ANALYSIS OF LANGUAGE VARIATION AND WORD SEGMENTATION FOR A CORPUS OF VIETNAMESE BLOGS: A SOCIOLINGUISTIC APPROACH

by

HEATHER LEE MELLO

(Under the Direction of William Kretzschmar)

ABSTRACT

This dissertation examines issues of units of meaning, word segmentation and language variation in a corpus of Vietnamese language blogs collected from publicly accessible internet sources originating in Viet Nam, the US, and Australia. Research using corpus linguistics techniques for study of the Vietnamese language have begun to proliferate in western sources in the past decade, however, studies using language-in-use data remain rare. Analysis of the corpus as a whole and by comments and blogs and Viet Nam, US, and Australia subcorpora used the Vietnamese syllable, or tiếng, as the basic unit of meaning, with subsequent iterations of one- through 5-tiếng. While results support previous research asserting the Vietnamese syllable as the basic distributional element in Vietnamese discourse, claims about Vietnamese as a monosyllabic language are not supported by results. Tiếng collocate and colligate meaningfully and regularly throughout the corpora in clusters larger than one syllable, indicating that syllable combinations, the union of tiếng (Nguyen, 1984), are also primary distributional patterns for the Vietnamese language. Varieties of Vietnamese by country show similarity in a variety of distributional patterns, including by a-curve

(frequency of frequencies), structural, content, and units of meaning analyses. Variations of Vietnamese by country are primarily limited to collocational and colligational content and topical patterns.

INDEX WORDS: Vietnamese Language, Corpus Linguistics, Sociolinguistics, Word Segmentation, Unit of Meaning, A-Curve, Blogs, Internet, Diaspora, Language Variety, Tiếng

ANALYSIS OF LANGUAGE VARIATION AND WORD SEGMENTATION FOR A CORPUS OF VIETNAMESE BLOGS: A SOCIOLINGUISTIC APPROACH

by

HEATHER LEE MELLO

B.S., University of the State of New York, Regents College, 1994

M.A., Georgia Southern University, 2003

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

ANALYSIS OF LANGUAGE VARIATION AND WORD SEGMENTATION FOR A CORPUS OF VIETNAMESE BLOGS: A SOCIOLINGUISTIC APPROACH

by

HEATHER LEE MELLO

Major Professor:	William Kretzschmar
Committee:	Dezso Benedek
	Lewis Chad Howe

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

DEDICATION

This dissertation is dedicated with love and gratefulness to my parents and my teachers.

ACKNOWLEDGEMENTS

I would like to give my thanks to all those who have helped me reach this goal. To my committee, past and present–Drs. William Kretzschmar, Linda Harklau, Dezso Benedek, and L. Chad Howe. To Trang Nguyen for generous hours of discussion and assistance with analysis. To the UGA Graduate School for award of the Summer 2012 Doctoral Research Fellowship. To the Center for Research on Behavioral Health & Human Services Delivery and the UGA Departments of Comparative Literature and Language & Literacy Education for assistantship support throughout my graduate studies. To the U.S. Department of Education Fulbright-Hays Group Projects Abroad program for a Vietnamese Advanced Summer Institute Fellowship to study advanced Vietnamese in Viet Nam. To the UGA Linguistics Program for taking me in and believing in my future as a linguist. And most especially to my friends and family for helping me through this process.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Despite the overwhelming dominance of English language corpora in the field of corpus linguistics today, the practice of recording language data in the form of corpora began with languages other than English (McEnery & Ostler, 2000). This fledgling field included Boas' recording of dying languages on paper in the 1940s, humanities computing and the development of machine-readable corpora for the study of historical languages like Latin and Aramaic in the 1940s and 50s. It also includes so-called "mechanolinguists" who produced machine-readable corpora in the 1950s for such typologically different languages as Chinese and Romanian (McEnery & Ostler, 2000). The field of linguistics veered away from broad use of multilingual corpora in the 1960s with the advent of the Chomskian paradigm, which advocated against the use of corpora as language data, especially among North American linguists (McEnery & Ostler, 2000). Nevertheless, Firthian linguistics, with its focus on language in use as data, remained popular in Britain and northern Europe. As a result, English language corpora began to predominate in the 1960s, a trend that has continued to the present day.

In recent years, the global study of corpus linguistics has seen a resurgence. While Chomsky's theories remain popular, they no longer dominate the field as they did in the early days of his new theory of generational grammar. Technology has developed dramatically, enabling the storage and analysis of banks of corpus data to an

extent unimaginable in the 1960s. Where once a million-word corpus seemed large, standard corpora are at least 100-million to 1-billion words and larger (Hundt, Nesselhauf, & Biewer, 2007). Processing speeds have gotten markedly faster and the facility with which one may tag and analyze corpus data has improved drastically; no longer is a simple concordance sufficient as analysis of a language corpus.

Another long overdue update since those early days of corpus linguistics is the resurgence of language corpora and corpus analytic research methods for languages other than English. While many technological developments in corpus analysis remain focused on English, updates in the ability to process writing systems other than English have also grown. For example, the development of Unicode character sets for human language encoding allows representation of potentially all the world's languages, lessening barriers to creating corpora for non-Latin scripts (McEnery & Olster 2000).

1.2 Corpus Linguistics and Vietnamese

Vietnamese is one of those world languages that was largely left out of initial corpus building efforts (McEnery & Ostler, 2000). Vietnamese uses a modified Latin alphabet. At present there is a small, but growing body of corpus research for Vietnamese. It is a smaller national language, with only about 88 million speakers nationally and a diaspora estimated at around two million speakers.

In the United States, Vietnamese are the 4th largest of all Asian groups living in the US (US Census Bureau, 2010). The Vietnamese population, at 1.55 million persons, is 0.5% of the total US population at 309 million persons (2010). Studies of the Vietnamese language are rare compared to other languages with longer traditions in

the US and with a much larger diaspora presence. For example, compare Vietnamese to Italian, a country with a smaller number of speakers – 88 million compared to just over 60 million speakers within country albeit with a much larger diaspora and much of that diaspora living in the US. Yet, Italian is considered a classical language and will fall into most universities' Romance language departments, whereas Vietnamese language studies in North America tend to be limited to areas where there is a significant heritage language student population, a reflection of the fact that size is not everything.

Much of the literature on Vietnamese language processing by Vietnamese and Vietnamese diaspora scholars has yet to find its way to mainstream Western scholarly venues. For example, a search using the scholarly database Linguistics and Language Behavior Abstracts revealed only one article concerning Vietnamese corpus linguistics as of January 2012 by Pham, Carney and Kohnert (2008). However, an expanding body of work is beginning to appear on the web and in regional publications. Google searches in English and Vietnamese reveal at least twenty corpus linguistics (CL) and natural language processing (NLP) related articles written in English after 2002 by Vietnamese language scholars within and outside of Viet Nam. Studies in this area use corpus methods to investigate linguistic phenomena, describe methods for corpus creation, and examine the best natural language processing techniques for Vietnamese. This dissertation aims to contribute to this nascent body of research using Vietnamese corpora.

1.3 Meaningful Units and Methodology

This dissertation focuses on the following three research questions:

1. *What are the most common syllable forms and collocational, colligational, and topical patterns as revealed in a corpus of in-country and US and Australian Vietnamese language blogs?*
2. *What are the most common syllable forms and collocational, colligational and topical patterns as revealed in a comparison between in-country and US and Australian Vietnamese language blogs?*
3. *In corpus analysis of Vietnamese language varieties, what are the implications for how we analyze data? What are the implications for existing theory concerning segmentation into meaningful units in Vietnamese? How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?*

Through analysis of a corpus of informal web-based Vietnamese, Vietnamese language variation in in-country and diasporic contexts, and an exploration of the appropriate unit of analysis for linguistic study of the Vietnamese language, this new research provides insight into several key issues in contemporary corpus language study.

As mentioned above, one issue of abiding concern in this burgeoning work concerns the appropriate unit of analysis for study and parsing of the Vietnamese language. Studies throughout the history of Vietnamese language research and much of the current NLP literature concern if and how the construct of "word" -a fundamental unit of analysis in English corpus studies-applies to Vietnamese. *Tiếng* is the Vietnamese term for a one-syllable form, separated by spaces in writing, which may or may not correspond to the notion of 'word' in English. As Vietnamese is an isolating language, wherein spaces are used to separate syllables, it may seem natural to assume that Vietnamese is a monosyllabic language. In this view, each monosyllable would correspond to the English concept of a word, a complete discursive unit able to stand on its own in each instance where there are spaces between words. On the other hand, much of the literature concerned with the Vietnamese lexicon and word segmentation has suggested that 'words' may consist of 1, 2, 3, 4, or extremely rarely, even 5 'tiếng' or syllables (Nguyen, Nguyen, Phan, Nguyen, & Ha, 2006; Dinh, Hoang & Nguyen, 2001; Le, 2003).

As such, the true nature of how to segment the Vietnamese language into meaningful units is even now still being defined. Hieu, Vu, and Kien (2010) state, "There does not exist an algorithm that segments a given Vietnamese sentence into words exactly according to its meaning if the sentence is considered isolated." As an example, the article gives the following sentence: "*Cái bàn là của tôi*" (2010). If one segments the sentence in this way; "*Cái / bàn / là / của / tôi*," then the sentence means, "The table is mine." If one segments the sentence in this way; "*Cái / bàn là / của / tôi*," the sentence means, "The iron is mine."

While questions regarding word segmentation, whether Vietnamese is a monosyllabic language, and the nature of the word in Vietnamese have become the focus of much of the CL and NLP literature, this dissertation will take a new approach. In Chapter 2 of Trust the Text, Sinclair (2004) makes a case for analysis based on "units of meaning". While the word has traditionally been the basic unit of meaning for much CL literature, especially for European languages, some American linguists put the morpheme forward as the most basic unit of meaning. For Vietnamese, the concept of word is not a given. Morphemes themselves may consist of one or multiple syllables. Conversely, one syllable may equal one unit of meaning which would correspond to both notions of morpheme and word.

The question of meaningful units for the Vietnamese language then, is a complicated one. Taking a usage perspective, this dissertation will provide a classic corpus analysis format featuring general description of co-occurring language forms and topical characteristics of the entire corpus as a whole, with the syllable and subsequent iterations of syllables as the basic units of analysis.

1.4 Vietnamese: In-country and in the Diaspora

It has been said that the United States is where languages go to die. Indeed despite some folk beliefs that immigrants are not learning English, research shows that immigrants today, like their forebears, tend to shift towards English monolingualism in the U.S. within three generations of immigration (see, e.g., Achugar, 2008). However, while this may be the norm across immigrant communities generally, the factors relating

to maintenance and shift for various immigrant groups and the outcomes derived may differ quite a bit.

In her dissertation about language attitudes and behaviors among a Vietnamese refugee community in the northern Virginia area, Clare O'Leary (1989, p. 66) suggests that Vietnamese Americans may present a special case for language maintenance. She says,

> "Because of the positive attitudes towards Vietnamese, and the factors contributing to its maintenance, the language may be maintained longer in this community than has been the case in some other immigrant situations. However, this group is under different pressures than many of the previous immigrants were because they are not from a European background, they initially arrived as refugees with no option to return and were required to acclimate abruptly to the changes in their social and cultural environment. Because of its uniqueness, further research should be done to determine the outcome of these factors on language maintenance and shift" (O'Leary, 1989, 66).

It has been over twenty years since O'Leary wrote her dissertation, but the question remains, what is the outcome of the sociolinguistic factors she observed and what is the current state of the Vietnamese language in the United States?

The same sociopolitical forces that shaped the sizable US Vietnamese diaspora also created a sizable Vietnamese speaking diaspora where English is dominant in Australia. The history of the Vietnamese in Australia shares many similarities to that of the US. Points of difference are typically related to Viet Nam's proximity to Australia.

For example, the first thirty-eight Vietnamese immigrants to land in Australia sought refuge after a storm blew their ship off course in 1920 (Tuc, 2003).

This dissertation thus explores whether Vietnamese-language use between these populations has begun to quantifiably differentiate. It asks whether contact phenomena with English in the US and Australia and rapid additions to the language occurring in Viet Nam have caused divergences in Anglophone diasporic and in-country varieties. It uses language-in-use data and examples to provide empirical documentation of and substantiates whether hypothesized and anecdotally observed linguistic differences actually exist on a broader scale.

1.5 Blogs as Corpus

This study also breaks new ground by exploring the linguistic features of collected informal Vietnamese-language writing on the internet, specifically blogs and blog comments. Some of the advantages to these forms of communication in examining language change and difference are that although primarily a written medium, blogs and the comments to their entries are highly personalized text types and thus tend to feature both formal and informal aspects of language use. The immediate, interpersonal nature of written blogs reveals features also found in informal spoken genres, to include dynamic, creative use of language; ad hoc neologisms, borrowings, code-switching and other uses of language not found in more formal and restricted discourse types. Hundt, Nesselhauf, and Biewer (2007) describe these forms of language in use as new text types unimagined by early corpus creators, types that use a "written medium, but (that) are obviously much closer to the patterns we expect to see in spoken interaction" (p. 1).

Using a snowball sample technique, blogs accessed initially were then checked for references to other blogs. On most blogs, users have the option of posting links to blogs they themselves read, either in the profile section for some sites or as a list in a column featured alongside the main blog content column. Also, fellow bloggers and blog links could be found in the content of the blogs themselves or referenced in the comments section, especially when the referenced blogger had written about similar topics. As such, data in this corpus could not be considered a random sample of Vietnamese language data across in-country and English-speaking diaspora sources, but rather a representation of language used in a series of intersecting networks.

1.6 Dissertation Format

This introductory chapter has provided some background and motivation for this present study. Chapter 2 examines the literature associated with corpus linguistics, especially for Vietnamese, sociolinguistics associated with diaspora community languages and the sociolinguistics of social networks in general and in relation to blogging and internet communication in particular. Chapter 3 lays out the sampling procedures and methods used to create and analyze this corpus of Vietnamese language. Chapters 4 and 5 discuss findings for the corpus as a whole and as divided into subcorpora representing blogs and comments. Chapter 6 discusses findings for the three subcorpora representing in-country and diaspora Vietnamese varieties. Chapter 7 concludes this dissertation by discussing the implications of this study for studies of the Vietnamese language in particular and internet writing and corpus analysis techniques in general.

1.7 Conclusion

In all then, there still exist gaps regarding the development of corpus analytic techniques to study varieties of the Vietnamese language. Further, while there is a growing body of research concerning corpus techniques as applied to the Vietnamese language, due to proposed divergence between so-called 'in-country Vietnamese' varieties and overseas Vietnamese varieties, the use of corpus techniques to compare varieties is still nascent. This current work is a partial attempt to further this investigation. Using a combination of corpus analysis, linguistic content analysis and discourse analysis methodologies, this dissertation examines informal Vietnamese-language writing on the internet originating in Viet Nam and in the United States and Australia.

CHAPTER 2

REVIEW OF LITERATURE

This chapter provides a review of the literature related to this study's research questions. The first section reviews corpus linguistics and relevant aspects of the Vietnamese language in detail. It explores several methods associated with Vietnamese language research using corpus and other computer-mediated methods of analysis, with a particular focus on units of meaning for analysis of Vietnamese. The second section examines issues associated with language variation in both in-country and diaspora community contexts, especially related to the Vietnamese language. In the final section, the sociolinguistics of social networks in general and in relation to blogging and internet communication are discussed.

2.2 Corpus Linguistics and Vietnamese

Much of the literature on Vietnamese language processing for corpus linguistics (CL) and natural language processing (NLP) has been focused on the issue of the proper unit of meaning and how to segment the language into meaningful units for analysis. In his seminal study on Vietnamese grammar, Thompson (1965) discusses the three units of analysis which comprise much of the discussion regarding how to parse Vietnamese morphology and syntax; namely, the syllable, the morpheme and the word.[1]

[1] "While morphemes are the smallest isolable units which convey meaning and are the ultimate constituents of any sentence in any language, it is difficult or impossible to understand the structure of sentences as simple sequences of morphemes. It is necessary to identify larger units, themselves composed of morphemes, which appear in varying

Tiếng is the Vietnamese term for the one-syllable form which has traditionally been considered the most basic unit of meaning in Vietnamese. As Vietnamese is written using spaces to separate syllables, it may seem natural to assume that Vietnamese is a monosyllabic language. In this tradition, each syllable would equate to a full unit of meaning most closely aligned with the concept of word, which Thompson (1965, p. 116) defines as, "the minimum freely distributed units of which sentences are composed." According to Thompson (1965), this attitude is likely based on factors related to the first Vietnamese writing systems, borrowed from classical Chinese, wherein one character represents one syllable and where syllables tend to be considered words, in most cases. In this view then, for Vietnamese each monosyllable would correspond to the English concept of a word, a complete discursive unit able to stand on its own in each instance where there are spaces between words. Supporting this notion, Ngo and Hoai write in 2001, "In Vietnamese, each morpheme, which in phonetic respects in most cases is a syllable, tends to form a separate word" (p. 11).

The use of 'tends' above should be highlighted; however. While a large proportion of the Vietnamese lexicon is monosyllabic and monomorphemic (especially for the spoken language), many morphemes, considered the smallest indivisible forms of meaning, consist of more than one syllable. Examples from Thompson (1965) include *thình lình* 'sudden', *Sài Gòn* 'Saigon', and *com mi nít* 'communist' (as a

---

relationships to one another as constituents of sentences. These larger units are words: they are the minimum freely distributed units of which sentences are composed.

Traditionally each Vietnamese syllable has been regarded as a word. As a matter of fact, a great proportion of Vietnamese words-especially those most current in the spoken langauge-are just one syllable long. The generalization is presumably to some extent based on this fact. Probably equally strong in the traditional attitude is another factor; Vietnamese was first written with symbols from the Chinese, and Chinese characters typically represent one syllable each and are traditionally considered to represent as well one word.

However, it is not really accurate to say that each Vietnamese syllable is a word or each word a syllable" (Thompson, 1965, 116-7).

borrowed term, *com mi nít* may not be analyzed according to its French roots). Conversely, there exist polysyllabic forms containing individual syllables that carry no meaning, especially for reduplicative forms such as *thình lình* above, where neither form individually, *thình* or *lình,* has meaning of its own. The concept of 'word' for Vietnamese, then may be defined as either mono- or polysyllabic forms which represent complete discursive units able to stand on their own.

Taking this definition as a starting point,, 'words' in Vietnamese may consist of 1, 2, 3, 4, or extremely rarely, even 5 *tiếng* or syllables (Nguyen, Nguyen, Phan, Nguyen, & Ha 2006; Dinh, Hoang & Nguyen 2001; Le 2003). Ngo and Hoai (2001) further refine this discussion, listing word types in Vietnamese, including poly-syllabic simple words and those that necessarily consist of two or more syllables such as the various forms of compounds. They also include reduplicative forms like those mentioned above – an extremely productive and creative word class in Vietnamese. Le (2003) claims that the Vietnamese lexicon is mostly disyllabic, especially concerning adjectives and reduplicatives. Likewise, in writing about compounding processes in Vietnamese specifically, Nguyen (1997) asserted that up to 80% of Vietnamese words consisted of two syllables.

This issue is further complicated when considering the difference between spoken and written Vietnamese. While monosyllables do comprise a large part of spoken Vietnamese, written Vietnamese derives heavily from Sino-Vietnamese forms and is more likely to contain two-, three and even four-syllable word forms. According to Ngo and Hoai (2001), up to seventy percent of written Vietnamese was borrowed from classical Chinese. In referring to earlier practices of using hyphens to connect

multi-syllable forms in Vietnamese writing, Thompson (1965) notes that there was quite a lack of uniformity, relating to "the problem of just what a word is in Vietnamese" (p. 74).

According to Pham, Tran and Pham (2009), there are two types of ambiguity that arise when segmenting sentences into words in Vietnamese, cross ambiguity and overlap ambiguity. The first type, cross ambiguity, occurs when the individual syllables in a phrase or potential compound have meaning and could stand as individual words themselves. Using the example sentence from Chapter 1, *Cái bàn là của tôi*[2], the table (*bàn)* versus iron (*bàn là)* ambiguity would require an understanding of the larger context in order to segment the sentence correctly (Hieu, Vu, & Kien, 2010). One could imagine instances, however where even context would not completely resolve the issue, if for example, one is merely listening to a conversation about dividing up possessions in a house, then one might still be unsure whether the thing involved is a table or an iron.

The second type of ambiguity, overlap ambiguity, occurs when individual syllables could meaningfully be segmented with previous or following syllables (D.D. Pham, et al., 2009). As an example, the authors provide this sentence, "*Tốc độ truyền thông tin ngày càng cao,"* where "*truyền,"* "*truyền thông,*" "*thông tin,*" and "*tin*" are all possible words (2009). The sentence could mean, "*The speed of information communication grows daily*" or "*The transmission speed of information grows daily,"* among other meanings. According to the authors, cross ambiguity is the rarer and

[2] From Chapter 1, section 1.3: "As an example of this, the article gives the following sentence: "*Cái bàn là của tôi*" (Hieu, Vu, & Kien, 2010). If one segments the sentence in this way; "*Cái / bàn / là / của / tôi*," then the sentence means, "The table is mine." If one segments the sentence in this way; "*Cái / bàn là / của / tôi*," the sentence means, "The iron is mine.""

more easily parsed of the two forms, while overlap ambiguity is the more common and more complex form. They assert that resolution of overlap ambiguity problems in particular would improve the issue of word segmentation in natural language processing tasks for Vietnamese.

This problem becomes even larger and more complex in taking new words into account. In an article by the head of the Vietnamese Corpus Linguistics Group (Corling), Dao Hong Thu (2011) examines the problem of automatic machine analysis of new scientific and technological terminology for the Vietnamese language. She states that the body of general and specialized scientific terminology grows daily under the influence of rapid scientific and technological growth in Viet Nam. According to the Corling group's research, 87% of new "word units" are scientific and technological terms, especially regarding the field of information science. Currently, according to Dao, there are as yet no unified standards for the use of such terminology (2011). Terms from the past tended to be from Sino-Vietnamese origin, which meant that they lacked an essential Vietnamese character[3], but also that the terms lacked scientific specificity. Terms introduced and used more currently tend to be borrowed primarily from English, as well as other languages, but still, according to the author, these terms lack specificity and at times are even applied incorrectly despite their widespread dissemination and use.

