts typically directed toward children shared a similar number of words (80 of 100). Fewer words were shared across texts directed to different audiences (adult vs. child), ranging from 53 to 57 of 100 words.

One-tailed Spearman rank correlations were calculated to examine how frequent words were ranked across subcorpora (see Table 4). All correlations were statistically significant ($p < .005$), indicating a relationship between the ranking of frequent words of each subcorpus on the basis of sampling of the 100 most common words. This finding seemed reasonable, given that the CVT is made up of one language (Vietnamese). It was important to note that texts directed toward adults were highly correlated ($r = .850$), texts directed toward children were highly correlated ($r = .791$), whereas texts intended for different audiences (adult, child) exhibited relatively lower correlations of around .50. Raw frequency counts of shared words (Table 3) as well as Spearman rank correlations (Table 4) highlighted overall differences between adult- and child-directed texts at the lexical level. However, these measures did not indicate differences on the basis of place of publication.

To further investigate lexical characteristics across subcorpora, we estimated distributions of word classes on the basis of the 100 most frequent words (see Table 5). The 100 most frequent words were listed separately for each subcorpus. Words were then classified into general categories of nouns, verbs, adjectives, numerators, pronouns, adverbs, conjunctions, and prepositions. As mentioned earlier, parsing tools were not available for Vietnamese, and manual calculations based on line-by-line sentential context were not feasible in this large sample. Therefore, in this analysis, words that could belong to more than one word class were counted in each possible category; total percentages were greater than 100%. Table 5 displays estimated distributions across word class in raw frequency counts and percentages.

As shown in Table 5, the most common word classes across all subcorpora were nouns, accounting for approximately 40% of words, followed by verbs (about 35%), and adjectives (about 25%). Similarities in proportion of the three main word classes indicated a consistent level of major word classes across subcorpora.

Table 6
Number of 100 Most Frequent Words That Belong to One or More Word Classes

Number of Word Classes	Adult VN	Adult Other	Child VN	Child Other
1	60	63	51	52
2	32	29	36	35
3	8	8	13	13

This agreement can also be seen in the number of words that belong to one or more word classes (see Table 6). Across subcorpora, the number of words that belonged to a single word class ranged from 51–63 of 100; words that potentially belonged to two word classes ranged from 29–36 of 100; and words that potentially belonged to three word classes ranged from 8–13 of 100. These estimations suggested that for certain types of corpus analyses, it may be possible to collapse across subcorpora to investigate major word classes such as nouns, verbs, and adjectives.

At the same time, differences between adult-directed and child-directed texts suggest that the CVT should be divided for certain analyses that could be unduly influenced by intended audience. For instance, the proportion of pronouns/kinship terms and prepositions differed between adult-directed and child-directed texts (see Table 5). The occurrence of pronouns/kinship terms ranged from 17%–28% in child-directed texts (children's literature), whereas they occurred in only 1%–3% of adult-directed texts (newspapers). A possible explanation is that kinship terms are often used in children's books with human or animal characters, such as *chú mèo* [(uncle cat) "Mr. Cat"]. In addition, there may be more dialogue in children's books, in which kinship terms are used to refer to the speakers and listeners. As shown in Table 5, prepositions occurred more often in adult-directed texts (11%) than in child-directed texts (6%). A possible explanation is that newspapers describe events in which explicit details of location and transactions are needed.