Another challenge in defining words in Vietnamese is that the Vietnamese lexicon is rapidly growing, and the state of Vietnamese lexicography in general and

[3] Something noted as such by Dao, which implies the perceived necessity of finding Vietnamese forms for new ideas.

Vietnamese dictionaries in particular are at this point inadequate to the task of a full and widely accepted representation of this Vietnamese lexicon (Dao 2011, Hieu, et al. 2010; D.D. Pham, et al. 2009; D. Nguyen 2009; Q.T. Dinh, et.al. 2008; T.V. Nguyen, et al. 2005; Dinh & Hoang 2003; Vu, et al. 2003; Le 2003; D. Dinh, et al. 2001).

Current methods for assessing and describing the full lexicon in specialized CL and NLP studies generally rely on one or a combination of several basic approaches (D.D. Pham, et al., 2009; Dao, 2011). As described by Dao (2011), the *linguistic* approach involves manually handling the data, usually by linguistic specialists, in order to extract syllables defined as words. In extraction, the researcher must analyze collocational and colligational patterns in single syllable and increasing iterations up to multi-syllable units. In order to refine the analysis for specialized analyses, filters are used to block syllable combinations that are not the focus of the study.

Automated approaches include those termed *dictionary* and *statistical* methods. The dictionary approach uses an automated matching algorithm to identify words by comparing them against a dictionary inputs (D.D. Pham, et al., 2009). One noted weakness of this approach is limitations in the state of Vietnamese lexicography in general and Vietnamese dictionaries in particular. The third approach, termed the statistical method by Pham, Tran and Pham (2009) as well as Dao (2011), involves automated identification of patterns based on frequency of occurrence and matching against concepts expressed for similar article types. Word identification is derived through probabilistic determination from common patterns for similar linguistic content wherein n-grams are derived from recurring continuous sequences of items in a corpus. As described for Vietnamese, an n-gram of one unit, or a 1-gram for Vietnamese would

consist of one syllable, a 2-gram would consist of 2 syllables of recurring continuous text, and on up to the proposed 5-gram sequence which correlates to the highest possible number of –grams for Vietnamese words.

The fourth and final approach is called the *hybrid* approach (D.D. Pham, et al., 2009; Dao, 2011) and consists of a blending of one or more methods as given above. For the linguistic method, use of the dictionary method is a common way of confirming word segmentation approaches. Pham, Tran, & Pham (2009) used a combination of the dictionary and statistical approaches to reach a 97% accuracy rate for word segmentation processes that matched words as noted in their dictionary files. While the bulk of the NLP literature uses the automated dictionary and statistical approaches, or some hybrid thereof, often linguistic judgments are used to confirm the accuracy of word segmentation results. This correlates to the linguistic approach inasmuch as linguists manually confirm and manipulate the corpus to confirm findings for the other approaches (Dao, 2011; Q.T. Dinh, et al., 2008).

Whether one uses a linguistic, automated or a hybrid approach in CL or NLP research, accurate segmentation is crucial in any research on discourse and meaning in Vietnamese. In the earliest Vietnamese corpus article currently available in the Linguistics and Language Behavior Abstracts database, Pham, Carney, and Kohnert (2008) investigate the effects of “publication place” on linguistic content by comparing Vietnamese language texts published in Viet Nam and in the West. Despite noting that Vietnamese words may contain from one up to two or more syllables, the authors chose to only examine and compare language content at the one-syllable level. As to differences between the language forms obtained between newspaper articles in Viet

Nam and the West, the article found none (Pham, Carney, & Kohnert, 2008). The lack of more clear findings may be attributable to the paper's methodology of using only one-syllable forms in assessing language difference. By overlooking subsequent iterations of syllable combinations, the authors likely missed much semantic and discursive content beyond the most common one-syllable forms.

Questions regarding differences in language content and not simply syllable lists would find only a limited answer in a paper that overlooked assertions by linguists as to the nature of Vietnamese as more than a monosyllabic language. For example, the word *người* 'person' in Vietnamese may indeed be a mono-syllabic form. If; however, that word is modified by a syllable entailing a nationality, as in '*người Mỹ'* 'American', as in an American person, then to restrict analyses to only a monosyllable is to overlook discourse as meaning and to only use the barest form of syllable iterations possible. As such, there is a significant need for studies that use keywords, collocates and colligations, semantic clusters and syntactic patterns making use of potential 1-, 2-, 3-, 4-, and 5- syllable forms and perhaps even longer strings of text.

One addressed weakness as noted above in discussion of different approaches to segmentation in Vietnamese is the issue of determining meaning, if lexis is to be the central focus in this study. Most studies note the insufficiency of current Vietnamese lexicography and available dictionaries. Therefore, using available dictionaries to determine virtually all semantic boundaries within a corpus would be incomplete as Viet Nam is a rapidly developing nation, growing its lexicon amidst intense change. One advantage of a corpus is to reveal patterns that do not exist in current dictionaries, especially if the patterns recur and the environments make the meaning transparent.

This may not be available. However, the fact that studies coming from Viet Nam often use linguists for inter-rater reliability in semanticity, lexical and word boundary studies indicates that while the focus is currently on natural language processing techniques for an available easily semantically, lexically and grammatically parsable corpus, the contribution of the individual linguist is still of benefit.

2.3 Sociolinguistics and Diaspora Community Languages

In the past, language contact theory consisted primarily of contrastive analyses of differences between languages as their speakers come into contact with one another and within the minds of bilinguals (Weinreich, 1963). These theories have been updated to include concepts of space, which take into account both migration and the growing use of media that closes gaps between persons and communities and makes communication between distant groups possible (Collins, Slembrouk, & Baynam, 2009, Tuc, 2003). To connect language use within Vietnamese-speaking communities and language contact theory, this dissertation will use corpus-analytic methods to ask the question, 'How does Vietnamese language use differ among contemporary Vietnamese language users in informal electronic media?' Specifically, one question that accounts for space and the effects of contact asks how does language differ between speakers in Viet Nam and speakers in the wider Vietnamese speaking diaspora.

Language contact pressures in the US are intense and the statistics relating to notions of cultural and civic assimilation – defined by Vigdor (2008) as the tendencies for groups to reach statistical parity with the mainstream (or aggregate) on cultural and civic factors – are comparatively high for Vietnamese in the US. As part of the definition

of assimilation, Vigdor notes statistics relating to English speaking, homeownership and rates of naturalization for Vietnamese immigrants that are among the highest for all named groups, especially for a group with such a recent history in the United States. Ho Dac Tuc (2003), in his dissertation on Vietnamese-English bilingualism and code-switching in Australia noted much the same effects. Migration as much as time is a significant factor in language change and language contact theory notes the pressure of being a minority language community within a language majority community as matter of fact variables towards language change. In this, it would seem that Vietnamese language in the Australian context is under the same sorts of pressures as Vietnamese in the US, especially as most migration occurred to both countries in the aftermath of war and a clamp down by the Vietnamese government on travel outside Viet Nam until opening the borders significantly in the early 1990s.

According to the Ben-Moshe and Pyke (2012), 2006 census numbers place the Vietnamese population in Australia as the 5th largest in the world outside of Viet Nam at over one hundred fifty nine thousand persons, after the US, Cambodia, France and Taiwan. The Vietnamese represent 0.65% of the overall Australian population of around 22.9 million persons (Australian Bureau of Statistics 2013). In comparison, the Vietnamese population in the US is over 1.5 million persons at 0.5% of the total US population.

Vietnamese is the seventh largest spoken language in Australia (Tuc, 2003) and over 90% of the Vietnamese population reports speaking Vietnamese well or very well (Ben-Moshe & Pyke, 2012). Conversely, only 56.5% of Vietnamese Australians report speaking English well or very well by 2006 census results (2012). Furthermore, as

English is the dominant language, Vietnamese “is by no means free of English influence” (Tuc, 2003).

While contact with English is stronger and more immediate for Vietnamese speakers in the West, changes are also occurring to Vietnamese in Viet Nam as well. Internet conventions, the existence of English as a global and world wide web lingua franca, the Socialist Republic of Viet Nam’s language policy of promoting the learning and use of English, and the rapid scientific and technological development accompanied by additions to the Vietnamese lexicon from other languages in Viet Nam, especially in cities and at universities, point to language contact phenomena occurring for Vietnamese speakers in country as well. This author expects that Vietnamese-language use between the two populations has begun to quantifiably differentiate, with much of the difference caused by contact phenomena with English in the US and rapid additions to the lexicon occurring in Viet Nam, changes which may or may not be transferring into Vietnamese in the US.

In comparing the use of Vietnamese language in Vietnamese-ethnic homes, Clare O’Leary’s dissertation on language maintenance in Vietnamese families (1989) found that although families generally assigned Vietnamese language maintenance a high priority, when it came to actual behavior, these same respondents noted that they tended to focus on English or to not use Vietnamese as much or in as many domains as they felt ideal to maintain ties and skill in Vietnamese language. This practice was even more marked for children in these homes, children who tended to use English more with siblings, peers and when helping parents to learn English. In the twenty-two years

since the dissertation was written, these children may now be parents themselves, without the skills to pass the language intact onto their own children.

O'Leary's (1989) report is based on self-report, interviews and observations, but contains no actual language data with which to compare language change over time. Although she notes informant reports of language use, such as census data, we, the readers must rely on that methodology as evidence of use in the home, the quality and quantity of the Vietnamese language used, and for concepts of dialect and language description. While reports of the difference between language attitudes and language behavior are important, the lack of data provides no baseline with which to compare. Pham (2002), attributing changes to the gender and kinship reference system in Vietnamese, describes the general pronominal referent system in use, where relations genders and social classes have become increasingly characterized by qualities of inclusion and equality. The "symmetries," as Pham (2002) calls them, between reference pairs in dialogue and social interaction have changed over time, and especially with French, Vietnamese Communist and modernist ideals of equality between the sexes, in the sense of hierarchical relations within families and between persons living in proximity with one another, with more equal relations between insider and outsider and employer and employee even without specific reference to gender.

While most of the referents noted by Pham (2002) refer to changes over time within Viet Nam, she further argues that the range of terms and the frequency of their use are being changed by notions of equality between the genders and on the levels between worker and boss. As such, we might then expect to see change in American Vietnamese as well; in the frequency and use of terms and more similar sorts of

pairings apart from expected reduction in the contact situation with English, with its comparatively very limited system of person reference. Address forms, to include the proposed changing gender and kinship symmetries given in the paper are provided as exemplars only however, no other language data are provided against which later research may be compared.

In a later paper by Pham (2008), which addresses the teaching of Vietnamese dialects, she again asserts that changes have taken place between the Vietnamese as spoken in Viet Nam and by overseas Vietnamese. Here, she notes changes to dialects within Viet Nam and alludes to the beginning of an overseas Vietnamese dialect that differs from the language as spoken in Viet Nam, changes attributed to migrations and institutional changes since reunification of the two Viet Nams (North and South) in 1975 as well as the emigration out of country of many Vietnamese after the end of the war. Changes to internal dialects, Pham (2008) indicates, are due to the convergence of formerly more divergent northern and southern dialects in Viet Nam as many northern speakers migrated into the south as well as the influence of the media and educational institutions in daily life after the war which privilege northern dialect varieties. In addition, the prevalence of Chinese words in Vietnamese varieties, as noted by Ngo and Hoai (2001), may be expected to differ qualitatively and quantitatively. As noted above, while it might be expected that Vietnamese in the West will have more English forms, to include lexical and syntactic patterns, these speakers, especially those who left Viet Nam before and immediately after 1975, might also use more of the traditional Chinese forms that had been discouraged in Viet Nam under the national language planning policy, which encouraged use of Vietnamese forms for concepts formerly

rendered in Chinese. Therefore, Vietnamese in Viet Nam would be likely to use less of the older Chinese forms and more of the newer Chinese forms, which have resulted from contact between Viet Nam and China since the major 1975 migration and language planning policy.

As to the nascent overseas Vietnamese variety, Pham (2008) writes matter of factly about expected changes when political strife necessitates out migration, such as in the cases of Cubans and Hungarians who also left their home countries during political turmoil. Language changes for overseas Vietnamese have become "a sensitive issue," whereby newly coined words in-country are associated with communism and so the language for these overseas speakers remains "frozen" in time (Pham 2008). If her prediction that "after as little as 30 or 40 years, such differences can be quite noticeable" (Pham 2008) holds, then language data comparing the two major populations should be evident. I would add; however, that with the waning of post-1975 reunification communist fervor, many forms that had gone out of favor may have returned, such as with formal hierarchical terms like *ngài* 'you' (as used in a highly formal or diplomatic sense) and those formerly discouraged Chinese forms mentioned above. That these terms had gone out of use and then come back into use in Viet Nam due to unique political pressures not experienced by overseas Vietnamese in the US indicates that the use of the term, as evidence of changes to the lexicon, may be quite different than expected.

In the conclusion to her 2002 paper, Pham refers to the new referent system order in Vietnamese generally as "more flexible" and that "the changes we see in use of terms of address, self-reference and reference are an indicator of social change"

(p.310). The article is a detailed description of the reference system with notes to how changes have occurred over time with changes in the political structure and national make-up of the government. Her 2008 article focuses on pressures within the Vietnamese teaching field in the US that link social and political identity to dialect, especially linking a pre-1975 northern dialect with the old communist regime. For both articles, changes are attributed to time periods and dialectal area, but as with the O'Leary dissertation (1989), no data sampling with which to confirm these assertions is provided.

Having language-in-use examples of the above phenomena would go far to demonstrate the existence of qualitative differences between in-country and diaspora varieties of Vietnamese. Clearly there is a need for research providing empirical documentation of whether hypothesized and anecdotally observed linguistic differences actually exist on a broader scale. In all then, there still exist gaps regarding the development of corpus analytic techniques to study varieties of the Vietnamese language. Further, while there is an exploding body of research concerning corpus techniques as applied to the Vietnamese language, due to proposed divergence between so-called 'in-country Vietnamese' varieties and overseas Vietnamese varieties, the use of corpus techniques to compare the two varieties is still nascent.

2.4 Blogs and Social Networking

The use of corpus linguistics for this study reflects linguistics as a social science, wherein language is a means for human interaction, bound by time and culture. In addition, as in Tuc (2003) and Collins, Slembrouk, & Baynham (2009), the concept of

space is another factor affecting variability in language. In linguistic studies of language and dialect, patterns of language across speakers is usually attributed to the social and geographic space in which people live their daily lives (Kretzschmar, 2009). People who live near or interact in meaningful ways with one another tend to share more language characteristics, while persons who do not may not.

For this study the digital arena of social media is another factor affecting language patterns. Geographical space distances can be mediated by the use of social media and the internet, with distance as a factor affecting language behavior that may be lessened by the use of the internet data and participation in social media and especially, digital social networks. From the perspective of language as a reflection of human interaction, this study uses language in use, derived from naturally occurring samples. These samples, though do not derive from random sampling or from surveys administered to enlisted populations, but rather represent a snowball sample of blogs, wherein lists of fellow bloggers provided on the page of the initial blog and subsequent lists on those bloggers' pages formed the whole of the sample. As such, the data for this dissertation represent an informal series of intersecting social networks.

Beginning with Milroy and Li (1995), social networks are defined as "a boundless web of ties which reaches out through a whole society, linking people to one another, however remotely" (p. 138). Generally speaking, these patterns of social interaction affect the language patterns of individuals within these social networks - how persons in these networks speak with each other and with those in the larger society. Variations in language behavior depend in part on the degree to which ties within these social networks are strong or weak (Tuc, 2003). Milroy and Li (1995) have used a "network

strength scale" to examine the link between network links and their relative strengths and language behavior patterns. Language behaviors such as code-switching and script choice (Androutspuolos, 2011) may be tied to the peculiarities of language norms for differing social groups, especially within social networks in digital social environments.

In his dissertation on code-switching among Australian Vietnamese speakers, Tuc (2003) found that the despite the distance between the three networks of Vietnamese speakers in Australia under study, they tended to share similar code-switching behaviors. He concluded that this "may be due to the likelihood that Vietnamese-English bilinguals in the three areas of Melbourne are involved in relatively similar types of interpersonal networks, and are in contact with one another" (p. 137).

It is expected for this paper that despite the distance between social groups under study for this dissertation, the closeness that social media networks provide, through daily interaction on blogs and the opportunities for interaction among otherwise isolated individuals, that groups that share characteristics would share similar language behaviors and patterns. For this study, this means that by identifying as Vietnamese, bloggers would be expected to be more likely to use Vietnamese in similar ways. Moreover, by identifying as Vietnamese in a particular location-Viet Nam, the US and Australia-bloggers might also use language in similar ways to their respective country counterparts.

CHAPTER 3

BLOG SAMPLING AND METHODS

Over 1.5 million persons living in the United States claim Vietnamese ethnicity alone (US Census Bureau, 2011). Adding the number of persons who claim some combination of Vietnamese ancestry, this number totals to over 1.6 million people. While this represents only 0.5% of the total American population, these 1.6 million persons are estimated to comprise roughly half of the total Vietnamese overseas diaspora (Ben-Moshe & Pyke, 2012).

A newer immigrant group, the largest number having immigrated to the US after 1975, most Vietnamese living in the US are first-, second- and third- generation American citizens and residents. Indeed, the US Census Bureau's American Community Survey (2010) finds that sixty-seven percent of persons claiming Vietnamese ancestry are foreign born.

Although the political situation has changed for the Vietnamese population in the United States specifically regarding the resumption of ties between the US and Viet Nam, enabling travel back and forth between the two countries, the fact of the two populations' separation by the space of an ocean remains. Within this context, that of being a relatively new immigrant group to the U.S., with renewed ties to and travel back and forth to Viet Nam itself, the Vietnamese community has retained a significant connection to the Vietnamese language. Furthermore, the invention of the internet and the increasing availability of internet access in the US and in Viet Nam have enabled

direct communication between Vietnamese language speakers and learners in both countries despite the geographic space between them. Despite pressures towards English monolingualism and despite the long gap in relations between the US and Viet Nam, according to the US Census Bureau (2010), as of 2010, 87.5% of Vietnamese households claim to speak a language other than English at home.

Conversely, English use in Vietnamese households is variable, as reported by the US Census Bureau's American Community Survey (2000). Of the 1.02 million persons claiming Vietnamese ethnicity alone in this survey, over seventy thousand claim to speak only English, while over six hundred thousand claim to speak English 'less than very well' (US Census Bureau, 2000). For those 1.11 million persons claiming both Vietnamese and another ethnicity, the number claiming to speak only English rises to over one hundred thousand, while over six hundred seventy thousand claim to speak English 'less than very well' (US Census Bureau, 2000). These statistics mean that roughly eight percent of the American Vietnamese community speaks only English and another sixty percent have some English speaking ability, even if it is not fluent.

In the past, language contact theory consisted primarily of contrastive analyses of differences between languages as their speakers come into contact with one another and within the minds of bilinguals (Weinreich, 1963). These theories have been updated to include concepts of space, which take into account both migration and the growing use of media that closes gaps between persons and communities and makes communication between distant groups possible (Collins, Slembrouk & Baynam; 2009, Tuc; 2003). To connect language use within Vietnamese-speaking communities and

language contact theory, this dissertation will use corpus-analytic methods to ask the question, ‘How does Vietnamese language use differ among contemporary Vietnamese language users in informal electronic media?’ Specifically, one question that accounts for space and the effects of contact asks how does language differ between speakers in Viet Nam and speakers in the wider Vietnamese speaking diaspora.

Language contact pressures in the US are intense and the statistics relating to notions of cultural and civic assimilation – defined by Vigdor (2008) as the tendencies for groups to reach statistical parity with the mainstream (or aggregate) on cultural and civic factors – are comparatively high for Vietnamese in the US. As part of the definition of assimilation, Vigdor notes statistics relating to English speaking, homeownership, and rates of naturalization for Vietnamese immigrants that are among the highest for all named groups, especially for a group with such a recent history in the United States.

Ho Dac Tuc (2003), in his dissertation on Vietnamese-English bilingualism and code-switching in Australia noted much the same effects. Migration as much as time is a significant factor in language change and language contact theory notes the pressure of being a minority language community within a language majority community as matter of fact variables towards language change. In this, it would seem that Vietnamese language in the Australian context is under the same sorts of pressures as Vietnamese in the US, especially as most migration occurred to both countries in the aftermath of war and a clamp down by the Vietnamese government on travel outside Viet Nam until opening the borders significantly in the early 1990s.

According to the Ben-Moshe and Pyke (2012), 2006 census numbers place the Vietnamese population in Australia as the 5th largest in the world outside of Viet Nam at

over one hundred fifty nine thousand persons, after the US, Cambodia, France and Taiwan. The Vietnamese represent 0.65% of the overall Australian population of around 22.9 million persons (Australian Bureau of Statistics, 2013). In comparison, the Vietnamese population in the US is over 1.5 million persons at 0.5% of the total US population.

Vietnamese is the seventh largest spoken language in Australia (Tuc, 2003) and over 90% of the Vietnamese population reports speaking Vietnamese well or very well (Australian Research Council, 2012). Conversely, only 56.5% of Vietnamese Australians report speaking English well or very well by 2006 census results (2012). Furthermore, as English is the dominant language, Vietnamese "is by no means free of English influence" (Tuc, 2003).

While contact with English is stronger and more immediate for Vietnamese speakers in the West, changes are also occurring to Vietnamese in Viet Nam as well. Internet conventions, the existence of English as a global and world wide web lingua franca, the Socialist Republic of Viet Nam's language policy of promoting the learning and use of English, and the rapid scientific and technological development accompanied by additions to the Vietnamese lexicon from other languages in Viet Nam, especially in cities and at universities, point to language contact phenomena occurring for Vietnamese speakers in country as well. This author expects that Vietnamese-language use between the two populations has begun to quantifiably differentiate, with much of the difference caused by contact phenomena with English in the US and rapid additions to the lexicon occurring in Viet Nam, changes which may or may not be transferring into Vietnamese in the US.

3.2 Corpus Linguistics Methods

At present, in the field of linguistics generally and within studies of the Vietnamese language more specifically, methodologies using corpus linguistics methods are comparatively new, despite the fact that corpus linguistics methodologies have been in use since before Firth's time. As discussed in Chapter 2, the question of meaningful units for the Vietnamese language is complicated and in need of more empirical investigation. Questions regarding units of meaning, word segmentation, whether Vietnamese is a monosyllabic language, and the nature of the word in Vietnamese have become the focus of much of the CL and NLP literature. This dissertation, however, will diverge from these questions as they are currently being addressed.

In Chapter 2 of Trust the Text, Sinclair (2004) makes a case for analysis based on "units of meaning". For Vietnamese, the concept of word is not a given, individual syllables may or may not equate to isolated units of meaning and morphemes themselves may consist of from one up to multiple syllables. In addition, the lexicon is still being developed in some domains and is reportedly going through a period of rapid growth. Furthermore, this dissertation uses blogs and blog comments as data, which are recognized to feature creativity in language use, to include ad-hoc neologisms, spelling alternations, borrowing from other languages and code-switching. To try to determine how the language is being used without predetermined categories beyond the syllable, this dissertation will provide a corpus analysis format featuring the general description of co-occurring language forms and topical characteristics of the entire corpus, with the syllable and subsequent iterations of syllables as the basic units of analysis.

This dissertation will examine informal varieties of Vietnamese as used on a series of intersecting social networks found on internet blogs. The language-in-use data contribute evidence of patterns attested to in studies that attribute change from this form an invaluable part of The fact that some of the studies on Vietnamese, even those attributing change over time, contain no language data but rely on self-report of use or general descriptions of language behavior makes the need for actual language in use data necessary. The language used in blogs is generally intended to for public consumption, more or less. The fact that it is collected in a situation of use means that it is an actual form of use data rather than examples of how language forms might be used. This study comprises one among few current studies in English of the Vietnamese language that examines actual language data that was produced for use and not specifically for research.

3.3 Data Collection

Corpus data were collected using a snowball sample starting with publicly accessible blog sources named on the website http://www.saigonbao.com/. The SaigonBao website offers sources by category, such as by dissemination source, e.g., radio, newspapers, blogs; by genre, e.g., religion and economics; and also by ideology, e.g., *Cộng Sản* 'communist' and *Dân Chủ* 'Democratic.[4]' These named divisions should not be understood as exhaustive; however, as for example, both Vietnamese and Western sources are listed under the *Tuổi Trẻ* 'Youth' category.

[4] These divisions in essence represent sources coming out of Viet Nam, termed Communist, and sources coming out of the West, usually termed (Democratic) .

Using the aforementioned snowball sample technique, blogs accessed initially through the SaigonBao site were then checked for references to other blogs. On most blogs, users have the option of posting links to blogs they themselves read, either in the profile section for some sites or as a list in a column featured alongside the main blog content column. Also, fellow bloggers and blog links could be found in the content of the blogs themselves or referenced in the comments section, especially when the referenced blogger had written about similar topics. As such, data in this corpus could not be considered a random sample of Vietnamese language data across in-country and English-speaking diaspora sources, but rather a representation of language used in a series of intersecting networks. Even if the data used for this dissertation does not represent a random sample, the use of blog data and the aforementioned snowball sampling technique afforded the best means to access this variety of the Vietnamese language as used by bloggers throughout the social networks posed in Viet Nam, the US and Australia, albeit with the limitations as discussed in below.

This study consists of an over six million-tiếng corpus, which includes three roughly 1.7-2-million syllable subcorpora from blogs originating in Viet Nam, the US and Australia. To compare, in the CL and NLP research reviewed in previous chapters, the smallest described corpus contained one million syllables (Pham, 2009) and the largest to date included 131, 318, 974 syllables (Hieu, Vu, & Kien, 2010). Details about the individual country corpora and the full blog and comment corpus appears below in Table 1. This table provides descriptive statistics on word counts and counts broken down by blog and comments entries themselves. The last column in the table gives a ratio of

comments to blogs. Table 2 appears in Appendix A and includes the detailed blog and comment counts per blog by country.

Table 1: Blog and Comment Totals

Country	Total Blogs	Total Words	Blog Words	Comment Words	Percent: Comments/Blogs
All	75	6,090,946	3,265,186	2,825,760	86.54%
Viet Nam	25	1,759,902	1,111,449	648,453	58.34%
Australia	25	2,293,832	974,873	1,318,959	135.30%
US	25	2,037,212	1,178,864	858,348	72.81%

Only publicly available blogs and blog content, accessible without passwords or signing into websites were included into the corpus. Blogs identified in links from prior blogs that led to sites normally requiring logins and passwords, such as at Yahoo.com blogging sites and groups, were not incorporated into the sample. Blogs were further identified for inclusion based on two primary criteria, 1) bloggers had to have identified themselves as Vietnamese and, 2) they had to have identified their living location as in Viet Nam, the US or Australia, either in the profile of their blog or in the content of blog entries.

Once particular blogs were identified for inclusion, secondary filtering retained only blog entries made in Vietnamese, even if these blog entries contained content in another language. Blog entries that incorporated articles from other sources into the body of the blog entry, no matter which language the articles were in, were kept for

inclusion. Articles appended to blog entries or linked to or referenced at the bottom of articles were not included in the corpus, no matter which language they were in. This was done to ensure that blog content especially would represent the particular social (Vietnamese) and geographic (Viet Nam, US, Australia) variety of the language as closely as possible.

Blogs and comments entries posted over a thirteen month period were included in the respective corpora, from December 2011 to December 2012. As may be noted from Table 2 in Appendix A, some bloggers and commenters are quite prolific and thus, some blogs are quite large and include a lot of comment content. Other bloggers write less long entries or do not blog as often and so other blogs are not as large and do not include as many entries. Also, for some blogs, no comments appeared, with some possible explanations to include a lack of popularity of the blog or the fact that comment functions were turned off for that particular blog.

Virtually all blog content written by bloggers who identified themselves as Vietnamese were written in standard Vietnamese orthography, using standard fonts available for Vietnamese. Blog comments in particular; however, featured a variety of writing styles, including the use of symbols to represent Vietnamese orthography. For example, the word *tiếng* 'syllable' was sometimes represented with keyboard symbols as *tie^'ng*. Other authors used a standard keyboard without using any special font program for Vietnamese, thus rendering no diacritical or tone symbols at all, as in *tieng*. While such non-standard orthography did not comprise a large part of the overall corpus, it is mentioned in analysis where features are salient.

For both blog data in particular (which here mean blogs and comments), the use of alternate forms in the semiotic environment may be indexical of other factors than simply meaning. The choice to use an unannotated corpus was also deliberate in order to preserve features that may relate to form rather than meaning, with form including features that may be visible or audible. For example, two tiếng form *só rì* 'sorry' is borrowed from English, but rendered using Vietnamese orthography. In seeing this word as used in context, it resembles any other Vietnamese word, but hearing it, it even adheres to English pitch, with the first tiếng rising – *só* and the second tiếng falling – *rì*. Lemmatizing this form into other meaning based head words such as *sorry* (English) or *xin lỗi* 'sorry' would preclude identification of creative forms.

In another example, the forms *hugs* and *hugsssss* both appeared, among others. Again, a decision was made not to fold these two forms into one entry in an attempt to preserve all forms, without prioritizing function over form. After all, the question arises as to whether the use of the additional 's' in the second form is indicative of a bigger hug or a more firm hug. Is there a pragmatic difference between the forms? What sociolinguistic variables might the two different forms represent? In these examples are also decisions not to lemmatize misspellings or correct for them as well as decisions not to correct for the lack of spaces between forms which would render two syllable forms into one form. Again, the inability to determine intent drove each of these decisions to work with a raw corpus.

Blog and comment entries were cut and pasted into separate files for each blogger and for comments to each bloggers' entries. Cleaning of the corpus consisted of removing all blog tags which appeared in the pasted blog content, such as time

stamps, reply links which allow commenters to upload comments, and any other metatags that did not feature as actual blog or comment content. As mentioned previously, no part-of-speech tagging or other form of predetermined unit tagging occurred beyond the exploitation of pre-existing spaces between syllables as division into units for analysis. In order to avoid top-down assumptions about syllable meaning, grammatical function and lexical and syntactic analysis and to deal with both cross and overlap ambiguities, which no program at present is able to accurately overcome for Vietnamese, no pre-tagging of the corpus was required.

An untagged corpus was also used in order to deal with a corpus created from English-speaking diaspora Vietnamese sources and thus may be expected to have some English language influence. Tagging and annotating parameters using specifications listed in some of the NLP research on Vietnamese overlook foreign words and not include them in the tagging and word-segmentation processes. Were there examples where a Vietnamese language phrase is created with an English language 'island' inside, these words might be missed or create confusion for processing software. For example, in a case there the syllable 'sự', used to derive nouns from adjectives and verbs, is used to modify an English-derived word, such as sự formal 'formality' or even sự formality 'formality' a naïve tagging program would be expected to ignore the word 'formal' or even the phrase 'sự formal/ity' entirely.

3.4 Data Analysis

Three major analyses were conducted, corresponding with the study's three research questions.

Question 1. What are the most common syllable forms and collocational, colligational, and topical patterns as revealed in a corpus of in-country and US and Australian Vietnamese language blogs?

The first analysis provides a general description of co-occurring language forms and topical characteristics of the entire corpus as a whole, with the syllable and subsequent iterations of up to 5-syllable units as the basic units of analysis. Analysis based on iterations of syllables and larger strings allowed a wider understanding of exactly how Vietnamese works in this sample of blog discourse, taking into account the problematic nature of meaningful unit segmentation in Vietnamese.

Syllable iterations were analyzed as 1-5 unit clusters as both collocational and colligational analyses for the aforementioned 1-, 2-, 3-, 4- and 5-syllable combinations. Extending the notion of a collocation as a non-verbal form, this analysis would include grammatical words, which are characterized by Firth as colligations (McEnery, Xiao & Tono, p.146). Collocational and colligational analyses were extended to up to five syllable clusters to assure maximal assessment of semantic and discursive properties. This seemed to make the most sense in hope of finding common forms used as well as neologisms within language varieties and those that mix the two and other language varieties in novel ways. This is also assumed to be the best method for analyzing blogs and other forms of informal language that have not been professionally edited to remove such novelties as described above. The human judgments involved in manual processing and analysis in this case are justified given that human judgments are already used in the literature as raters to check the validity of NLP coded and tagged corpora.

That said, for purposes of syllable and word verification and analysis, bound paper dictionaries, on-line dictionaries and corpus dictionaries available at the CORLING site were used. Despite the problems discussed previously regarding the incomplete and insufficient state of Vietnamese lexicography, these remained the best options for verification, and the CL and NLP research generally takes advantage of these media as well. Luckily, several newer CL and NLP study-derived dictionaries were available to supplement traditional bound and on-line dictionaries for this analysis. As in the CL and NLP research, independent raters, sometimes linguists themselves, are used to verify and assess accuracy for meaning and word segmentation boundaries.

In this sense the literature makes it quite clear that for the Vietnamese language, dictionaries are improving, but as yet insufficient for the exhaustive task of word segmentation and meaning confirmation. It should be noted, though, that as in the general corpus analysis literature, new words are continually being discovered for all languages and in all corpora, therefore it is expected that use of dictionaries are by their very nature incomplete and that native speakers should not be understood to be aware of all potential language forms, no matter how well educated or well-read these raters may be. In these cases, context and analogy may be the only means by which to determine meaning for any patterns found.

Question 2. What are the most common syllable forms and collocational, colligational and topical patterns as revealed in a comparison between in-country and US and Australian Vietnamese language blogs?

The second analysis used the previously conducted analysis methods for question 1, but divided the corpus by source in order to compare two subcorpora by area of origin, diaspora (US and Australian) and in-country (Viet Nam). Its purpose was to look for similarities and differences in Vietnamese and diaspora varieties of Vietnamese. In addition, the corpus was further divided into three parts, Viet Nam, US, and Australia and analyzed separately using the same methods.

Question 3. In corpus analysis of Vietnamese language varieties, what are the implications for how we analyze data? What are the implications for existing theory concerning segmentation into meaningful units in Vietnamese? How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?

Based on the previous analyses, the third analysis will look at issues of unit segmentation, lexical, grammar and discursive forms associated with the overall corpus, for the three subcorpora individually. The purpose of this third analysis is to examine the overall collocational and colligational processes and to compare and contrast these techniques and findings against the growing body of CL and NLP literature. Further analysis of the subcorpora individually would examine whether unit segmentation and collocational and colligational patterns diverge from each other as well as from the overall corpus and whether these analyses may be of use in analyzing patterns related to the notion of segmentation for Vietnamese varieties overall and for the named varieties individually.

CHAPTER 4

LISTS, COUNTS, CURVES

To depart from previous studies, which have sought to predetermine units of meaning and parts of speech and morpho-syntactic operations for Vietnamese, this study will use a corpus-driven approach to study of a corpus of Vietnamese language blogs. The use of an inductive approach, without pretagging of the corpus has been chosen, to avoid deductive pre-analysis of the corpus through any form of annotation and editing, apart from the aforementioned use of spaces between units of analysis, here referred to as tiếng, as a 'naturally' occurring feature of word processing and corpus analysis software.

As such, this analysis will attempt to derive meaning from recurring contexts of use, in the model of Firth, where meaning is lexis driven and the divide between lexis and grammar is not a given. Conversely; however, this analysis will not seek to create an entirely new set of categorizations of the language and its features, but rather explore the progression of units and how these may be classified based on preexisting categories, such as open-class and closed-class, etc. In this way, this dissertation explores the realities as they occur in this sample of Vietnamese blogs, recognizing as in Saussure and Kretzschmar, that selection and classification of units is a necessarily arbitrary process (Saussure; 1983, Kretzschmar; 2009).

This paper describes one corpus of Vietnamese language data derived from a series of intersecting networks of internet blogs originating in Viet Nam, the US and

Australia. For additional analysis, this corpus is divided into two subcorpora, the blog entries themselves and blog comments.

This is a study of a variety of the Vietnamese language using a corpus of language in use data that include Vietnamese from in-country and diaspora sources. Although this study will be looking only at a sample of the full range of language data possible, and only through the lens of a small available internet sample of that language, this method will provide a look at one such informal intersecting network of internet language logs as to how the language as used in the informal context of blogs. This paper will address these gaps with the following research question:

What are the most common syllable forms and collocational, colligational, and topical patterns as revealed in a corpus of in-country and US and Australian Vietnamese language blogs?

4.2 Methods

The first analysis provides a general description of co-occurring language forms and topical characteristics of the entire corpus as a whole, with the tiếng and subsequent iterations of up to 5-tiếng units as the basic units of analysis. Analysis based on iterations of syllables allowed a wider understanding of exactly how Vietnamese works in this sample of blog discourse, taking into account the problematic nature of meaningful unit segmentation in Vietnamese.

Syllable iterations were analyzed as 1-5 unit clusters as both collocational and colligational analyses for the aforementioned 1-, 2-, 3-, 4- and 5-tiếng combinations. Extending the notion of a collocation as a non-verbal form, this analysis would include

grammatical words, which are characterized by Firth as colligations (McEnery, Xiao & Tono, 2006, p.146). Collocational and colligational analyses were extended to up to five tiếng clusters to assure maximal assessment of semantic and discursive properties. This seemed to make the most sense in hope of finding common forms used as well as any other forms found within language varieties and even those that mix the two and other language varieties in novel ways. This is also assumed to be the best method for analyzing blogs and other forms of informal language that have not been professionally edited to remove such novelties as described above. The human judgments involved in manual processing and analysis in this case are justified given that human judgments are already used in the literature as raters to check the validity of NLP coded and tagged corpora.

That said, for purposes of tiếng verification and analysis, bound paper dictionaries, on-line dictionaries, and corpus dictionaries available at the online were used. Despite the problems discussed previously regarding the incomplete and insufficient state of Vietnamese lexicography, these remained the best options for verification, and the corpus linguistics (CL) and natural language processing (NLP) research generally takes advantage of these media as well. Luckily, several newer CL and NLP study-derived dictionaries were available to supplement traditional bound and on-line dictionaries for this analysis. As in the CL and NLP research, independent raters, sometimes linguists themselves, are used to verify and assess accuracy for meaning and word segmentation boundaries.

The literature makes it quite clear that for the Vietnamese language dictionaries are improving but, as yet insufficient for the exhaustive task of word segmentation and

meaning confirmation. It should be noted, though, that as in the general corpus analysis literature, new words are continually being discovered for all languages and in all corpora, therefore it is expected that use of dictionaries are by their very nature incomplete and that native speakers should not be understood to be aware of all potential language forms, no matter how well educated or well-read these raters may be. In these cases, context and analogy may be the only means by which to determine meaning for any patterns found.

4.3 Description of Corpora

Table 1 below describes some of the basic features of each of the three corpora, full corpus, blog corpus and comment corpus. *Types* refers to the number of unique forms that appear in each list. *Tokens* refers to the total number of times those unique forms appear, or the total number of forms for each list. *Top 20 Types % of Tokens* is calculated based on the sum of the frequency of occurrence for each of the top twenty types as a percentage of the entire token count for that list. *Top 100 Types % of Tokens* is calculated in the same manner as for the Top 20 list, but gives the percentage for the top one hundred forms for the respective lists.

The *Top 25% of Tokens* and *Top 80% of Tokens* columns give the number of types for which 25% and 80% of word usage, or sum of tokens, and the percentage of total types represented by that percentage of tokens. Finally, the *1-Freq Tail column* provides data showing the tail of the distributions where forms with a frequency of one occur. The sub-columns give the number of types with a frequency of one and the percentage of those types among all types for each list.

Lists of the top one hundred 1-, 2-, 3-, 4-, and 5-tiếng forms for each of the corpora; full corpus, blog corpus and comment corpus, may be found in Appendices B, D, and F. In addition, charts showing A-curve (rank and frequency) analyses for each of the form lists may be found in Appendices C, E, and G. While the lists presented in Appendices B, D, and F feature the top one hundred forms for each of the 1- through 5-tiếng data sets and corpora, the use of the A-curve as in Appendices C, E, and G shows the distribution of types and tokens based on the first 3000 forms in each list.

The full corpus, referred to within Table 1 as All BC (or All Blogs and Comments), has a total of 39,975 unique words, or types. If we compare that number to the number of types for the blogs (28,340) and comments (26,401) subcorpora, we see that while the two subcorpora may share some types, there are quite a number of types they do not share. The number of types for the full corpus is 11,635 more than the number of types for the blogs corpus and 13,574 more than the number of types for the comments corpus.

For the lists representing the 2-, 3-, 4-, and 5- tiếng forms, we see a similar pattern. The number for the full corpus is larger than the numbers for the subcorpora, indicating unique forms that appear in one, but not both subcorpora. Please note that these lists are not simple node forms and their collocates, these lists are derived in Wordsmith Tools, version 5.0 (Scott, 2008), as forms that appear in varying frequencies as clusters of 2-, 3-, 4-, and 5- tiếng forms. As with any analysis of word clusters, these lists include iterations of forms that appear in sequence within discourse that may not have any lexical or grammatical meaning, but rather appear only because discourse was broken up in sequential order to derive the required lists.

Table 1: Full Corpus, Blog Corpus, and Comment Corpus Descriptors

			Top 20 Types	Top 100 Types	Top 25% of Tokens		Top 80% of Tokens		1-Freq Tail	
Forms	**Types**	**Tokens**	**% of Tokens**	**% of Tokens**	**# of Types**	**% of Types**	**# of Types**	**% of Types**	**# of Types**	**% of Types**
All BC										
1-Tieng	39,975	6,105,756	15.61%	38.46%	43	0.11%	902	2.26%	16,728	41.85%
2-Tieng	1,089,252	5,678,175	1.52%	4.44%	3,019	0.28%	217,623	19.98%	657,192	60.33%
3-Tieng	3,340,382	5,289,209	0.25%	0.75%	108,952	3.26%	2,282,444	68.33%	2,726,534	81.62%
4-Tieng	4,328,359	4,936,162	0.09%	0.22%	582,611	13.46%	3,297,500	76.18%	4,001,651	92.45%
5-Tieng	4,366,059	4,607,359	0.03%	0.10%	910,540	20.85%	3,444,588	74.76%	4,196,626	96.12%
Blogs										
1-Tieng	28,340	3,279,202	16.73%	39.78%	39	0.14%	823	2.90%	10,949	38.63%
2-Tieng	670,296	3,049,900	1.76%	4.87%	2,504	0.37%	159,486	23.79%	413,354	61.67%
3-Tieng	1,896,963	2,838,072	0.31%	0.89%	77,260	4.07%	1,329,349	70.08%	1,579,533	83.27%
4-Tieng	2,376,966	2,643,989	0.12%	0.27%	393,975	16.57%	1,848,169	77.75%	2,229,247	93.79%
5-Tieng	2,373,662	2,462,773	0.05%	0.12%	526,583	22.18%	1,881,108	79.25%	2,310,101	97.32%
Comments										
1-Tieng	26,401	2,826,554	15.29%	34.74%	46	0.17%	911	3.45%	11,090	42.01%
2-Tieng	662,764	2,628,275	1.50%	4.80%	3,100	0.47%	193,418	29.08%	413,038	62.32%
3-Tieng	1,721,754	2,451,137	0.33%	0.85%	90,444	5.25%	1,231,517	71.53%	1,445,576	83.96%
4-Tieng	2,052,872	2,292,173	0.11%	0.30%	333,743	16.26%	1,594,438	77.67%	1,911,828	93.13%
5-Tieng	2,026,482	2,144,586	0.05%	0.15%	418,043	20.63%	1,597,565	78.83%	1,941,627	95.81%

For example, the sentence ‘I am a cat,’ will provide the following 2- word forms; I am, am a, a cat. However, when considering the Vietnamese language, although it is written monosyllabically, units of meaning may be from one up to five tiếng. For example, with the sentence ‘*Tôi là người Việt Nam* ‘I am Vietnamese’, a 2-unit cluster list will provide the following forms; *Tôi là, là người, người Việt, Việt Nam* ‘I am’, ‘am person’ ‘person Viet’, Viet Nam’. The unit Việt Nam, although written monosyllabically as two tiếng, represents one concept, the country of Viet Nam. A deeper discussion of the content of these lists will be featured in Chapter 5.