Summary and Future Research

The CVT database represents a significant addition to Vietnamese corpora in part due to its large sample size (over 1 million words), current content (years 1976–2006), inclusion of large samples of daily language use (i.e., newspapers), and electronic accessibility (www.vnspeechtherapy.com/vi/CVT). It is a tool that will allow systematic investigation of frequency and distributional characteristics of the Vietnamese language at phoneme, word, and sentence levels. Results of the lexical analyses described here suggested that the CVT may be collapsed for linguistic analyses on general word classes including nouns, verbs, and adjectives. On the other hand, for certain types of linguistic analyses, such as investigating the role of kinship terms, researchers should consider the impact of genre type. The present analysis revealed no significant differences for language produced or published in the majority versus minority language countries. This null finding supports collapsing the CVT across place of publication. However, it is also possible that place of publication will have a greater influence at other language levels. One limitation of these analyses is that frequency counts were based on syllable forms. As mentioned earlier, the concept of "word" is an ongoing debate in Vietnamese linguistics. Furthermore, no parsing software is available to identify Vietnamese word units. Future parsing tools may enable deeper lexical analyses that include more accurate lexical counts as well as investigation of compound words and the phenomenon of reduplication.

Frequency and distributional information at sound, word, tone, and grammatical levels is needed for a variety of pedagogical, theoretical, and experimental reasons (Thomas & Short, 1996). For example, to develop stimuli that will allow researchers to profile or test selected aspects of language in individuals who learn Vietnamese as a first or primary language, information regarding frequency and distributional characteristics of linguistic features is needed to develop stimuli for empirical validation and elaboration. The collection and analysis of corpora data are essential to understanding language and language use.

AUTHOR NOTE

Funding for this project was provided by the Graduate Research Partnership Program at the University of Minnesota and was awarded to the first author under the faculty mentorship of the second author. We thank Hai Anh Nguyen, Xuan Tran Tang, and Tien Pham, who helped manually type children's books for the children's literature subcorpus. We thank Pui Fong Kan, Mahmoud Sadrai, and Brian Gordon for technical assistance with computer software for corpus analysis. We thank the Center for Cognitive Processes in Language for the use of equipment. Correspondence concerning this article should be addressed to G. Pham (formerly G. Tang), Department of Speech–Language–Hearing Sciences, 115 Shevlin Hall, 164 Pillsbury Drive SE, University of Minnesota, Minneapolis, MN 55455 (e-mail: tangx098@umn.edu) or to K. Kohnert (e-mail: kohne005@umn.edu).

REFERENCES

Barlow, M. (2003). MonoConc Professional 2.2: A professional concordance program [Computer software]. Houston, TX: Athelstan.

Bauer, L. (1983). *English word-formation*. Cambridge: Cambridge University Press.

Do, C. H. (1981). *Từ vựng ngữ nghĩa tiếng Việt* [Vietnamese lexicosemantics]. Hà Nội: Nhà Xuất Bản Giáo Dục.

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

Luong, H. V. (1990). *Discursive practices and linguistic meanings: The Vietnamese system of person reference*. Philadelphia: Benjamins.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Nguyen, C. T. (1999). *Ngữ pháp tiếng Việt, in lần thứ sáu* [Vietnamese grammar, 6th ed.]. Hà Nội: Nhà Xuất Bản Đại Học Quốc Gia.

Nguyen, D. D. (1980). *Dictionnaire de fréquence du Vietnamien* [Frequency dictionary of Vietnamese]. Paris: Université de Paris.

Nguyen, D. H. (1997). *Vietnamese*. Amsterdam: Benjamins.

Nguyen, D. H. (2001). Vietnamese. In J. Garry & C. Rubino (Eds.), *Facts about the world's languages: An encyclopedia of the world's major languages, past and present* (pp. 794-796). New York: Wilson.

Nguyen, G. T. (2003). *Từ vựng học tiếng Việt, tài bản lần thứ tư* [Vietnamese semantics, 4th ed.]. Ho Chi Minh City: Nhà Xuất Bản Giáo Dục.

Nguyen, K. L. (2004). *Giáo trình tiếng Việt II* [Teachings on Vietnamese II]. Huế: Đại Học Huế Trung Tâm Tạo Tứ Xa.

Rayson, P., & Garside, R. (2000, October). *Comparing corpora using frequency profiling*. Paper presented at the Workshop on Comparing Corpora and the 38th Annual Meeting of the Association of Computational Linguistics, Hong Kong.