The Top 20 Types column shows that the top 20 types for the 1- tiếng forms for each of the corpora equals roughly 15% of tokens, with the blogs corpus slightly higher than the comments corpus. The Top 100 Types column for these same forms contains more that 38% of all tokens, with the blogs corpus somewhat higher than the comments corpus, for 39.78% and 34.74% respectively, indicating that unique forms comprise more of the most common forms for the blogs corpus than for the comments corpus. In layperson’s terms, what this means is that for the 1- tiếng lists, 15% of the word usage is being done by only 20 types and 39.78% of the word usage is being done by 100 types. This issue is complicated when one considers that 1- tiếng forms may be units of meaning in themselves, but also part of larger units of meaning.

As we drill down to the cluster forms lists for Top 25 and 100 Types; however, we see these numbers dwindle rapidly, indicating that while 1- tiếng forms represent a larger number of the more frequent tokens, the subsequent iterations do not appear as frequently for any of the corpora. For the 2- tiếng forms we see 1.52% and 4.44% for the full corpus, 1.76% and 4.87% for the blogs corpus and 1.50% and 4.80% for the

comments corpus. Percentages for the 3-, 4-, and 5- tiếng corpora are even smaller. This indicates that there are an increasingly higher percentage of lower frequency forms for each of these lists. As such, the results show that for the larger form-clusters, the top forms represent less of the word usage as for the 1- and 2- tiếng forms. In other words, the bigger the cluster, the less talking it does in general.

For the Top 25% and Top 80% of Tokens columns, we see that twenty-five percent of the sum of tokens represents 43, 39 and 46 types for each of the corpora, respectively. These numbers rise in opposition to the pattern for the Top 20 and Top 100 Types columns. This indicates that the top types for each of the lists has a lower frequency as iterations of tiếng increases.

Supporting the assertion of an increasingly lower percentage of high frequency forms made in the discussion of Top Types is supported by looking at the 1-Freq Tail columns. These columns show the number of types and percentage of total types represented by types with a frequency of one. For the 4- and 5- tiếng lists for each of the corpora, these numbers are above 90%. As this corpus, and its two subcorpora, represent an unedited, unannotated corpus, this is unsurprising. Including the sequential iterations of forms as discussed in the 'I am a cat' example above, these lists also feature a variety of unique forms regardless the size of the cluster. These forms should not be understood to be unique based on rareness of use, but rather a reflection of the variety of alternative orthographies, misspellings, dialect-based spellings, pronunciation-based spellings, foreign forms, and other content discussed in Chapter 3.

4.4 Tiếng Lists and Curves

As described in Kretzschmar (2009), the A-curve showing “frequency of frequencies” (p. 158) represents the inverse relationship between the rank of a type and its frequency or the number of tokens for that given type. The A-Curve as given in Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. For analyses of many languages of the world, the above A-curve shows that inverse relationship between rank and frequency, wherein rank is the number of times a particular type appears – the number of tokens of that particular type, or its frequency. Kretzschmar states that Zipf’s law “extends to experimental data from survey research as well as to words in texts, and thus it stands as a primary characteristic of speech as a complex system” (2009, p. 159). For this corpus, analysis shows that for this sample of the Vietnamese language, the distribution of forms according to the A-curve holds not only for 1- tiếng lists, but also for the 2-, 3-, 4- and 5- tiếng lists.

The charts in Appendix C show the A-curve for the full corpus, Appendix E for the blogs corpus and Appendix G for the comments corpus. As discussed in the above analysis of Table 1, we can see that for each list, the inverse relationship between rank and frequency obtains. The y-axis represents the number of tokens or the frequency for types. The x-axis represents the rank of the form based on its frequency. For this analysis, rank could also stand in for the actual named form, but for purposes of this study, numbers are shown in order to demonstrate the rapidity with which rank declines based on frequency.

The following series of charts represents the full corpus:

Chart 1a: Full Corpus 1-Tiếng Chart

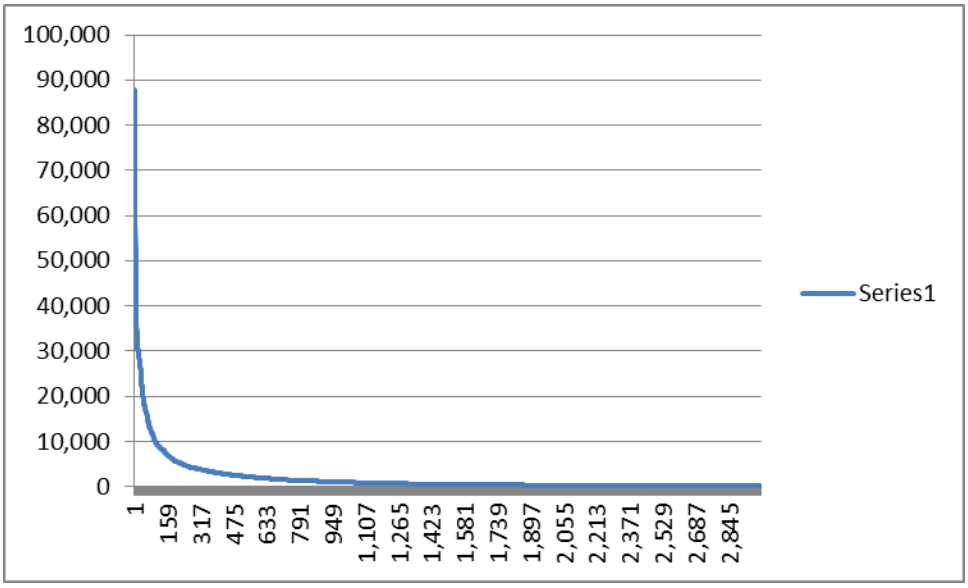


Chart 1b: Full Corpus 2-Tiếng Chart

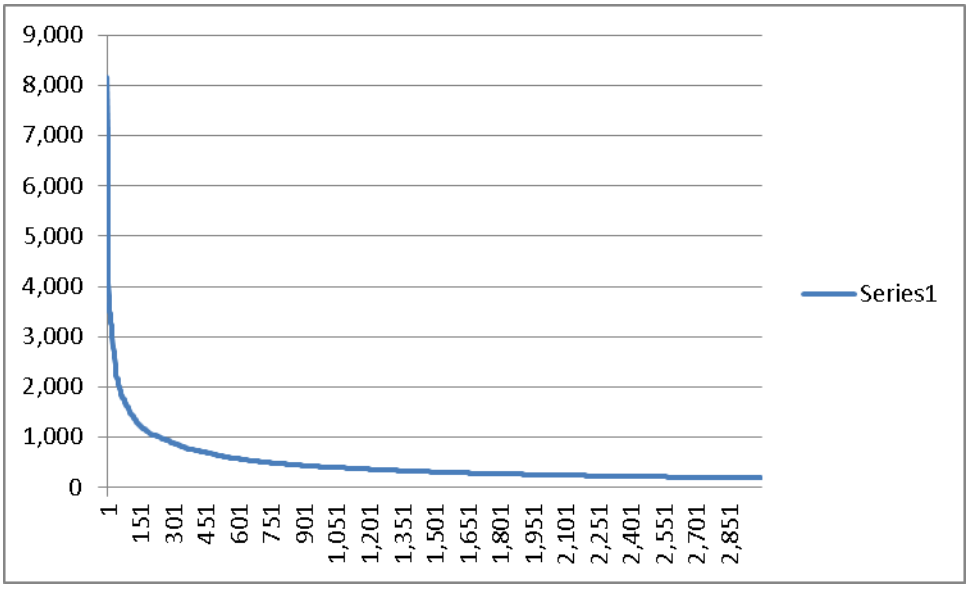


Chart 1c: Full Corpus 3-Tiếng Chart

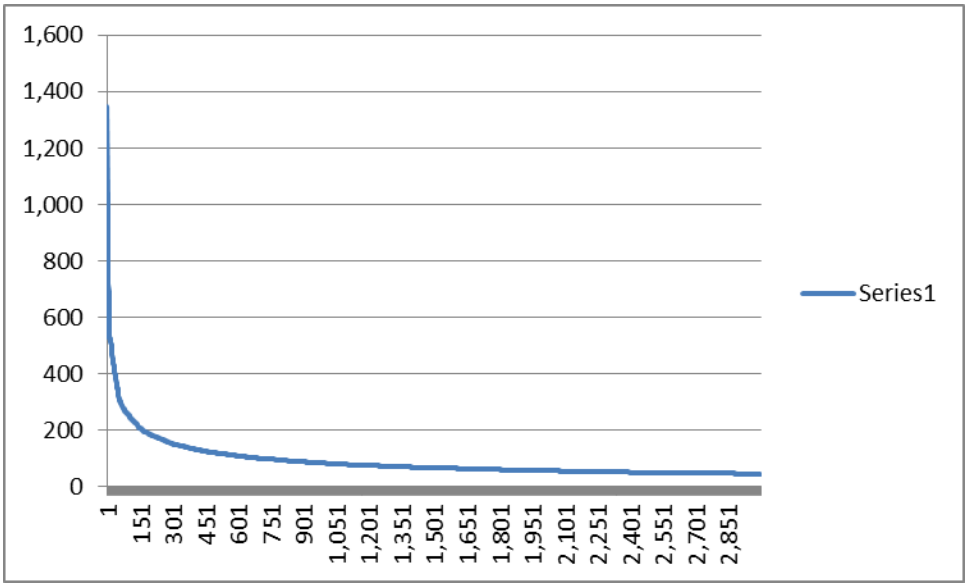


Chart 1d: Full Corpus 4-Tiếng Chart

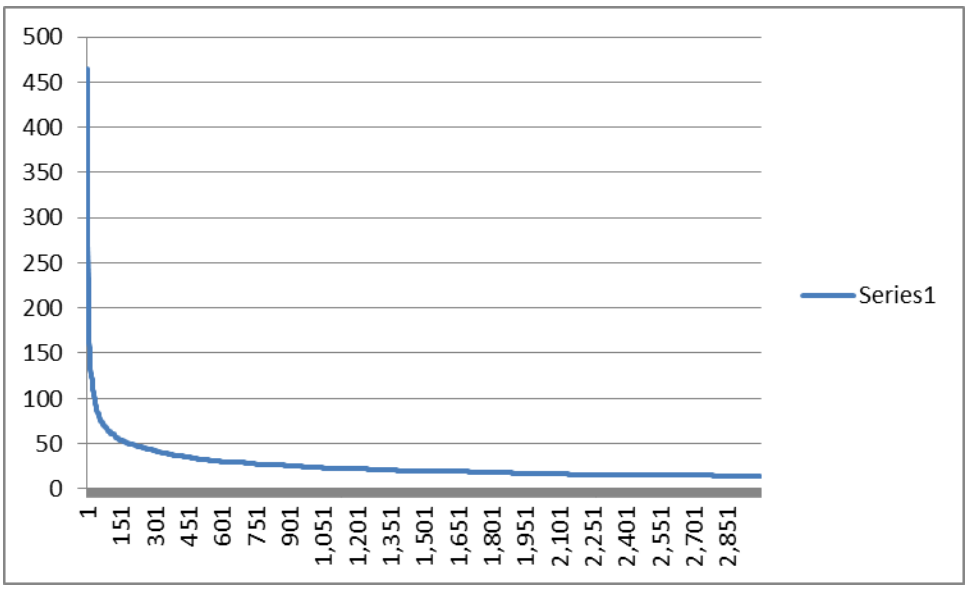
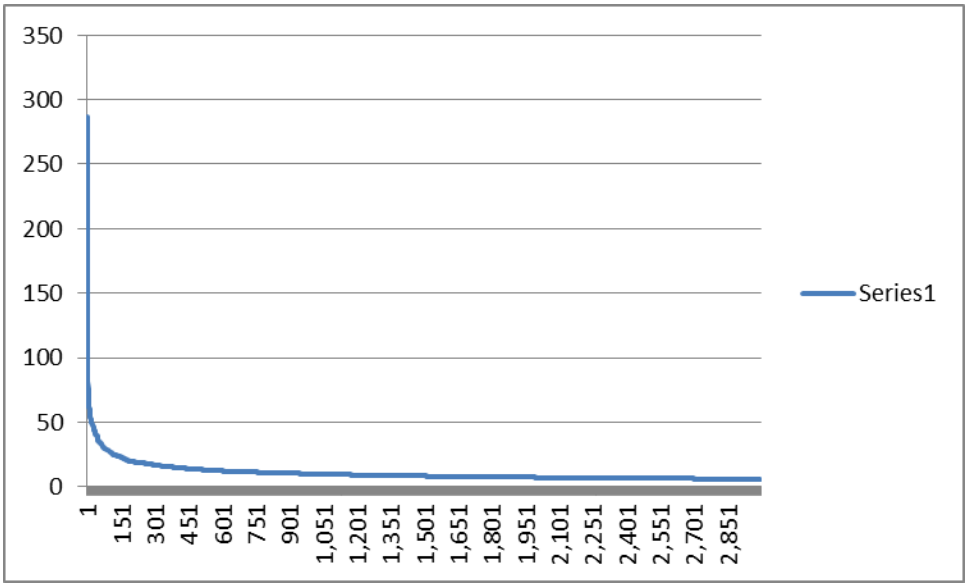


Chart 1e: Full Corpus 5-Tiếng Chart



As shown, the a-curve begins on the y-axis with the frequency for the top ranked type. Actual frequencies for the top 100 types for all corpora are also featured on the lists in Appendices B, D, and F. For Chart 1a, we see that the top type *LÀ* has 87,789 tokens. As discussed on section 4.3, we further see that for the subsequent lists, top types do not appear as commonly for the 2-, 3-, 4-, and 5- tiếng forms, as the top types have a much lower number of tokens. Also, we can see how rapidly the tokens decline in frequency, with a much less sharply declining curve for the 1- tiếng list compared to the 2- tiếng list. Which then, declines much less sharply than the 3-tiếng list and so on.

The following series of charts represents the blogs corpus:

Chart 2a: Blog Corpus 1-Tiếng Chart

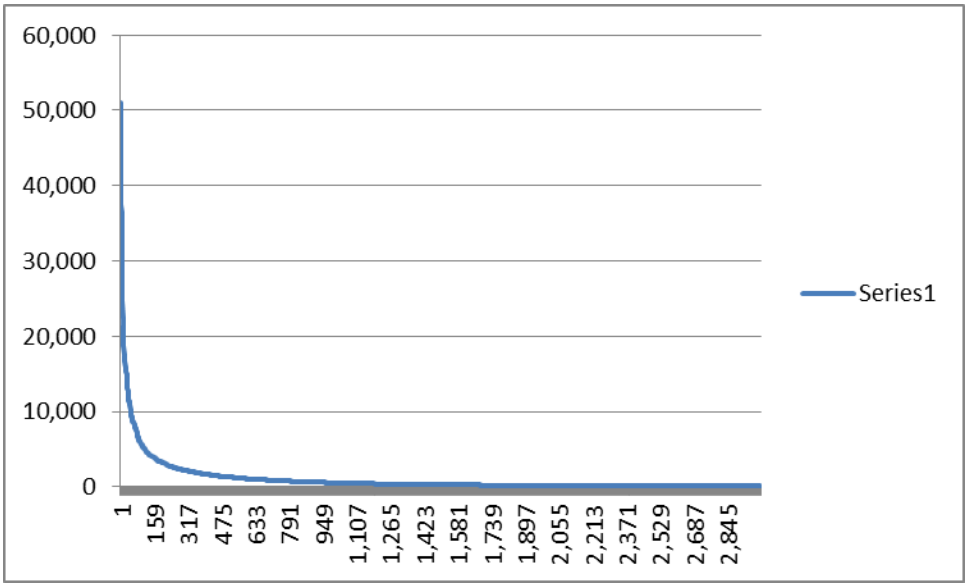


Chart 2b: Blog Corpus 2-Tiếng Chart

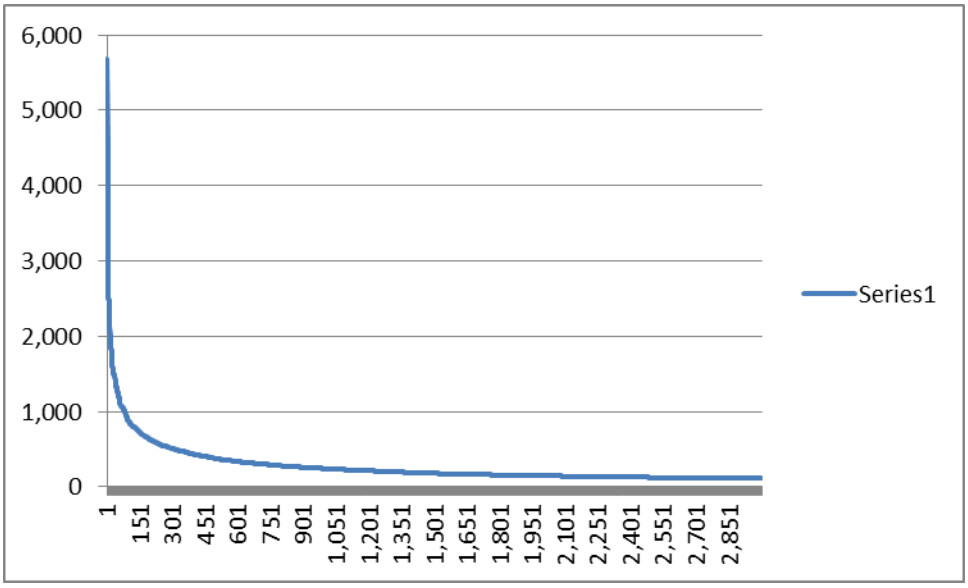


Chart 2c: Blog Corpus 3-Tiếng Chart

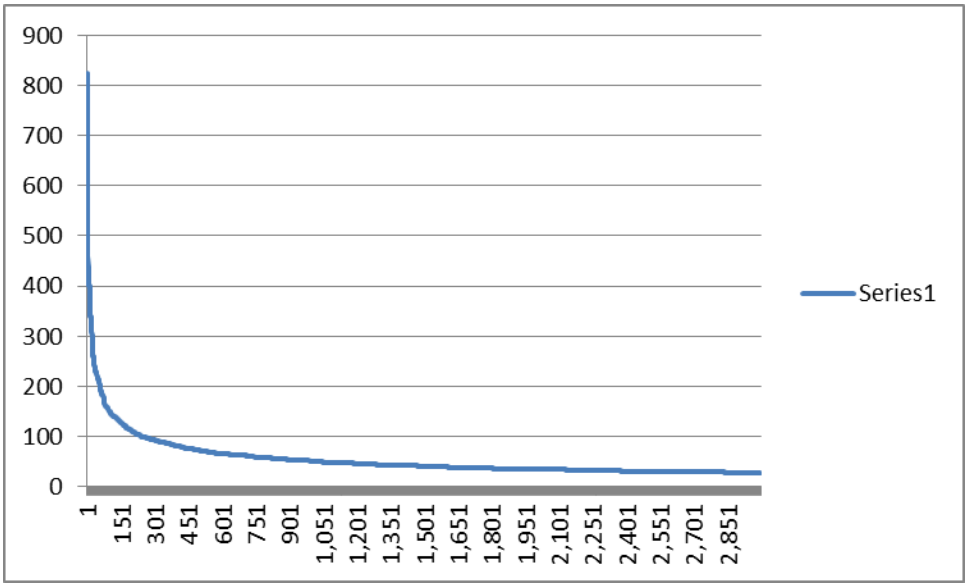


Chart 2d: Blog Corpus 4-Tiếng Chart

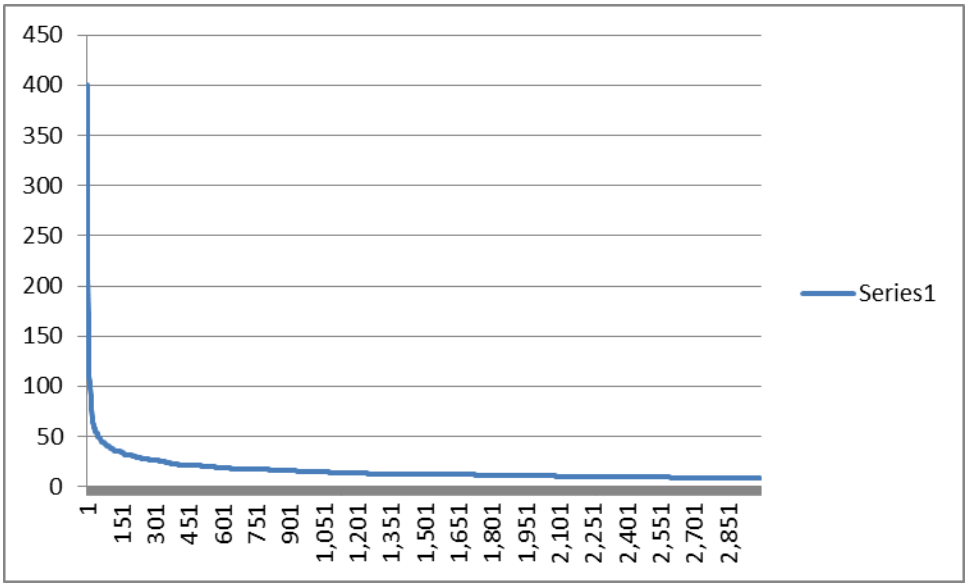
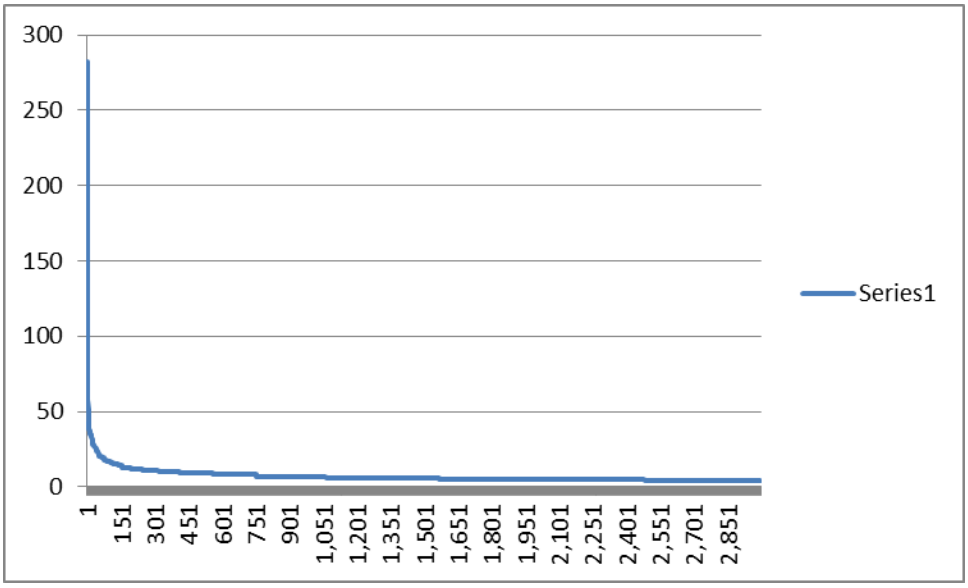


Chart 2e: Blog Corpus 5-Tiếng Chart



Here, we see the same pattern as for the full corpus lists. The a-curve begins on the y-axis with the frequency for the top ranked type. For Chart 2a, we see that the top type #, representing the use of digits in the corpus, has 51, 019 tokens. As happened for the full corpus, we further see that for the subsequent lists, top types do not appear as commonly for the 2-, 3-, 4-, and 5- tiếng forms, as the top types have a much lower number of tokens. Also, we can see how rapidly the tokens decline in frequency, with a much less sharply declining curve for the 1- tiếng list compared to the 2- tiếng list. Which then, declines much less sharply than the 3-tiếng list and so on.