Reeves, T. J., & Bennett, C. E. (2004). We the people: Asians in the United States. Census 2000 Special Report (U.S. Census Bureau Report No. ASI 2004 2326-31.16). Washington, DC: U.S. Department of Commerce, Economics, and Statistics Administration.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Tan, V. (1994). *Từ điển tiếng Việt* [Vietnamese dictionary]. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

TANG, G. (2006a). Corpora of Vietnamese Texts. Retrieved October 7, 2006, from www.vnspeechtherapy.com/vi/CVT.

TANG, G. (2006b). Cross-linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States. *Journal of Southeast Asian-American Education & Advancement*, 2, 1-33.

THOMAS, J., & SHORT, M. (EDS.) (1996). *Using corpora for language research: Studies in honour of Geoffrey Leech.* London: Longman.

THOMPSON, L. (1965). *A Vietnamese grammar.* Seattle: University of Washington Press.

VIETNAMESE DICTIONARY AND TRANSLATION (2006). Retrieved January 15, 2007, from vdict.com/.

VNDOCR (2006). Version 2.2 [Vietnamese text-scanning software]. Retrieved October 1, 2006, from www.vndocr.itgo.com/.

WILSON, A., ARCHER, D., & RAYSON, P. (2006). *Corpus linguistics around the world.* New York: Rodopi.

APPENDIX A
Shared Words Across Tang (2006a) and D. D. Nguyen (1980)

	Tang (2006a)		D. D. Nguyen (1980)		
Item	Frequency	Rank	Frequency	Rank	LL
anh	2,659	64	3,854	9	13,533.08
ta	690	72	2,720	21	3,146.034
đi	3,495	37	3,408	13	785.4055
những	6,583	13	5,236	5	649.3777
khi	4,127	29	1,053	76	410.6416
lên	2,979	56	2,501	25	375.153
để	5,303	15	1,543	45	363.4923
ông	4,315	23	1,235	62	311.7079
phải	3,872	34	2,906	18	285.4529
nhà	4,155	28	1,241	61	260.989
thì	3,152	46	2,418	28	260.381
mới	1,887	99	1,617	43	259.2179
ở	4,204	27	3,052	16	258.002
mà	2,974	57	2,278	31	243.7917
còn	3,048	51	2,257	32	208.9722
làm	4,044	32	2,825	19	197.0004
tôi	4,278	25	2,916	17	177.6291
rồi	1,862	100	1,465	47	174.3677
mình	2,400	71	1,768	39	159.9267
thế	3,341	42	1,074	74	159.1296
sẽ	3,385	40	1,108	70	149.3626
năm	3,499	36	1,206	63	121.3333
và	13,710	1	7,903	1	120.7843
các	8,803	8	5,257	4	118.901
bị	3,451	38	1,203	64	112.9223
trên	3,113	48	2,074	35	110.2647
nước	3,368	41	2,203	33	104.0976
đã	8,047	10	4,771	7	100.2636
ra	5,091	16	3,091	15	81.9014
thấy	2,141	86	1,437	48	79.93346
này	4,545	20	1,758	41	77.18171
đang	2,237	83	1,470	46	71.60603
cũng	4,295	24	2,601	24	67.30049
vào	4,729	19	2,775	20	52.21448
cho	8,807	7	3,814	11	45.4523
từ	3,395	39	1,354	53	44.77509
sau	2,720	60	1,060	75	43.5632
nào	1,994	91	1,243	60	41.31653
theo	2,507	70	974	81	41.13648
nói	3,201	45	1,890	37	38.0248
cả	2,680	62	1,577	44	30.58672
với	5,723	14	2,481	26	29.08203
về	4,274	26	2,419	27	29.05309
trong	8,576	9	3,821	10	27.41079
nhưng	3,068	49	1,762	40	25.71427
đầu	2,724	59	1,129	68	24.58826

APPENDIX A (Continued)

	Tang (2006a)		D. D. Nguyen (1980)		
Item	Frequency	Rank	Frequency	Rank	LL
đến	4,921	17	2,141	34	23.67577
chỉ	2,999	55	1,262	58	22.72962
như	4,334	22	2,410	29	22.18372
là	10,904	5	4,963	6	21.96329
đây	2,020	90	1,178	67	20.66833
ngày	3,059	50	1,307	56	19.09984
trước	1,910	98	1,109	69	18.4953
lại	4,085	31	1,801	38	15.82077
của	12,998	2	6,058	2	13.11715
rất	1,981	92	1,091	72	8.534645
hơn	2,185	84	965	82	8.209146
qua	2,099	87	941	88	5.934528
nhiều	3,021	53	1,377	51	5.872452
người	6,963	12	3,598	12	5.220412
được	6,990	11	3,309	14	3.714559
đó	3,924	33	1,961	36	0.241675
không	8,995	6	4,396	8	0.224307
biết	2,245	81	1,094	71	0.099147
con	4,857	18	2,408	30	0.051142
việc	2,621	65	1,286	57	0.019531
có	11,052	3	5,441	3	0.007005

Note—Based on the 100 most frequent words of each corpus (*n* = 67). LL, log likelihood ratios, an estimate of the relative frequency difference between two corpora. High LL ratios indicate large differences in frequency rankings. Low LL ratios indicate high similarity in frequency ranking across corpora.

APPENDIX B
List of 100 Most Frequent Words in Each Subcorpus of the CVT

	Adult VN	Freq	Adult O	Freq	Child VN	Freq	Child O	Freq
1	anh	691	an	406	ăn	593	ăn	136
2	bị	687	anh	424	ấy	385	anh	213
3	biết	610	bị	850	anh	642	ba	130
4	bộ	627	bộ	420	bà	631	bà	172
5	cả	559	cả	543	bạn	537	bạn	160
6	các	2,231	các	1,806	bác	389	bác	149
7	chỉ	930	chỉ	571	bay	311	bé	208
8	chính	656	chính	885	bé	661	bên	86
9	cho	2,016	cho	1,579	bị	446	biết	118
10	có	3,094	chức	403	biết	395	cá	97
11	cơ	493	chủ	796	cả	577	cả	134
12	con	643	chúng	453	các	728	các	182
13	còn	754	có	1,712	cái	549	cái	231
14	công	1,362	cộng	421	cây	515	cho	322
15	của	3,229	con	424	chạy	335	chơi	126
16	cũng	1,075	còn	553	chàng	315	chú	229
17	dân	541	công	1,115	chỉ	354	chúng	295
18	đã	1,899	của	2,498	chim	355	có	502
19	đầu	808	cũng	668	cho	1,289	cô	273
20	đang	534	dân	1,491	chú	391	con	853
21	đây	519	đã	1,620	chúng	518	còn	107
22	đến	1,168	đầu	418	có	1,573	của	470
23	để	1,180	đảng	500	cô	655	cũng	135
24	đi	594	đại	424	con	2,232	đá	97
25	định	575	đến	722	còn	493	đã	259
26	điều	498	để	897	công	308	đầu	92

APPENDIX B (Continued)