The following series of charts represents the comments corpus:

Chart 3a: Comment Corpus 1-Tiếng Chart

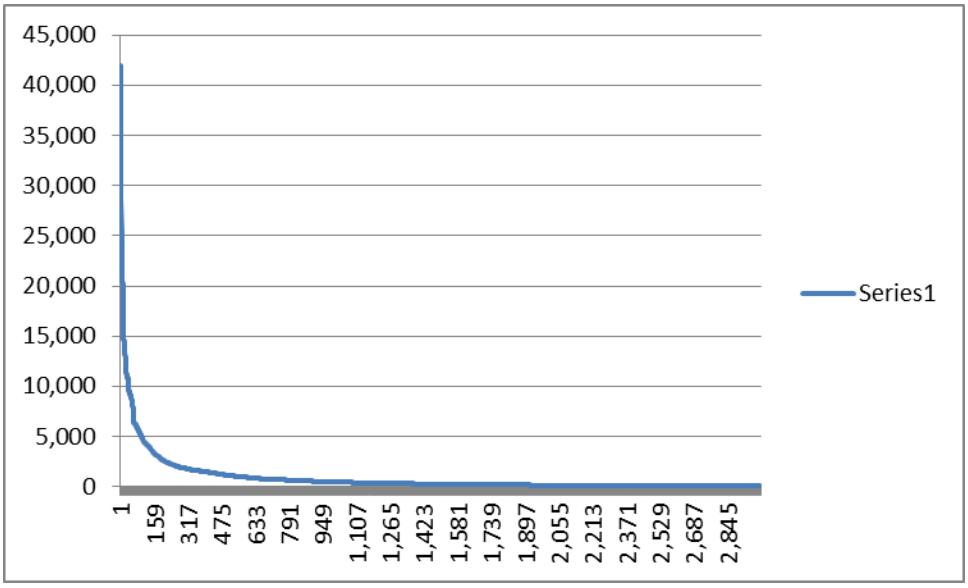


Chart 3b: Comment Corpus 2-Tiếng Chart

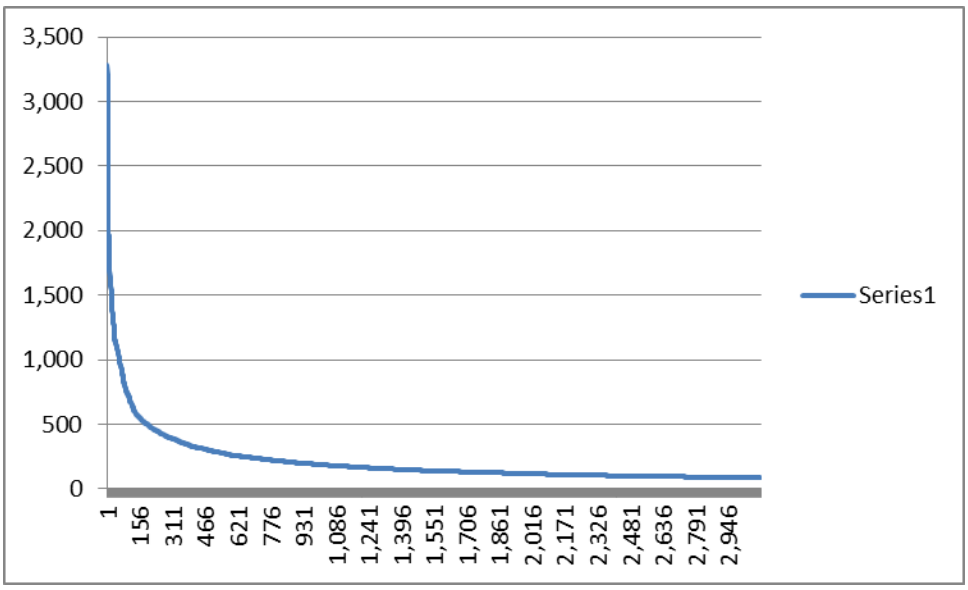


Chart 3c: Comment Corpus 3-Tiếng Chart

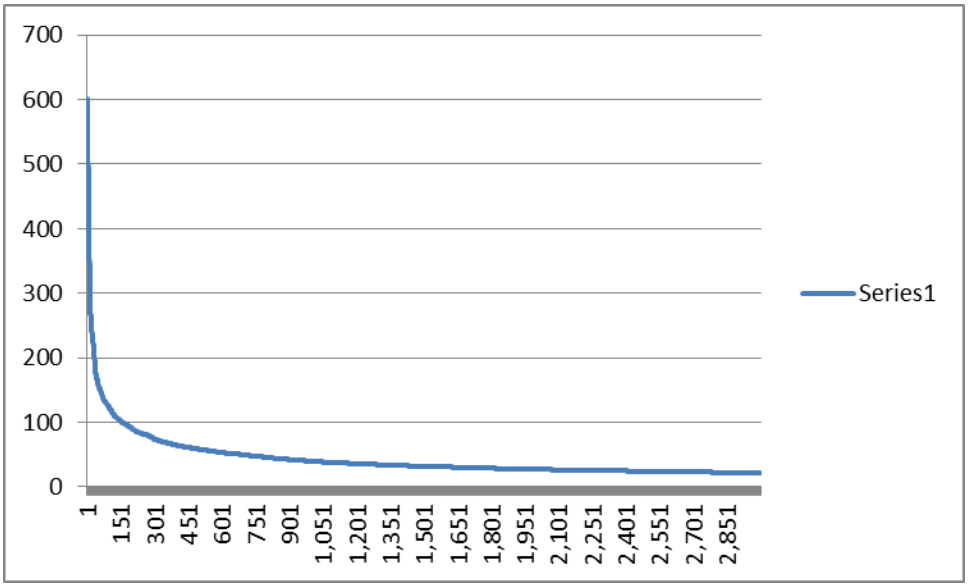


Chart 3d: Comment Corpus 4-Tiếng Chart

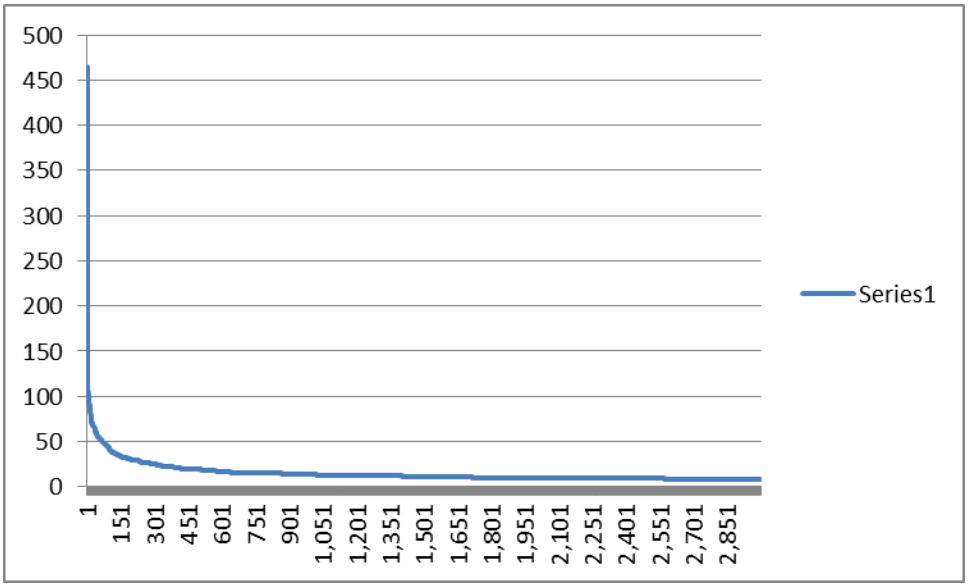
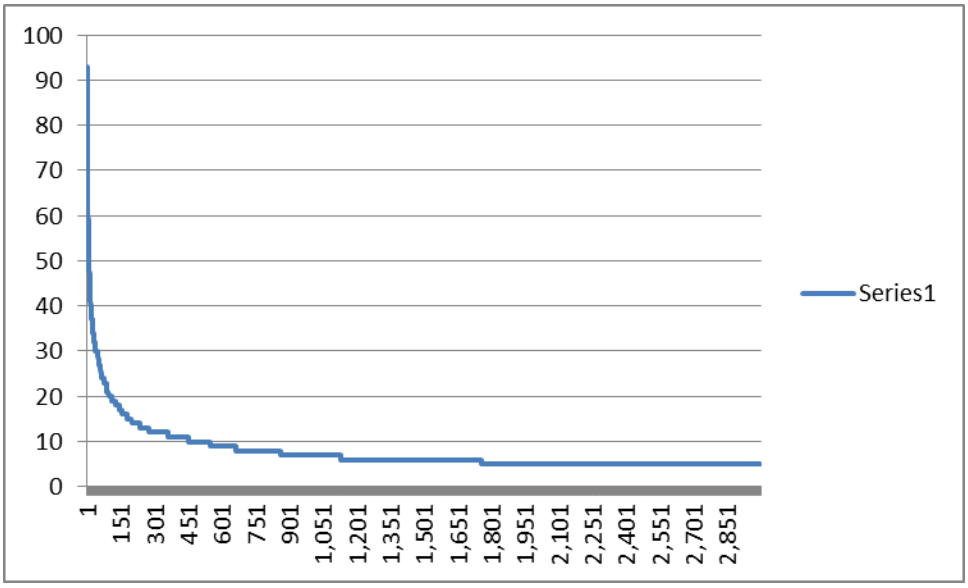


Chart 3e: Comment Corpus 5-Tiếng Chart



And, again, the same pattern persists with the comments corpus as for the full and blogs corpora. The same A-curve pattern showing the inverse relationship between rank and frequency with the 1- tiếng list showing the highest initial token frequency and the least sharp decline down to the 5- tiếng list showing the lowest initial token frequency and the sharpest decline in frequency for subsequent types.

4.5 Conclusion

This analysis provided a description of the full corpus for this study as well as for the two subcorpora, blogs and comments. Results indicate that while much of the content between the two subcorpora is shared, much is not, as revealed by the total number of types for the full corpus, which is over 11,000 more than for either of the two subcorpora just for 1- tiếng types alone. These results also obtain for the subsequent iterations of

2-, 3-, 4-, and 5- tiếng lists, with much shared content, but also much unique content between the two subcorpora.

Results also indicate the much higher frequency rate for unique forms for the 1-tiếng lists, with lower initial token rates and more sharply declining frequency rates for the subsequent form iterations. The prevalence of 1-frequency forms is highest as the number of tiếng in the forms increases, with the 4- and 5- tiếng forms lists being comprised of over 90% 1-frequency forms. This indicates that 1- and 2- tiếng are relatively more commonly used and represent more typical language patterns than 3-, 4- and 5- tiếng forms.

A-curve charts further indicate the ubiquity of the inverse relationship between rank and frequency of forms, no matter the size of the tiếng cluster, supporting assertions of this sample of Vietnamese blogs as representing “speech as a complex system” (Kretzschmar, 2009, p. 159).

CHAPTER 5

CLUSTERS, FORMS, COUNTS

This chapter continues the analysis begun in Chapter 4, specifically regarding the lists of forms provided in Appendices B, D and F. In this analysis, the cluster lists appearing in the respective full and subcorpus lists from one to five-tiếng will be examined. As stated in the previous chapter, this analysis will not seek to create an entirely new set of categorizations of the language and its features, but rather explore the progression of units and how these may be classified based on preexisting categories. The first analysis begins with traditional structural categories, starting at the morpheme level and extending through word, phrase, clause and onto the sentence level. The second analysis explores the clusters from another perspective, labeling the relevant clusters in three ways, content/function, free/bound, and open/closed classes. Although there are many ways linguistic data may be classified, any choice of criteria should be considered as necessarily arbitrary, as one of many possible ways of exploring the linguistic realities at hand (de Saussure 1983, Kretzschmar 2009). Because such analysis is inherently arbitrary and furthermore, because notions of a perfect divide between lexis and grammar, collocation and colligation are not a given (Nguyen, et.al. 2006, Tognini-Bonelli 2001), interrater reliability analysis was also used in order to provide support for any such form classifications provided.

This paper will add onto the description of the full corpus of Vietnamese language data derived from a series of intersecting networks of internet blogs

originating in Viet Nam, the US and Australia. For additional analysis, this corpus was divided into two subcorpora, the blog entries themselves and blog comments.

This chapter will go deeper into the following research questions:

What are the most common syllable forms and collocational, colligational, and topical patterns as revealed in a corpus of in-country and US and Australian Vietnamese language blogs? In corpus analysis of Vietnamese language varieties, what are the implications for how we analyze data? What are the implications for existing theory concerning segmentation into meaningful units in Vietnamese? How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?

5.2 Methods

The chapter provides a structural, content, and meaningful units description of co-occurring language forms and topical characteristics of the entire corpus as a whole, and as divided into the two subcorpora, blogs and comments with the tiếng and subsequent iterations of up to 5-tiếng units as the basic units of analysis. Analysis based on iterations of syllables allowed a wider understanding of exactly how Vietnamese works in this sample of blog discourse, taking into account the problematic nature of meaningful unit segmentation in Vietnamese.

Syllable iterations were analyzed as 1-5 unit clusters as both collocational and colligational analyses for the aforementioned 1-, 2-, 3-, 4-, and 5-tiếng combinations. Extending the notion of a collocation as a non-verbal form, this analysis would include

grammatical words, which are characterized by Firth as colligations as a special case of collocation (McEnery, Xiao & Tono 2006: p.146). Because this paper prioritizes "linguistic realities" (Kretzschmar 2009), no pretagging or exclusion of forms based on any pre-existing criteria such as lexical word or part of speech (noun, verb, etc.) was used in order to understand exactly how the respective clusters of forms were distributed.

For this chapter, the three word lists as given in Chapter 4 Appendices B, D and F, were analyzed along three major dimensions. The first dimension is an analysis of structural forms, including the labels Morpheme, Word, Phrase, Clause, and Sentence. For the first two categorizations, Morpheme and Word, the forms were analyzed and a count was made of how many morphemes and words were contained in the form. For the remaining categorizations, Phrase, Clause and Sentence, the forms were analyzed and any forms containing a phrase, a clause, or a sentence were labeled as such.

The second dimension comprised an analysis of the types of forms along the axes Content and Function, Free and Bound, and Open and Closed. Although the forms in the list contain larger units than mere morphemes or words, an attempt was made to categorize each full form according to whether the content within could be defined as a 'reality' that conformed to the label as such. In some cases, forms were included in both columns of the content could arguably be considered to reflect aspects of one or both sides. For example, a 5-tiếng form might be labeled both a content and a function form if it contained primarily grammatical forms with few meaningful content forms, such as for the form *TẤT CẢ MỌI NGƯỜI ĐỀU* 'every person'.

The third dimension comprised an analysis of the forms based on Sinclair's (2004) concept of units of meaning. Although Sinclair discusses varying definitions of units of meaning from the morpheme to the word to the phrasal verb and beyond, for this analysis, units occurring above the level of the morpheme were categorized based on whether they, as single tiếng or any combination of tiếng within a form, could be considered as one unit. Then, total units of meaning were counted for each one- through 5-tiếng form.

An additional analysis was performed to assess how these units of meaning distributed themselves within the forms. Units of meaning (UM) is a necessarily vague but, yet central topic for this analysis, and so this analysis was done in addition to the more traditional analyses of forms. As collocation and colligation are both topics of interest in this study, the units of meaning category reflects differing notions of unit for both the collocational and colligational forms. For content UM, the notion of morpheme, word, and larger units contrast with function UM, where the notion of grammar and unique colligational structures for the Vietnamese language necessitated somewhat different conclusions. Considering individual tiếng for a colligational UM would not equate to the single meaning that the tiếng acting together obtain, much as single morphemes are not always words.

Using the example above, *TẤT CẢ MỌI NGƯỜI ĐỀU* 'every person' is described as having one unit of meaning with a structure of 5. The presence of the word *NGƯỜI* 'person,' in this form indicates content, however, the surrounding tiếng work together to create the meaning of 'all people' or 'every person.' The 1-tiếng form *ĐỀU* 'every, in all cases' acts as a distributive referring back to the topic to reinforce or indicate all of the

topic content as included. The form *ĐỀU* also marks the division between the topic and comment in Vietnamese sentence structure, a sentence structure type that stands alongside SVO order as typical for Vietnamese, especially spoken Vietnamese (Nguyen, et.al. 2006, Nguyen 2010). Despite the fact that this form would be considered a phrase in the structural analysis, for collocational UM analysis, each of the tiếng here act together to create one UM.

Because categorization was not a given for these analyses, interrater reliability analysis was also used. Before rating, both raters discussed their own understanding of the labels with the other. The first rater, the author of this dissertation, is a linguistics graduate student who has worked with Vietnamese language as a translator, university-level teacher, and linguist for over twenty-eight years. Other than one and a half years of formal Vietnamese language education, however, the first rater has no other formal education in Vietnamese. The second rater is a native speaker of Vietnamese, having attended Vietnamese schools through the first year of graduate school in linguistics. The second rater currently works as a teacher of the Vietnamese language to university undergraduate students. The full range of interrater reliability (IRR) analysis charts appears in Appendix H.

After the raters completed their analyses, they again discussed issues with analyzing the data as it was. As both raters live in the US, have received graduate language and linguistics education in English and work through the medium of English, both noted that it was hard to separate labels of Vietnamese items from education in linguistics that prioritized English ways of understanding language, theory, and research. For example, the form *XÃ HỘI CHỦ NGHĨA* ‘socialism’ is rendered in English

as one word. But, these items may also be defined as ‘the doctrine of socialism’. And, of course, the question remains whether these items should be considered as one word for Vietnamese itself, without taking English into consideration.

For IRR analysis, SPSS version 21 was used. Count variables, Morpheme and Word, were converted to categorical variables by labeling counts 1 through 6 as discrete categories. This reflects a way of understanding of the respective 1- through 5-tiếng forms as realities having certain properties, such as that a 1-tiếng form may also be a 1-morpheme form or a 1-word form. As variables Phrase, Clause, and Sentence were already analyzed as dichotomous, with 1 indicating the presence of a phrase, clause, or sentence, no conversion was necessary.

Names of bloggers and names of persons mentioned in blogs have been redacted from the data for this chapter in order to protect privacy. Public figures, such as Aung San Suu Kyi and Vietnamese military generals mentioned in regular news sources were not redacted.

5.3 Description 1

Table 1, below, describes the relevant one- through five-tiếng forms for each of the respective corpora, the full corpus (labeled as BC) and the two subcorpora, blogs and comments. It should be noted that because the lists provided in Chapter 4 Appendices B, D, and F contained forms including only numbers, additional forms were included in this analysis, so the number of forms totals more than one hundred.

The first analysis column, Morphemes, shows that for each list, there was almost a one-to-one ratio between tiếng and morphemes. In the places where there are fewer

morphemes than tiếng, this is attributable to the few bound morphemes, such as QUỐC ‘nation, national’, which were not considered words in themselves as these forms cannot appear on their own. In the comments corpus, there are two instances of more morphemes than tiếng, one in the 2-tiếng list and one in the 5-tiếng list. These forms include initials, which stand for what would have been two tiếng had the words been spelled out, with VN for Việt Nam ‘Vietnam’ and TQ for Trung Quốc ‘China’.

For the column Words, tiếng and combinations of tiếng were categorized and counted as words. For 1-tiếng forms, there is again almost a one-to-one ratio between tiếng, morphemes and words. But, as the tiếng forms get larger, the number of possible words per form does not keep up with the tiếng and morpheme counts. The column Full Form Words gives a count of how many forms may be considered as one word covering the entire form. The numbers are high for the 1-tiếng lists, but begin to decrease with the 2- and 3-tiếng lists.

Table 1: Structural Analysis of Forms

Form	# of Forms	Tieng	Morphemes	Words	Full Form Words	Phrases	Clauses	Sentences
All BC								
1-Tieng	101	101	101	97	97	0	0	0
2-Tieng	102	204	204	151	53	47	0	0
3-Tieng	102	306	305	230	8	84	6	7
4-Tieng	100	404	404	285	1	98	9	3
5-Tieng	101	505	505	359	1	99	4	2
Blogs								
1-Tieng	101	101	101	97	97	0	0	0
2-Tieng	100	200	199	153	45	52	1	1
3-Tieng	103	310	309	233	7	88	6	2
4-Tieng	100	400	400	276	0	100	7	1
5-Tieng	101	505	503	353	1	101	0	4
Comments								
1-Tieng	101	101	102	95	95	0	0	0
2-Tieng	101	202	203	146	60	38	0	0
3-Tieng	101	303	304	233	6	77	2	13
4-Tieng	101	404	404	300	0	97	6	2
5-Tieng	101	505	506	385	0	99	5	2

Appendix L gives a list of full form words across the corpora. It should be noted that a number of the 3-tiếng forms include reduplicative onomatopoeic interjections, such as HA HA HA or HÌ HÌ HÌ. And, despite attestations in the literature as to the existence of 4- and 5-tiếng forms, none appeared in the top 100+ forms lists from which this analysis derived. Digits appeared in the top 100 forms lists at each tiếng form size. For example, a number 1 appearing in the corpus list would be considered by the computer program to be a 1-tiếng form. A 4-digit form, 2013 for example, would be considered to be a 4-tiếng form. Despite their appearance in the full form list below, digits should not

be confused with tiếng, because there is no definite correspondence between the form and the number of syllables that form would equal were the number spelled out. For example, the number 10,000 could be either *một vạn* ‘one ten thousand’ or *mười nghìn* ‘ten thousand,’ each of these only two tiếng. The number 10,555 would be *mười nghìn năm trăm năm mươi lăm* ‘ten thousand five hundred fifty five.’

As mentioned in the methods section of this chapter, the columns Phrases, Clauses, and Sentences are not counts, but each number represents the presence of a phrase, clause or sentence within the form, whether it takes up the entire form or not. In a result opposite of the Word column, phrases, clauses and sentences appear most often in the 3- through 5-tiếng forms lists, with virtually no clauses or sentences in the 1- and 2-tiếng lists.

IRR analysis indicated high agreement for morphemes, with a Kappa score of 0.99. When it came to words, however, agreement declined to 0.72. This should not be a surprise, as this paper has laid out the difficulty with identifying unit segmentation above the level of the morpheme for Vietnamese. Some of the difference between the two raters includes whether plural markers should be considered as part of the word they modify or whether they remain as separate words themselves as well as whether pronouns attached to names become one word unit or whether they remain as two words. Interestingly, despite recognizing that such forms as *QUỐC* ‘nation, national’ are bound, the second rater, hereafter referred to as ‘T’, labeled these as words.

IRR Kappa ratings for Phrase, Clause, and Sentence are 0.87, 0.77, and 0.35 respectively. While the Phrase and Clause ratings are adequately high, the Sentence rating is notably low. While most items for both raters were not labeled as possible

sentences, comparison of the two raters' determinations reveals a tendency for the first rater to accept sentences with pro-drop while the second rater did not accept such forms as full sentences. For example, *CHÚC MỪNG SINH NHẬT* 'wish a happy birthday' was labeled by the first rater as a possible sentence. Conversely, the second rater labeled sentences with a subject and a modal verb only as sentences while the first rater did not. For example, *CHÚNG TA CÓ THỂ* 'we could', was labeled by the second rater as a possible sentence.

5.4 Description 2

Table 2: Content Analysis of Forms

Form	# of Forms	Content	Function	Ratio Percent	Free	Bound	Ratio Percent	Open	Closed	Ratio Percent
All BC										
1-Tieng	101	64	64	100.00%	95	8	8.42%	66	52	78.79%
2-Tieng	102	82	23	28.05%	91	10	10.99%	48	54	112.50%
3-Tieng	102	91	11	12.09%	95	7	7.37%	63	39	61.90%
4-Tieng	100	99	2	2.02%	98	3	3.06%	85	16	18.82%
5-Tieng	101	101	0	0.00%	100	1	1.00%	98	3	3.06%
Blogs										
1-Tieng	101	65	63	96.92%	97	8	8.25%	66	51	77.27%
2-Tieng	100	79	24	30.38%	88	11	12.50%	44	56	127.27%
3-Tieng	103	88	15	17.05%	95	8	8.42%	61	42	68.85%
4-Tieng	100	95	5	5.26%	96	4	4.17%	77	23	29.87%
5-Tieng	101	97	4	4.12%	99	2	2.02%	96	5	5.21%
Comments										
1-Tieng	101	65	62	95.38%	92	11	11.96%	66	51	77.27%
2-Tieng	101	86	17	19.77%	92	8	8.70%	56	45	80.36%
3-Tieng	101	92	9	9.78%	97	4	4.12%	71	30	42.25%
4-Tieng	101	101	0	0.00%	99	2	2.02%	97	4	4.12%
5-Tieng	101	101	0	0.00%	101	0	0.00%	101	0	0.00%

Table 2 above describes the relevant forms from along three axes, Content and Function, Free and Bound, and Open and Closed. As noted in the methods section of

this chapter, although the forms in the list contain larger units than mere morphemes or words, an attempt was made to categorize each full form according to whether the content within could be defined as a 'reality' that conformed to the label as such. There are many polysemous forms in Vietnamese. Forms, especially 1-tiếng forms, can function in a variety of ways, so at times forms were included in both sides of each axis. For example, some prepositions can also act as verbs, such as the word *VÀO* which can mean 'into' as well as 'to go into' therefore, it was included in both the Function and Content columns. Forms such as *BỊ* can mean 'bag' or be a 'modal verb showing lack of fortune' or be a 'passive marker', and it was used in context in each of these three ways throughout the corpus. As such, it was included in all six columns. For the column Content/Function, the likelihood that a form could be considered as one or the other is almost evenly split for 1-tiếng forms, but begins to decline sharply as the forms get larger, with very few forms considered as function forms or closed class forms at the 5-tiếng size. For Open/Closed, the likelihood that a form could be considered as one or the other is roughly even for 1-, 2--tiếng forms, but begins to decline as the forms get larger. For the columns Free/Bound; however, there is a low likelihood that any forms, from 1- to 5-tiếng, could be considered bound. In looking at the data, it seems that a few Sino-Vietnamese forms, such as *QUỐC* 'nation, national' as described above might be considered bound.

Ratio percentages for each axis are provided to give additional description. These percentages were derived by dividing the second part of each axis by the first, Function as divided by Content, Free as divided by Bound, and Closed as divided by Open. Percentages above 100% indicate that the second part has a larger count than

the first, such as for All BC, 2-tiếng, Open/Closed, where the number of closed class items is larger than the number of open class items.

IRR analysis indicated high agreement for Content and Function forms with Kappa scores of 0.89 and 0.89 respectively. IRR analysis indicated moderately high agreement for Open and Closed, with Kappa scores of 0.80 and 0.79 respectively. Agreement for Free and Bound forms was lower, with Kappa scores of 0.68 and 0.75. For the second rater, forms with partial words or where it was deemed difficult to categorize the exact meaning of the phrase as given, were not labeled at all. At first, the researcher labeled these as missing items, but then a decision was made to leave the items as is, indicating disagreement as to whether the form might be included in either side of the axis. In fact, the second rater tended to leave off the task of labeling forms along all three of the above axes if she felt there was not enough information in the form, if the form began with a partial item, or if the distribution of the items left her unable to determine the exact intent of the blogger.

5.5 Description 3

Table 3 below introduces counts for Units of Meaning. These are provided alongside Morpheme and Word columns for perspective. While counts for Units of Meaning (UM) are not quite as high as for Words, they do have the highest correspondence to this categorization. In all cases and for all forms, from 1- through 5-tiếng, the counts for UM are only a few less than for Words. One explanation for this is due to the inclusion of some units which include a noun and their respective modifier as one unit, as for example with *NGƯỜI VIỆT* 'Vietnamese person, people.' This was labeled as two

words for Word analysis, *NGƯỜI* ‘person’ *VIỆT* ‘Vietnamese,‘ but one unit of meaning for UM analysis. Also, full three-part, last, middle, and first names were labeled as three words for Word analysis, but as one unit for UM analysis.

Table 3: Units of Meaning Comparison

Form	# of Forms	Tieng	Morphemes	Words	Units of Meaning
All BC					
1-Tieng	101	101	101	97	101
2-Tieng	102	204	204	151	146
3-Tieng	102	306	305	230	207
4-Tieng	100	404	404	285	254
5-Tieng	101	505	505	359	340
Blogs					
1-Tieng	101	101	101	97	101
2-Tieng	100	200	199	153	149
3-Tieng	103	310	309	233	213
4-Tieng	100	400	400	276	253
5-Tieng	101	505	503	353	334
Comments					
1-Tieng	101	101	102	95	101
2-Tieng	101	202	203	146	143
3-Tieng	101	303	304	233	211
4-Tieng	101	404	404	300	269
5-Tieng	101	505	506	385	370

IRR analysis of UM shows a Kappa score of 0.75, indicating moderately high agreement. This is only slightly higher than the score for Words at 0.72. Again, for a language such as Vietnamese, where issues of unit segmentation have been traditional foci for linguists in the past and which are still not adequately resolved today, this is not a surprise. Comparison indicates that the second rater tended to label kinship

pronominal/name combinations as both one word and one unit, while the second rater tended to keep these separate for both categories. Also, items like the 4-tiếng unit were considered differently between the two raters. For example, *MỘT NGƯỜI ĐÀN ÔNG* ‘a man’ was considered to be two units by the first rater separated as *MỘT* ‘one’ and *NGƯỜI ĐÀN ÔNG* ‘man.’ and two units separated as *MỘT NGƯỜI* ‘a person’ and *ĐÀN ÔNG* ‘man’ by the second rater.

5.6 Description 4

Tables 4, 5, and 6 below give an analysis of structure types, counts and examples for Units of Meaning (UM). Variation in structure types expanded as the number of tiếng in the forms grew. As the different corpora included different forms, there are also different structural possibilities noted for each list.

For 1-tiếng forms, there is only one possible distribution type within the form. For 2-tiếng forms, the possible distributional structures expanded with up to three possibilities, 2, 1+1 and 1+1(2). For 3-tiếng forms, the structural possibilities included 3, 1+2, 2+1, 2+1(2), and 1+1+1. Forms with a number in parenthesis as in 1+1(2) and 2+1(2) above indicate one of those initial forms as mentioned in section 5.3 above: VN for *Việt Nam* ‘Vietnam’ or TQ for *Trung Quốc* ‘China’.

For 4-tiếng forms, the structural possibilities included 4, 1+3, 2+2, 3+1, 1+1+2, 1+2+1, 2+1+1, 2+1+1(2), and 1+1+1+1. The structure 2+1+1(2), the type Mekong, has what would traditionally be spelled as two tiếng spelled here as a one tiếng unit. Although there were no forms that might be considered to be 4-tiếng full words, there are many forms to be found in the respective corpora if considered from a Units of

Meaning perspective. For 5-tiếng forms, the structural possibilities included 1+4, 2+3, 3+2, 4+1, 1+1+3, 1+2+2, 2+1+2, 2+2+1, 1+1+1+2, 1+1+2+1, 1+2+1+1, 2+1+1+1, 1+1+1+1+1, and 1+1+1(2)+1+1. There were no 5-tiếng full form Units of Meaning.

Examples provided in column Example 1 were those that tended to be characterized as collocational/content forms. Examples provided in column Example 2 were those that tended to be characterized as colligational/function forms. The lack of available function forms for column Example 2 indicates that there were no strictly function forms for this column within that structure.

Underlined forms indicate either a form that includes initials or a partial word, such as *QUỐC* 'nation' 'national' and *ƠN* 'favor' in *ƠN EM CHÚC EM LUÔN.* As may be seen from the lists in Appendices B, D, and F for Chapter 4, there was variation in the spelling of forms indicating 'thanks' such as *cám ơn* and *cảm ơn.* This partial collocate, *ƠN* 'favor' in its full form would render the form to mean *Cám ơn em chúc em luôn...* 'thank you/em (little sister, brother, wife, younger person), wish you/em always….'.

Table 4: Full Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
All BC					
1-Tieng	1	1	101	VIỆT	THEO
2-Tieng	1	2	56	TỰ DO	THẾ NÀO
	2	1+1	46	TÔI KHÔNG	ĐÓ LÀ
3-Tieng	1	3	24	NỀN KINH TẾ	BAO GIỜ CŨNG
	2	1+2	29	CÓ Ý NGHĨA	BIẾT BAO NHIÊU
	2	2+1	23	NÓI CHUYỆN VỚI	
	3	1+1+1	25	HƠN # NĂM	NẾU KHÔNG CÓ
4-Tieng	1	4	2		TẤT CẢ MỌI NGƯỜI
	2	1+3	3	CÁC NHÀ KHOA HỌC	LÀ MỘT TRONG NHỮNG
	2	2+2	36	THẾ LỰC THÙ ĐỊCH	CHÚNG TA CÓ THỂ
	2	3+1	3	SỰ PHÁT TRIỂN CỦA	LÚC NÀO CŨNG CÓ
	3	1+1+2	14	ĐẢNG VÀ NHÀ NƯỚC	SẼ KHÔNG BAO GIỜ
	3	1+2+1	7	TỪ NĂM # ĐẾN	CÓ LIÊN QUAN ĐẾN
	3	2+1+1	12	THÂN ÁI GỬI ANH	
	4	1+1+1+1	13	TỪ # ĐẾN #	ƠN EM ĐÃ CHIA
5-Tieng	2	1+4	3	BÀ AUNG SAN SUU KYI	LÀ MỘT TRONG NHỮNG NGƯỜI
	2	2+3	4	THỦ TƯỚNG NGUYỄN TẤN DŨNG	
	2	3+2	2	NHÀ CẦM QUYỀN VIỆT NAM	
	2	4+1	1	XÃ HỘI CHỦ NGHĨA VIỆT	
	3	1+1+3	5	GỬI ANH {LastName} {MiddleName} {First Name}	KHÔNG PHẢI LÚC NÀO CŨNG
	3	1+2+2	21	SỐ ẤN PHẨM KHOA HỌC	
	3	2+1+2	13	NHÂN QUYỀN Ở VIỆT NAM	
	3	2+2+1	8	HOÀN TOÀN ĐỒNG Ý VỚI	
	4	1+1+1+2	14	NGÀY MỚI NHIỀU NIỀM VUI	ƠN EM ĐÃ ĐỒNG CẢM
	4	1+1+2+1	6	# NĂM TRỞ LẠI ĐÂY	TRÊN CÁC TẬP SAN QUỐC
	4	1+2+1+1	4	VÀ GIA ĐÌNH NĂM MỚI	
	4	2+1+1+1	8	BÀI VIẾT CỦA ÔNG {NAME}	
	5	1+1+1+1+1	11	CHỈ BIẾT CÒN ĐẢNG CÒN	

Table 5: Blogs Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
			Blogs		
1-Tieng	1	1	101	ĂN	VÀ
2-Tieng	1	2	44	CHÍNH TRỊ	TẤT CẢ
	2	1+1	56	Ở NHÀ	CÓ NHỮNG
3-Tieng	1	3	23	ĐÔNG NAM Á	LÚC NÀO CŨNG
	2	1+2	29	TRONG TRƯỜNG HỢP	NHƯ THẾ NÀO
	2	2+1	25	CHỦ YẾU LÀ	CÔNG TRÌNH NGHIÊN
	3	1+1+1	25	ANH EM {NAME}	
4-Tieng	1	4	9	MỘT NGƯỜI ĐÀN ÔNG	MỘT CÁI GÌ ĐÓ
	2	1+3	9	# TIẾNG ĐỒNG HỒ	CHỨ KHÔNG PHẢI LÀ
	2	2+2	40	THU NHẬP BÌNH QUÂN	
	2	3+1	5	SỰ PHÁT TRIỂN CỦA	
	3	1+1+2	9	NHƯNG TRONG THỰC TẾ	
	3	1+2+1	10	ĐÃ TRỞ THÀNH MỘT	LÀM THẾ NÀO ĐỂ
	3	2+1+1	8	CÓ THỂ NÓI RẰNG	QUAN TRỌNG NHẤT LÀ
	4	2+1+1(2)	1	ỦY HỘI SÔNG MEKONG	
	4	1+1+1+1	14	ANH EM NHÀ {NAME}	
5-Tieng	1	5	1		TẤT CẢ MỌI NGƯỜI ĐỀU
	2	1+4	1		LÀ MỘT TRONG NHỮNG NGƯỜI
	2	2+3	3	CHÂU Á THÁI BÌNH DƯƠNG	
	2	3+2	2	BAN CHẤP HÀNH TRUNG ƯƠNG	
	3	1+1+3	7	HƠN # TIẾNG ĐỒNG HỒ	CŨNG LÀ MỘT TRONG NHỮNG
	3	1+2+2	30	VIẾT BÀI BÁO KHOA HỌC	VÀO NGÀY # THÁNG #
	3	2+1+2	14	HOÀNG SA VÀ TRƯỜNG SA	BÂY GIỜ LÀ THÁNG #
	3	2+2+1	9	VẤN ĐỀ LIÊN QUAN ĐẾN	CÓ THỂ CHẤP NHẬN ĐƯỢC
	3	3+1+1	3	BÀI DỰ THI SỐ #	BÁO KHOA HỌC TRÊN CÁC
	4	1+1+1+2	4	VIỆT NHƯ MỘT NGÔN NGỮ	MỖI NGÀY MỘT TẤM HÌNH
	4	1+1+2+1	7	# NĂM TRỞ LẠI ĐÂY	
	4	1+2+1+1	5	EM LÀM ƠN IM ĐI	GIỮA VIỆT NAM VÀ TRUNG
	4	2+1+1+1	5	CÔNG BỐ TRÊN CÁC TẠP	
	5	1+1+1+1+1	9	ANH {NAME} VÀ CHỊ {NAME}	

Table 6: Comments Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
			Comments		
1-Tieng	1	1	101	VUI	VỚI
2-Tieng	1	2	56	VIỆT NAM	CÁI GÌ
	2	1+1	44	CHÚC ANH	CHỨ KHÔNG
	2	1+1(2)	1	Ở VN	
3-Tieng	1	3	19	{LastName} {MiddleName} {FirstName}	HÌ HÌ HÌ
	2	1+2	32	Ở VIỆT NAM	THÌ LÀM SAO
	2	2+1	24	NGƯỜI TA KHÔNG	NÀO CŨNG CÓ
	2	2+1(2)	2	NHÂN DÂN VN	
	3	1+1+1	23	GỬI ANH {NAME}	ƠN ANH ĐÃ
4-Tieng	1	4	3	CHA TRUYỀN CON NỐI	TẤT CẢ MỌI NGƯỜI
	2	1+3	15	GỬI {LastName} {Middle Name} {First Name}	LÀ MỘT TRONG NHỮNG
	2	2+2	24	BUỔI TỐI VUI VẺ	THÁNG # NĂM #
	3	1+1+2	15	EM ĐÃ CHIA SẺ	SẼ KHÔNG BAO GIỜ
	3	1+2+1	11	VÀ GIA ĐÌNH MỘT	# BÀI BÁO KHOA
	3	2+1+1	20	GIA ĐÌNH NĂM MỚI	
	4	1+1+1+1	12	CHÚC ANH LUÔN VUI	KHÔNG PHẢI LÀ NGƯỜI
5-Tieng	2	1+4	2	BÀ AUNG SAN SUU KYI	HÒA XÃ HỘI CHỦ NGHĨA
	2	2+3	2	TỰ DO CÁI CON C	
	2	4+1	1	XÃ HỘI CHỦ NGHĨA VIỆT	
	3	1+1+3	6	GÌ TRÊN SỰ DỐI TRÁ	
	3	1+2+2	11	HƠN ĐỘC LẬP TỰ DO	
	3	2+1+2	11	MẠNH KHỎE VÀ HẠNH PHÚC	
	3	2+2+1	6	HOÀN TOÀN ĐỒNG Ý VỚI	
	4	1+1+1+2	16	ƠN EM ĐÃ ĐỒNG CẢM	
	4	1+1+2+1	7	CHỈ CÓ TRUNG QUỐC LÀ	
	4	1+2+1+1	8	VÀ GIA ĐÌNH NĂM MỚI	
	4	2+1+1+1	12	KHOA HỌC TRÊN CÁC TẬP	CẢM ƠN ANH ĐÃ GHÉ
	5	1+1+1+1+1	15	BLOG FOR YOU GREAT HTTP	ƠN EM CHÚC EM LUÔN
	5	1+1+1(2)+1+1	1	CHỈ CÓ TQ LÀ TỐT	

5.7 Analysis

Section 2.2 of Chapter 2 describes current methods used for assessing and describing the Vietnamese lexicon. This analysis will use two of these methods, the linguistic and the dictionary approaches. The linguistic approach relies on individual linguists to manually handle the data in order to assess collocational and colligational patterns. The dictionary approach involves the use of dictionaries in order to match entries with dictionary inputs. Although Chapter 2 described the dictionary method as one of the automated approaches, for this analysis, dictionaries will be used, but processing will

not be automated, rather the researcher will check data against dictionaries using manual look-up methods.

For 1-tiếng forms across all three corpora, determinations of morpheme, word and UM status is the easiest. At this level, a form either has a meaning of its own and can stand on its own or not. For morpheme and UM, correspondence is closest. The lack of correspondence for word indicates, as noted previously, that there were morphemes listed that could not be considered independent words. Whether a form is bound or not, when considering meaning, there were no forms in the top 100 lists that held no intrinsic meaning. Such units do exist for the language in general, such as the second syllable in a reduplicative form, which when taken alone would have no intrinsic meaning of its own as seen for the partial reduplicative form *vui vẻ* 'happy'. Here, the first syllable carries the meaning and the second form is the partial reduplicative syllable. As previously noted, there are many polysemous forms in Vietnamese. Although the syllable *vẻ* does carry meaning in other contexts, where it can mean 'appearance, look,' it does not carry meaning as the second syllable in the reduplicative form above. But, in the top 100 list, there were no 1-tiếng forms found that had no meaning of their own.

For the remaining 2- through 5-tiếng lists, determinations of word and UM status relied primarily on the collocations and colligations at hand, the researcher's intuition and dictionary look-up. As dictionaries tend to focus on lexis rather than grammar, it was a simple process to look up forms in order to corroborate collocations. For colligational analysis, most times it was not possible to find grammar patterns by head word or even as grouped entities having one corresponding collocational or functional

meaning. Appendices I, J and K provide lists of the Top 100 forms by corpus according to whether the forms are considered content or function forms. Appendix L provides a list of full form content words for each corpus.

As described in Chapter 4, the 2- through 5-tiếng lists are not simply forms in collocation or colligation with the 1-tiếng forms listed. The forms on the larger tiếng form lists are the top 2- through 5-tiếng forms as given and appearing at the frequencies listed in the appendices to Chapter 4. The very high frequencies of the forms on the 1-tiếng lists do include both the token rates for those 1-tiếng forms alone as well as the number of times these forms appear as part of larger forms. Using the form *là* as an example, the top 100 lists include this form as a 1-tiếng form meaning 'to be' and 'clause marker.' This form is also in collocation on the respective lists giving *rất là* 'very,' *đó là* 'there is,' *là một* 'is one,' and *nhất là* 'especially' among others.