	Adult VN	Freq	Adult O	Freq	Child VN	Freq	Child O	Freq
27	đó	952	điều	424	của	1,049	đang	120
28	đồng	647	đó	609	cũng	555	đâu	97
29	động	663	đồng	649	cùng	390	đây	117
30	được	2,041	động	450	đã	908	đến	232
31	do	523	được	1,200	đầu	392	để	145
32	gia	652	do	749	đang	454	đi	385
33	hàng	528	gia	537	đâu	308	đó	286
34	hiện	766	giới	432	đến	1,051	được	278
35	học	1,131	hà	470	để	696	em	152
36	hội	549	hai	400	đi	1,380	gấu	235
37	hơn	686	họ	419	đó	520	gì	147
38	khi	1,117	hội	872	được	1,087	giờ	107
39	không	2,104	khi	650	em	401	hai	99
40	là	2,590	không	1,482	gì	445	hỏi	136
41	lại	839	là	2,107	hai	425	họ	105
42	làm	1,109	lại	638	hoa	353	khi	209
43	lên	482	làm	629	hôm	377	không	505
44	lý	536	lên	414	khi	644	là	475
45	mà	634	mà	562	không	1,654	lại	208
46	mình	538	một	1,451	là	1,204	làm	203
47	mới	639	năm	847	lấy	341	lên	241
48	một	2,419	nam	1,029	lại	1,034	lớn	144
49	năm	1,081	này	757	làm	795	mà	96
50	nam	662	ngày	684	lên	1,012	màu	102
51	nay	520	người	1,364	lúc	422	mẹ	249
52	này	1,146	Nguyễn	428	mà	514	mình	191
53	ngày	737	nhà	854	mẹ	839	một	706
54	người	1,937	nhân	874	mình	729	nào	108
55	nhà	916	nhiều	449	một	2,213	này	128
56	nhất	522	như	765	nàng	345	ngày	137
56	nhân	633	những	1,083	nào	443	nghe	88
58	nhiều	1,000	nhưng	429	này	450	người	289
59	như	996	nội	552	ngày	481	nhà	235
60	những	1,655	nói	437	nghe	381	nhảy	90
61	nhưng	805	nước	898	người	1,015	nhiều	100
62	nước	888	ở	803	nhà	1,004	nhìn	99
63	ở	1,328	ông	710	nhìn	351	nhỏ	90
64	ông	813	phải	587	như	536	như	155
65	phải	964	pháp	449	những	637	những	246
66	phát	540	qua	409	nhưng	637	nhưng	192
67	qua	505	quan	427	nó	639	nó	194
67	quan	603	quốc	950	nói	955	nói	367
67	quốc	537	quyền	723	nữa	321	nữa	88
70	ra	968	ra	802	nước	398	ở	207
71	rất	678	sau	468	ở	598	ông	112
72	sau	585	sẽ	478	ông	927	phải	164
73	sẽ	934	số	553	phải	590	qua	86
74	sinh	735	sự	670	quá	311	quá	85
75	số	759	tại	712	ra	1,179	ra	245
76	sự	774	tháng	471	rất	452	rất	140
77	tại	995	thành	697	rồi	895	rồi	221
78	thành	901	thế	642	sao	304	sao	92
79	thế	711	thể	475	sau	429	sau	109
80	thể	948	theo	478	sẽ	602	sẽ	126
80	theo	745	thì	463	ta	782	ta	281
82	thi	635	thứ	478	thấy	953	thấy	195
83	thì	736	tôi	979	thật	322	thật	100

APPENDIX B (Continued)

	Adult VN	Freq	Adult O	Freq	Child VN	Freq	Child O	Freq
84	thời	497	trên	630	thế	566	thế	119
85	thông	485	trong	1,649	thì	683	thể	133
86	tôi	1,137	trung	527	thỏ	417	thì	139
87	trên	761	trước	456	tiếng	317	tiếng	109
87	trong	1,979	từ	696	tìm	320	tới	98
87	trung	600	tự	546	tôi	615	tôi	429
87	trước	487	và	2,956	trên	478	trên	148
91	trường	903	văn	502	trong	729	trong	294
92	từ	971	vào	877	từ	434	từ	107
93	và	3,198	về	780	và	1,608	tường	85
94	vào	988	vì	523	vào	911	và	881
95	về	1,069	việc	469	về	714	vậy	115
96	vì	577	việt	878	vì	309	vào	202
97	việc	820	viên	405	với	623	về	156
98	việt	632	với	1,040	vừa	419	với	192
99	viên	528	vụ	451	vua	410	vừa	98
100	với	1,657	2006	514	xuống	413	xuống	116

(Manuscript received March 3, 2007;
revision accepted for publication April 11, 2007.)