Beginning with collocations, determination of word and UM status depended on the researcher's intuition and experience as well as dictionary corroboration. Collocations here indicates content forms to include nouns, verbs, adjectives, and adverbs primarily. Definite pronouns were considered content forms for Vietnamese in this analysis as these forms index more specific relational and hierarchical information than pronouns do for English, but they were also placed in the function form list as well. Classifiers received the same treatment and were placed in both columns, as at times they work as nouns, but at other times they perform as function forms.

Content forms appearing in the respective lists include morphemes, words, phrase, clauses, and sentences as given in the structural forms list in Table 1 above. These lists also include incomplete forms. As noted in Chapter 2, while Vietnamese is

written monosyllabically, with spaces in between syllables, 2- and 3-tiếng form words were found in the Top 100 lists. From the lists, some of the full 2-tiếng forms include *CHẾ ĐỘ* 'regime,' *CHIA SẺ* 'to share,' *CHIẾN TRANH* 'war,' *CHÍNH QUYỀN* 'political power,' *CHÍNH TRỊ* 'politics, policy,' *CHÚC MỪNG* 'to wish, to congratulate,' *CỘNG SẢN* 'communism, to be communist,' *ĐẠI HỌC* 'university,' *DÂN CHỦ* 'democracy, democratic,' *DÂN TỘC* 'people (as a nation),' *ĐẤT NƯỚC* 'country, nation,' *GIA ĐÌNH* 'family', *KHOA HỌC* 'science,' *KINH TẾ* 'economy,' *NGHIÊN CỨU* 'research, to research,' *QUỐC GIA, TÁC GIẢ, THAM NHŨNG* 'corruption, to be corrupt,' *THÂN ÁI* 'affectionate, to be affectionate,' *THÀNH PHỐ* 'city,' *TRẢ LỜI* 'to answer,' *TỰ DO* 'freedom, to be free,' *VUI VẺ* 'happy, to be happy.'

These 2-tiếng forms appear as high frequency clusters in the corpora and are attested as words by their status as head words in the dictionaries consulted (Nguyen 1963, Ban Biên Soạn Chuyên Từ Điển: New Era 2001, Bùi Phụng 2003, Ban Biên Soạn Chuyên Từ Điển: New Era 2005, Ban Biên Soạn Từ Điển Ngọc-Xuân-Quỳnh 2006, Viện Khoa Học Xã Hội Việt Nam: Viện Ngôn Ngữ Học 2009). Most of these forms are also common to other Vietnamese language texts and as such, the researcher was readily able to discern their status as 2-tiếng UM.

For other 2- and 3-tiếng forms, status as words is not as easily determined. The typical prescriptive Vietnamese noun phrase is numeral/number-measure word/classifier-noun-adjective-determiner. An example of this would be *hai người đàn ông béo này*, with a breakdown of *hai* 'two' *người* 'person' *đàn ông* 'man' *béo* 'fat' *này* 'this' for a translation of 'these two fat men.' In this example, the noun form *người* 'person' is in the slot reserved for the classifier. When referring to the noun in question,

it is possible to drop the main noun and use the classifier term alone, as in *hai người béo này* ‘these two fat people,’ or one could even drop the other forms once context is established to continue to refer to these persons by simply saying *hai người này* ‘these two people.’

For this corpus, there are several 1-tiếng forms that are classifiers or nouns that act as classifiers. The definition for the form when it acts as a noun is often related to the range of items for which that form acts as a classifier. These listed 1-tiếng forms include *BÀI* defined as ‘text, lesson, script’ and ‘”classifier for nouns denoting speeches, newspaper articles, etc.’” (Nguyen 1967 p. 14), *CÁI* ‘object, thing, item, article’ and “‘classifier for most nouns denoting inanimate things and some nouns denoting small insects”’ (1967 p. 49), *CON* ‘child, to be small/young, girl’ and “‘classifier for animals and certain inanimate things’” (1967 p. 63), and *NGƯỜI* ‘man, person, individual’ and “’classifier for adult human beings, other people, others, body’” (1967 p. 387). Some researchers consider classifiers to be a special class in themselves, while others consider classifiers to be nouns (Nguyen, et.al. 2006, Cao 1985).

There are many classifier-noun forms in the list, to include such forms as *BÀI THƠ* ‘poem,’ *BÀI VIẾT* ‘written papers,’ and CON NGƯỜI ‘human being’. Taking *BÀI VIẾT* ‘written papers’ as an example, only one of the six consulted dictionaries listed this form (Bui 2003). Results were the same for the other two 2-tiếng forms here-they appeared in some dictionaries as head words, but not in others.

What is interesting about these forms is that at times, the main noun has the same meaning as the classifier-noun combination, as for *BÀI THƠ* ‘poem,’ where *THƠ* can mean both ‘poetry’ and ‘poem.’ The presence of the classifier *BÀI* indicates one

unit of poetry or a single poem. For the form *BÀI VIẾT* ‘written papers,’ however, one cannot separate the two units to get the same meaning as *VIẾT* is a verb meaning ‘to write.’ Therefore, in this case, the measure word or classifier is required not only to change word class for the form, but also to indicate the particular unit of writing as the measure word-form combination *câu viết* would indicate a ‘line of writing,’ for example. The decision, then, about whether the classifier-noun combination represents one or two words is not an easy one. For purposes of this analysis, each of these 2-tiếng forms was considered to be one word and one UM.

Some 3-tiếng examples of the above issues include the forms *NGHIÊN CỨU SINH* ‘student researcher’ and *NHÀ KHOA HỌC* ‘scientist.’ The first form exhibits *Hán-Việt* or Sino-Vietnamese word order of modifier-noun with the word *NGHIÊN CỨU* ‘to do research, study’ acting as a modifier for *SINH* ‘student.’ The second form exhibits Vietnamese classifier/measure word-noun word order, despite the fact that *KHOA HỌC* ‘science’ is a Chinese derived form that has internal modifier-noun order. *NHÀ* can mean ‘house, dwelling, abode, family, dynasty, household, home’ and “‘classifier for experts, authorities’” (Nguyen 1967 p 388). Put together the form means ‘a scientist.’ *NHÀ KHOA HỌC* appeared in Bui (2003) as a headword and *NGHIÊN CỨU SINH* appeared in Ban Biên Soạn Chuyên Từ Điển: New Era (2001) as a headword. Also, of the five forms mentioned thus far for this section of the analysis, only one, *NHÀ KHOA HỌC*, was considered to be two words by the second rater with a word and a structural UM analysis of 1+2. For the researcher and for this analysis, these forms were also considered to be one word and one UM.

There are some affixes in Vietnamese which can affect the form type, turning verbs into nouns or adjectives into nouns. These forms are sometimes called semi-affixes (bán phụ tố) because they are not strictly affixed to the form, but appear with a space in between reflecting the practice of writing single syllables individually for Vietnamese. Examples in the form list include *SỰ PHÁT TRIỂN* 'development' in the 4-tiếng form *SỰ PHÁT TRIỂN CỦA* 'development of.' The 2-tiếng form *PHÁT TRIỂN* means 'to develop, evolve, expand.' The form *SỰ* is a bound root denoting events and matters, but it also acts as a classifier for actions and states. When put with the verb *PHÁT TRIỂN* the full form denotes 'the state of or act of developing something' or simply, 'development.' For both researchers, this form represented four morphemes, but only two words with the affix/classifier as a separate form, but not as a full word. For UM purposes, the two raters diverged, with the first rater determination of the 3-tiếng form as one UM, whereas the second rater considered the form as two UM with the structure 1+2.

For 2-tiếng pronouns, the two raters and most dictionaries were in agreement (as exhibited by including the forms as headwords). These forms include *CHÚNG TA* 'we, inclusive,' *CHÚNG TÔI* 'we, exclusive,' and *NGƯỜI TA* '"people, one, they, we, you'" (Nguyen 1967 p. 387). These forms were each considered to be one word and one UM.

Table 1: Structural Analysis of Forms

Form	# of Forms	Tieng	Morphemes	Words	Full Form Words	Phrases	Clauses	Sentences
All BC								
1-Tieng	101	101	101	97	97	0	0	0
2-Tieng	102	204	204	151	53	47	0	0
3-Tieng	102	306	305	230	8	84	6	7
4-Tieng	100	404	404	285	1	98	9	3
5-Tieng	101	505	505	359	1	99	4	2
Blogs								
1-Tieng	101	101	101	97	97	0	0	0
2-Tieng	100	200	199	153	45	52	1	1
3-Tieng	103	310	309	233	7	88	6	2
4-Tieng	100	400	400	276	0	100	7	1
5-Tieng	101	505	503	353	1	101	0	4
Comments								
1-Tieng	101	101	102	95	95	0	0	0
2-Tieng	101	202	203	146	60	38	0	0
3-Tieng	101	303	304	233	6	77	2	13
4-Tieng	101	404	404	300	0	97	6	2
5-Tieng	101	505	506	385	0	99	5	2

Looking back to Table 1, reproduced below, in order to study the issue of units of analysis Vietnamese as relates to the question of unit segmentation and meaning, this study began with the tiếng and subsequent iterations of tiếng. Tiếng as a unit has correspondence with the categories of syllable, morpheme and word, however, while all tiếng are syllables, not all tiếng are morphemes or words. Much scholarship on the Vietnamese language and units of meaning centers around whether the notion of word as a unit larger than 1-tiếng applies to Vietnamese. As evidenced by the 1-tiếng forms in the three corpora, there is an almost one-to-one correspondence between tiếng and word, which would support assertions that Vietnamese is a monosyllabic language.

The presence of 2- and 3-tiếng words across the corpora, as attested by their presence as head words in various dictionaries, adds support to contrasting opinions that Vietnamese is not merely monosyllabic. Around half of the 2-tiếng forms and a few forms in the 3-tiếng list are considered to be full words. This includes 2-tiếng content forms that do not cross any theorized collocation/colligation boundary, such as *CHẾ ĐỘ* 'regime,' *CHÚC MỪNG* 'to wish, to congratulate,' and *VUI VẺ* 'happy, to be happy,' as well as attested forms such as *BÀI THƠ* 'poem' and *NHÀ KHOA HỌC* 'scientist' that do straddle the line between collocation and colligation into the larger noun phrase.

Looking again back to Table 1 above, while the 1-tiếng lists only feature morphemes and words, at the 2-tiếng level, the structures begin to feature a mix of forms including both full words and larger constituents such as phrases. At the 3--tiếng level, the number of full words declines sharply, still supporting the notion of Vietnamese as more than monosyllabic, but also indicating the transition in the basic language from collocational to more colligational structures. For the 4- and 5-tiếng forms, the corpus features no full form 4- and 5-tiếng words, but consists of multi-tiếng combinations which form these larger colligational structures as phrases, clauses and sentences. This gives support to the notion that while Vietnamese is not strictly monosyllabic, the structure of the language features primarily colligational forms at the 3-tiếng form size and above.

The varying levels of IRR agreement between the two raters reflect several different tensions, one of which represents prescriptive and descriptive language orientations, as in the case of whether to count pro-drop and subject/modal verb forms as full sentences. Another tension and the one of particular importance for this study is

the issue of the boundary between words and phrases as for the classifier/measure word-main noun combinations. Just as there is a lack of agreement between the dictionaries consulted, the two raters did not always agree in categorizing classifier/measure word-noun combinations as words. Agreement for forms that did not cross these boundaries was very high, however, with almost complete agreement for the 2-tiếng content and function forms that could not be included in the disputed categories.

Considering most function forms and colligations, the researcher was not able to rely on dictionary corroboration in many cases and so determination of word and UM status depended on the researcher's intuition and experience as well as the ability to consider the forms from the corpus in their regularity and as they worked together to constitute grammatical relations. In most cases, these forms were considered to be separate words where UM determination depended on the transparency of the form and its token rate in the corpus.

There were many 1- and 2-tiếng function forms, to include *TRƯỚC* and *TRƯỚC KHI* 'before,' *SAU* and *SAU KHI* 'after' and *TRONG* and *TRONG KHI* 'during, while.' Each of these forms is confirmed in each of the dictionaries as a head word. The difference in each of the pairs of words occurs in use – the 1-tiếng forms are prepositions which are always followed by nouns, while the 2-tiếng forms are conjunctions which precede verbs and clauses. Each of the 1-tiếng forms was considered to be one word and one UM. The 2-tiếng forms were also considered each to be one word and one UM.

Most colligational forms – grammar patterns that were clustered together were not counted as one word or one UM. Patterns such as *ĐÂY LÀ* 'here is' and *ĐÓ LÀ* 'there is' were found to be common, but were not considered as one UM. This is recognized to be an arbitrary distinction, as many of these forms do pattern commonly and work together. In this case, the two forms act as a dummy subject.

For another example, prepositional phrases were not considered to be one UM. These forms were considered to be content forms, primarily because the object of the preposition necessitates the form's placement into the content category. As units larger than the word, though, these forms also move into the colligational realm (but, were not included in function categories). The Top 100 lists include such phrases as *CỦA CHÚNG TA* 'of us, inclusive,' considered to be three morphemes, two words- *CỦA* and *CHÚNG TA*, and two units of meaning (1+2). At this point, it seems clear that if phrases were the specific starting definition for Units of Meaning, then all phrases would likely be considered Units of Meaning.

Plurals seemed to be the most contentious areas for determination of words and units of meaning. For the first rater and for this analysis, plurals were mostly considered to be separate words and separate Units of Meaning in themselves, whereas for the second rater, plural forms attached to the classifier as one word and one UM, with the main noun as the second form, word or UM in most cases. For example, *CÁC NHÀ KHOA HỌC* 'all scientists, scientists' represents two words and two UMs, where *CÁC* is a pluralizer meaning 'all' or "'the various'" (Nguyen 1967 p. 48). For the second rater, attaching the plural to the classifier would give *CÁC NHÀ* 'experts' *KHOA HỌC* 'science'.

Another function form is *MỘT CÁCH* ‘one way, a manner’. This form acts in conjunction with adverbs and adjectives to form adverb phrases, as in *yêu một cách lãng mạn* ‘love in a manner romantic’ or ‘love romantically.’ This form is believed by Nguyen & Nguyen (1980) to be a borrowing from French after contact with the bound affix *–ment* ‘as in *malheureusement* ‘unfortunately.’ While the form *CÁCH* ‘to be distant from, manner’ appears in isolation or in clusters with other forms in other contexts with other meanings, the specific form *MỘT CÁCH* ‘in a manner’ here has one specific meaning and purpose apart from the term ‘one’ and ‘manner’. Also, the function the form serves is not transparent from simply looking at the two parts alone. For purposes of this analysis, this form was considered to be two words and one UM.

A few other colligational or lexical forms were considered for UM status. The 4-tiếng form *BẤT CỨ LÚC NÀO* ‘anytime’ consisted of 2 sets of 2-tiếng words, as in *BẤT CỨ* ‘whichever’ and *LÚC NÀO* ‘when, what time.’ For UM purposes, this was considered to be one unit, as the four tiếng act together for one concept of ‘any possible time.’ This was also the case for the 3-tiếng colligation *MỘT TRONG NHỮNG* ‘one of ‘pluralizer’.’ This was considered to be three full words, but one UM as this set of forms acts together to represent one concept of ‘one of a number of things’ as in *một trong những người đẹp nhất thế giới* ‘one of the most beautiful people in the world’ from the full corpus.

The exception to the ‘most colligational forms are not considered to be one UM’ rule was made because these forms are considered by Vietnamese grammar to be ‘structures,’ special grammatical forms that pattern together in a way specific to Vietnamese language format. These structures are seen as more than simple issues of

word order, but as cases of forms appearing together to perform a specific grammatical or colligational function. As discussed in the methods section, when considering UMs it is the combination of tiếng that are required to render meaning for these particular colligational forms, with the individual tiếng for these forms as subunits, comparable to individual morphemes as subunits for words. In fact, from this standpoint, of tiếng patterning together for grammatical effect, the form *BẤT CỨ LÚC NÀO* could be widened to include the non-contiguous form *BẤT CỨ … NÀO* 'whichever X' as *LÚC* refers to time, but could be replaced by any noun to indicate anything as acceptable, such as *bất cứ người nào* 'whichever person.' Again, however, it is noted that the choice to consider Vietnamese specific structures as UMs, but not other colligational patterns is an arbitrary one.

5.8 Conclusion

This analysis provided a deeper investigation of the full corpus for this study as well as for the two subcorpora, blogs and comments. Analysis according to the three dimensions explored above indicate many differences depending on the size of the form and the forms therein. In the structural analysis, findings show that there is an almost one-to-one ration between tiếng and morpheme, with the exception of a few bound forms. Also, while the smaller tiếng forms include primarily one- and two-tiếng words, larger forms, to include two-tiếng forms, feature words within larger phrases, clauses and sentences.

In the form content analysis, we again see that most form content is free, with a few forms considered to be bound types. For open/closed class forms and

content/lexical class forms, patterns vary by form size. For 1-tiếng forms, numbers are closer to even between categorization by class, but as the tiếng size increases the types of forms found within increase to include more content and open class forms. There are very few entirely closed and lexical class 4- and 5-tiếng forms.

For the units of meaning analysis, we see that by virtue of defining units of meaning above the level of the morpheme, there is then a count for UM that corresponds most closely to counts for words. Counts of UM do, however, include phrases like *MỘT NHÀ KHOA HỌC* 'a scientist', which would be considered to be two words for this analysis; *MỘT* 'one' and *NHÀ KHOA HỌC* 'scientist' and certain lexical phrases like *BẤT CỨ LÚC NÀO* 'anytime, whenever', which would be considered to be two words, *BẤT CỨ* 'any' and *LÚC NÀO* 'when'.

Distributional analysis of UM reveals a variety of possible structures, with more structural forms appearing as the number of tiếng increases. 1-tiếng forms in the list naturally remain at the one UM stage, while differing clusters of UM within larger forms create many different structural distribution patterns. Also, depending on size and type of form, different patterns emerge. Despite most categorization and description of units as larger than a morpheme resulting in a primarily word and phrase level analysis, for the 4-tiếng form *CHA TRUYỀN CON NỐI,* categorization as one UM occurred because this is an idiom meaning 'hereditary' (father transmits child emerges) as seen in Table 6. Other than the example just given, there are virtually no 4- and 5-tiếng full form UM, which mirrors patterns for words, but only when including phrases and idioms.

Interrater reliability analysis reveals varying levels of agreement depending on the approach. IRR analysis showed high agreement for morphemes, but only moderate

agreement for words and units of meaning. Agreement for phrases and clauses was high, but for sentences, agreement dipped to its lowest, revealing differing patterns of discrimination between raters. The first rater included pro-drop sentences, whereas the second rater included sentences featuring only a subject and a modal verb. Both raters mentioned that the influence of English-language education in language and linguistics had some effect on coding decisions.

As for analysis and determinations of exactly what a word is and units of meaning for Vietnamese, there remain issues to be clarified. The status of the classifier in word phrases as both a descriptive form that may attach to the main noun and as a separate grammatical form in the noun phrase seems to be in flux. Dictionaries include some classifier-main noun clusters as head words, but not others. The two raters did not always agree as to which combinations may be included as full words. Also, the use of affixes, which change the form of the word from adjective to noun or verb to noun are most often not included in dictionary headings. This is a productive way to form units in Vietnamese and the choice of which affix to use is not arbitrary, as seen in the token rates for certain affix/form combinations as well as in prescriptive language texts such as Binh (2003).

Furthermore, for colligational analysis, the decision has to be made whether to include all common patterns for UM analysis, or if not, which to consider. A decision was made to include recognized multiple tiếng forms as words and UM if they were in the dictionary, such as the time prepositions mentioned above, but not other common patterns such as *ĐÂY LÀ* ‘here is’ and *ĐÓ LÀ* ‘there is,’ which rely more on word order than any special colligational relationship. On the other hand, other special grammatical

patterns like *MỘT CÁCH* ‘one way, in a manner’ and patterns termed ‘structures’ for example *BẤT CỨ LÚC NÀO* ‘anytime, whenever,’ were included as single UMs for this analysis, however, these patterns were not usually considered to be one-word units, but were segmented typically by tiếng or combinations of tiếng.

The above analysis gives numerous examples of recognized forms that consist of more than one tiếng, supporting the notion that Vietnamese is not a monosyllabic language. Dictionaries confirm the status of these multi-tiếng combinations as words despite the fact that many of the consulted dictionaries note in their forewards that the jury is still out on which forms should be given that status (Nguyen 1963, Bùi Phụng 2003, Ban Biên Soạn Chuyên Từ Điển: New Era 2005, Ban Biên Soạn Từ Điển Ngọc-Xuân-Quỳnh 2006, Viện Khoa Học Xã Hội Việt Nam: Viện Ngôn Ngữ Học 2009). In addition, while evidence suggests the status of Vietnamese as more than monosyllabic, the transition between collocational to colligational structures starting with half of the forms at the 2-tiếng level and extending to include most forms at the 3-tiếng level and all forms at the 4- and 5-tiếng levels indicates that Vietnamese also tends towards phrasal level at units as small as 2-tiếng and at the phrasal level and above for units larger than 2-tiếng.

Most importantly, while lexicographers, linguists and computer programmers in Viet Nam recognize that Vietnamese is not a true monosyllabic language, as is obvious from the number of recognized multiple-tiếng content and function words, it seems that it is Western linguists that must be convinced. This seems to be a somewhat difficult proposition when the state of dictionaries is not yet adequate, where there is a lack of uniformity in judgments among the aforementioned linguists and lexicographers as to

what constituents may be included in a full word, and where the use of spaces can seem confusing to the researcher.

CHAPTER 6

FROM COUNTRY OF ORIGIN PERSPECTIVES

This chapter will address gaps in the body of Vietnamese language variety research by looking at the full blogs corpus by country of origin using the same methods and approaches as used in Chapters 4 and 5. The comments corpus will not be used in order that the sample of blogs data may be more accurately described by country of origin and by identification as Vietnamese. It was not possible to identify the national or socio-ethnic identities and geographic location of commenters.

This chapter will address the following research questions:

What are the most common syllable forms and collocational, colligational and topical patterns as revealed in a comparison between in-country and US and Australian Vietnamese language blogs?

In corpus analysis of Vietnamese language varieties, what are the implications for how we analyze data? What are the implications for existing theory concerning segmentation into meaningful units in Vietnamese? How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?

6.2 Methods

The blogs corpus was divided into three subcorpora by country of origin, Australia (AUS), United States (US), and Viet Nam (VN). Blogs were sampled to ensure individual blogger samples were not overrepresented in each by-country corpus. Any blogs containing fewer than 32,000 words were included as is, whereas for blogs containing over 32,000 words, a sample of the first 10,000, the middle 10,000 and the final 10,000 words were included. Only entire blog posts were included, so totals for each sampled blog ran between 30,000 and 32,000 words. Table 1 below shows the sampled totals by country and Appendix M shows the final sample numbers for individual blogs.

Table 1: By Country Blogs Sample Totals

Country	Blog Words	Sampled Words
AUS	974,873	360,202
US	1,178,864	619,800
VN	1,111,449	614,085

The Australia sample contains several blogs containing fewer than five thousand words, while there are other blogs which contain more than one hundred thousand. Minimum blog sizes were higher on average for the VN and the US blogs. As a result, the total sampled words for the AUS corpus contains only half as many words as the US and VN corpora. The effect of this can be seen in the word lists as described in Appendices N, P, and R, where the number of texts in which a particular form appears is quite low for the AUS sample in particular, indicating that the language patterns of particular bloggers are more likely to rise to the top.

Just as in Chapter 4, the first analysis in section 6.3 provides a general description of co-occurring language forms and topical characteristics of each of the subcorpora, Australia (AUS) United States (US) and Viet Nam (VN), with the tiếng and subsequent iterations of up to 5-tiếng units as the basic units of analysis. Syllable iterations were analyzed as 1-5 unit clusters as both collocational and colligational analyses for the aforementioned 1-, 2-, 3-, 4- and 5-tiếng combinations. Section 6.4 revisits Zipf's Law and the A-curve to assess whether or not varieties of Vietnamese as given in the three subcorpora by country of origin show the same patterns for the full, blogs, and comments corpora.

Sections 6.5 through 6.8 replicate the approaches used in Chapter 5. The first dimension of analysis in Section 6.5 is an analysis of structural forms, including the labels Morpheme, Word, Phrase, Clause, and Sentence. For the first two categorizations, Morpheme and Word, the forms were analyzed and a count was made of how many morphemes and words were contained in the form. For the remaining categorizations, Phrase, Clause, and Sentence, the forms were analyzed and any forms containing a phrase, a clause, or a sentence were labeled as such.

Section 6.6 comprises an analysis of forms along the axes Content and Function, Free and Bound, and Open and Closed. Forms were categorized according to the content, the 'realities' according to the chosen paradigms. For the forms containing units larger than words, an attempt was made to categorize each full form. For example, the 4-tiếng form # *MUỖNG CÀ PHÊ*[5] '# spoon coffee' would be labeled a

[5] This noun phrase is considered for this analysis to consist of three morphemes, #, MUỖNG and CÀ PHÊ and three words. CÀ PHÊ' 'coffee' is an example of a two-morpheme word, borrowed and unanalyzable into smaller

content form and an open form as it consists of a noun phrase, referring to numbers of spoonfuls of coffee. The word *CÀ PHÊ* 'coffee' is a cognate itself.

Section 6.7 comprised an analysis of the forms based on Sinclair's (2004) concept of units of meaning. Units occurring above the level of the morpheme were categorized based on whether they, as single tiếng or any combination of tiếng within a form, could be considered as one unit. Subsequently, total units of meaning were counted for each one- through 5-tiếng form. An additional analysis was performed to assess how these units of meaning distributed themselves within the forms. Using the example above, # *MUỖNG CÀ PHÊ* '# spoon coffee' is described as having 3 units of meaning with a structure of 1+1+2 where CÀ PHÊ 'coffee' is defined as one unit of meaning. Units of meaning is a necessarily vague, but yet central topic for this analysis, and so this analysis was done in addition to the more traditional analyses of forms.

In Chapter 5, Interrater Reliability (IRR) analysis was used to assess the reliability of the methods and definitions used. There was very high agreement for morphemes, but agreement for words and units of meaning only reached to around the low 70%s. Agreement for sentences was very low at 31%. For this analysis, inclusion into the sentence category was expanded to include the second rater's determination that forms consisting of subject pronoun and a modal verb, as in *BẠN CÓ THỂ* 'you/friend could/can', are sentences. Interestingly, repeated face to face conversation during completion of Chapter 5 had definite influence on the first rater's perception of language constituents for this analysis. As the first rater ran the IRR analyses, T's

---

constuents. Thomason states; however that there are no absolute constraints when it comes to language contact situations (Thomason, 2008). CÀ PHÊ was borrowed into Vietnamese from French café. (Nguyen, 1967).

insights comprised another source of understanding of the language and how it works. Reading articles in Vietnamese and in English is a different way of understanding the language than direct interaction with a native speaker who also studies and teaches the language. Also, because the second rater had not read the entire corpus as the first rater had, the second rater proposed more potential definitions and situations and contexts of use than the first rater had considered. Any second guessing for the first rater has been assisted with dictionary checking as a source of understanding and will be mentioned in the concluding discussion of units of meaning in Chapter 7.

6.3 Description of Corpora

Table 1 below describes some of the basic features of each of the three corpora, AUS, US, and VN. Just as in Chapter 4, this chart feature totals for *Types* as unique words and *Tokens* as the total of word usage for types. Columns for the *Top 20 Types % of Tokens* and *Top 100 Types % of Tokens* provide a percentage of tokens that the top types represent. The *Top 25% of Tokens* and *Top 80% of Tokens* columns give the sum of tokens representing 25% and 80% of word usage and the percentage of total types represented by percentage of tokens. The *1-Freq Tail column* provides data showing the tail of the distributions where forms with a frequency of one occur.

Lists of the top one hundred 1-, 2-, 3-, 4-, and 5-tiếng forms for each of the corpora; AUS, US, and VN, may be found in Appendices N, P, and R. In addition, charts showing A-curve (rank and frequency) analyses for each of the form lists may be found in Appendices O, Q, and S. While the lists presented in Appendices N, P, and R feature the top one hundred forms for each of the 1- through 5-tiếng data sets and

corpora, the use of the A-curve as in Appendices O, Q, and S shows the distribution of types and tokens based on the first 3000 forms in each list.

For the lists representing the 1-tiếng forms, each of the corpora have between ten thousand and a little over thirteen thousand unique words or types, with the US showing the highest number of unique types at 13,882. The total word usage or word total for the 1-tiếng forms is roughly six hundred thousand for the US and VN corpora, and more than half that number for AUS at over three hundred thousand.

For the lists representing the 2-, 3-, 4-, and 5- tiếng forms, please be reminded that these lists are not simple node forms and their collocates, these lists are derived in Wordsmith Tools, version 5.0 (Scott, 2008), as forms that appear in varying frequencies as clusters of 2-, 3-, 4-, and 5- tiếng forms. As with any analysis of word clusters, these lists include iterations of forms that appear in sequence within discourse that may not have any lexical or grammatical meaning, but rather appear only because discourse was broken up in sequential order to derive the required lists as explained in Chapter 5.

The Top 20 Types column shows that the top 20 types for the 1- tiếng forms for each of the corpora equals roughly 15-17% of tokens, with the AUS corpus slightly higher than the US corpus which is again higher than the VN corpus. The Top 100 Types column for these same forms contains more that 39% of all tokens, with the US corpus somewhat higher than the other two corpora at around 39% each. This shows that that for the 1- tiếng lists, 15% of the word usage is being done by only 20 types and 39.78% of the word usage is being done by 100 types. These 1- tiếng forms may be units of meaning in themselves, but also part of larger units of meaning.

Table 2: AUS, US and VN Corpus Descriptors

			Top 20 Types	Top 100 Types	Top 25% of Tokens		Top 80% of Tokens		1-Freq Tail	
Forms	**Types**	**Tokens**	**% of Tokens**	**% of Tokens**	**# of Types**	**% of Types**	**# of Types**	**% of Types**	**# of Types**	**% of Types**
AUS										
1-Tieng	11,060	362,679	17.29%	39.97%	38	0.34%	825	7.46%	4,197	37.95%
2-Tieng	140,192	335,680	1.83%	5.06%	2,160	1.54%	73,057	52.11%	96,892	69.11%
3-Tieng	257,244	311,685	0.32%	1.00%	25,759	10.01%	194,907	75.77%	229,209	89.10%
4-Tieng	274,021	289,713	0.14%	0.42%	56,737	20.71%	216,079	78.85%	262,928	95.95%
5-Tieng	262,202	269,265	0.09%	0.27%	60,254	22.98%	208,350	79.46%	256,533	97.84%
US										
1-Tieng	13,882	621,801	16.85%	40.24%	39	0.28%	867	6.25%	4,933	35.54%
2-Tieng	222,955	581,941	1.54%	4.40%	2,669	1.20%	106,607	47.82%	153,036	68.64%
3-Tieng	449,905	545,042	0.26%	0.78%	44,364	9.86%	340,897	75.77%	402,303	89.42%
4-Tieng	488,614	510,837	0.10%	0.28%	105,487	21.59%	386,447	79.09%	472,730	96.75%
5-Tieng	470,365	478,787	0.05%	0.15%	111,275	23.66%	374,608	79.64%	463,460	98.53%
VN										
1-Tieng	10,065	612,769	15.58%	39.30%	41	0.41%	794	7.89%	3,281	32.60%
2-Tieng	213,613	569,248	1.58%	4.64%	2,554	1.20%	99,764	46.70%	144,006	67.41%
3-Tieng	434,421	527,824	0.25%	0.77%	43,067	9.91%	328,857	75.70%	386,841	89.05%
4-Tieng	467,388	489,736	0.10%	0.27%	100,086	21.41%	369,441	79.04%	450,860	96.46%
5-Tieng	444,859	454,073	0.05%	0.16%	104,305	23.45%	354,045	79.59%	436,970	98.23%

As we drill down to the cluster forms lists for the Top 25 and the Top 100 Types, however, we see these numbers dwindle rapidly, indicating that while 1- tiếng forms represent a larger number of the more frequent tokens, the subsequent iterations do not appear as frequently for any of the corpora. For the 2- tiếng forms we see 1.5-1.8% for the Top 25 types and 4.4-5.0% for the Top 100 Types lists for each corpus. Percentages for the 3-, 4-, and 5- tiếng corpora are even smaller. This indicates that there is an increasingly higher percentage of low frequency forms for each of these lists. As such, the results show that for the larger form-clusters, the top forms represent less of the word usage than they do for the 1- and 2- tiếng forms. This pattern is similar to the patterns exhibited in Chapter 4 for the full and blogs and comments corpora.

For the Top 25% and Top 80% of Tokens columns, we see that twenty-five percent of the sum of tokens represents 38, 39 and 41 types for each of the corpora, respectively. These numbers rise in opposition to the pattern for the Top 20 and Top 100 Types columns. This indicates that the top types for each of the lists has a lower frequency as iterations of tiếng increases. Once again, this patterns similarly to data for Chapter 4.

Supporting the assertion of an increasingly lower percentage of high frequency forms made in the discussion of Top Types is supported by looking at the 1-Freq Tail columns. These columns show the number of types and percentage of total types represented by types with a frequency of one. For the 3-, 4- and 5- tiếng lists for each of the corpora, these numbers are near 90%. As all data in this study represent an unedited, unannotated corpus, this continues to be unsurprising, as these particular subcorpora feature many of the same variations as the full and blogs and comments

corpora. Including the sequential iterations of forms these lists also feature a variety of unique forms including foreign words, for example, forms in Chinese, Japanese, English, and Thai. In fact, several English forms appear in Top 100 lists contained in the Appendices to this chapter. Other unique forms include rare forms as well as forms showing a variety of alternative orthographies, misspellings, dialect-based spellings, pronunciation-based spellings, web addresses, and other content as discussed in Chapter 3.

6.4 Tiếng Lists and Curves

The A-curve pattern for the full, blogs, and comments corpus as seen in Chapter 4 also holds for the corpus when divided by country of origin. In addition, for these individual corpora, analysis shows that the distribution of forms according to the A-curve holds not only for 1- tiếng lists, but also for the 2-, 3-, 4- and 5- tiếng lists.

The charts in Appendix O show the A-curve for the AUS corpus, Appendix Q for the US corpus and Appendix S for the VN corpus. For each list, the inverse relationship between rank and frequency obtains. The y-axis represents the number of tokens or the frequency for types. The x-axis represents the rank of the form based on its frequency. For this analysis, rank could also stand in for the actual named form, but for purposes of this study, numbers are shown in order to demonstrate the rapidity with which rank declines based on frequency.

The following series of charts represents the AUS corpus:

Chart 1a: Aus Corpus 1-Tiếng Chart

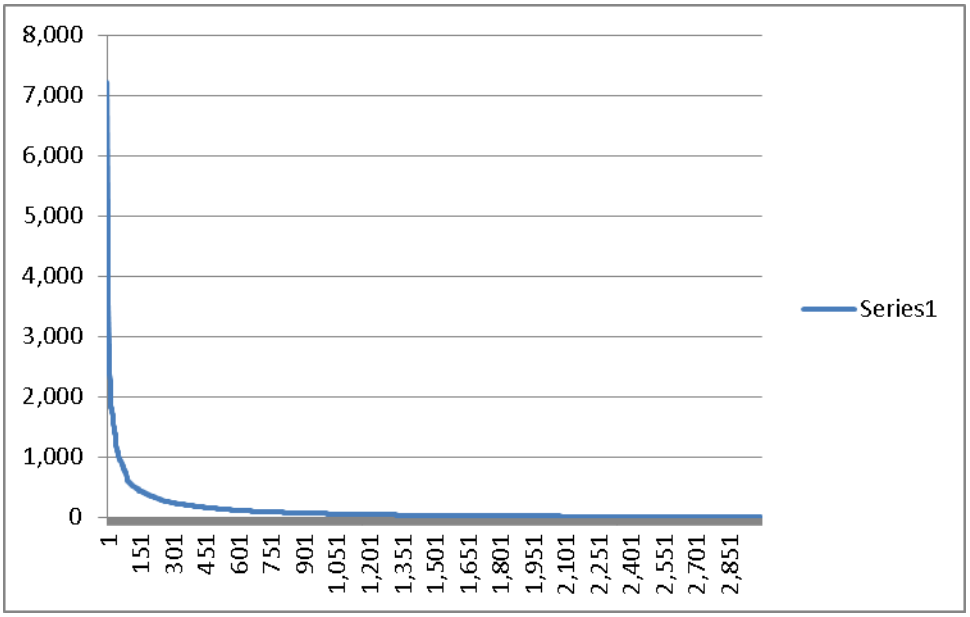


Chart 1b: Aus Corpus 2-Tiếng Chart

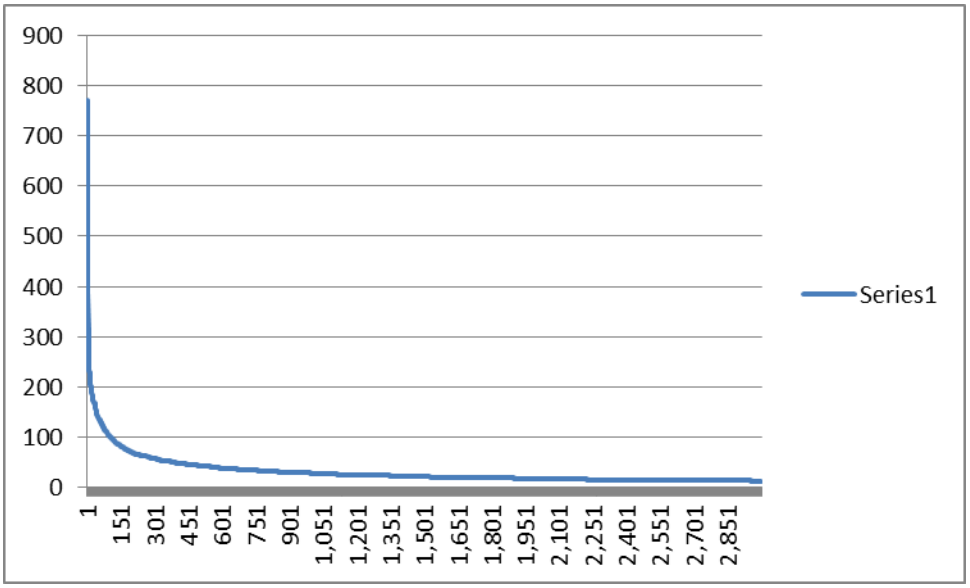


Chart 1c: Aus Corpus 3-Tiếng Chart

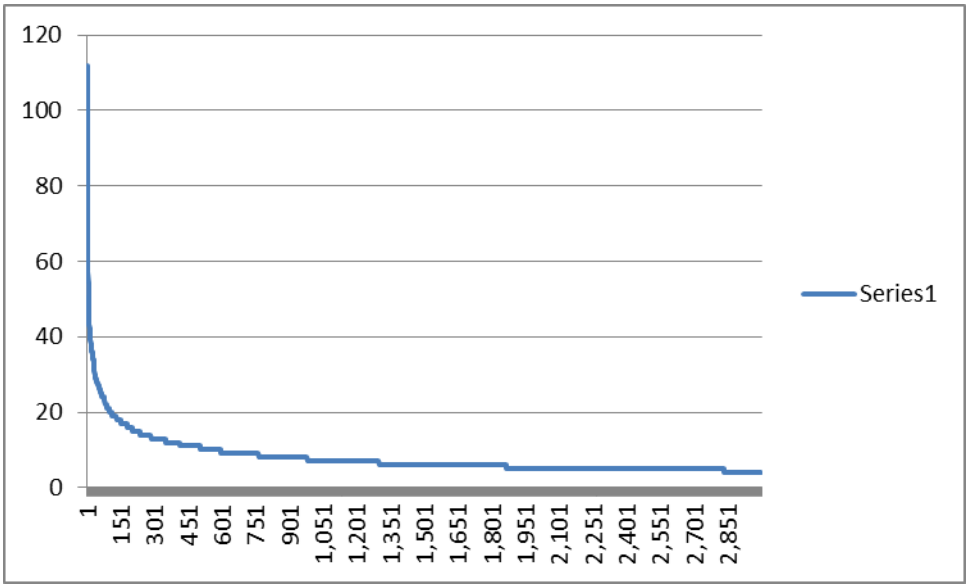


Chart 1d: Aus Corpus 4-Tiếng Chart

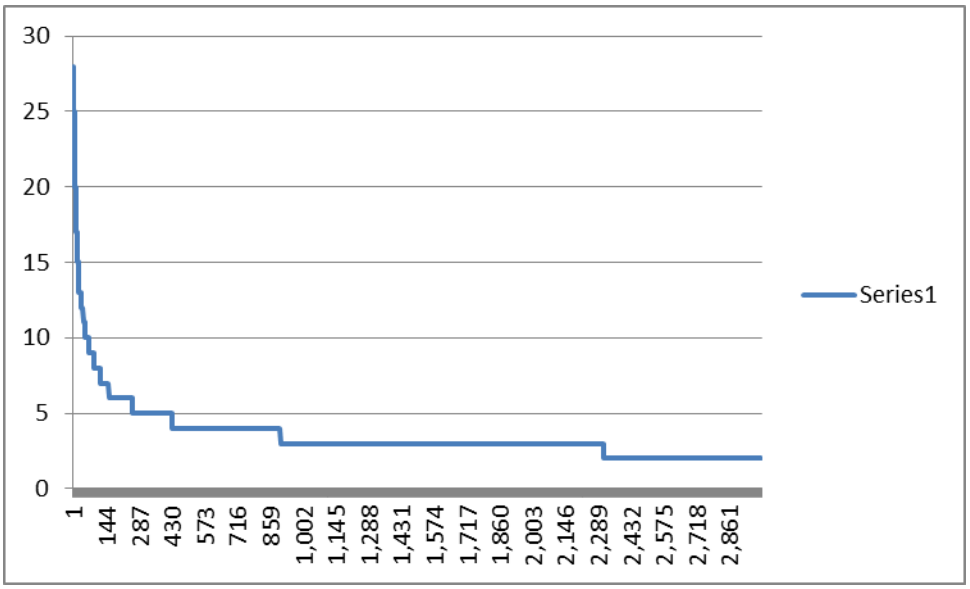
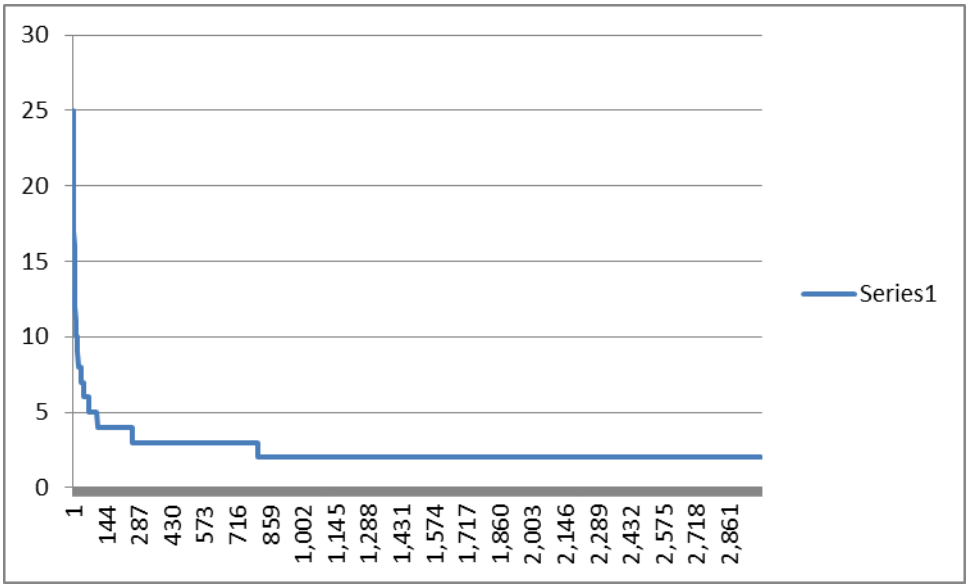


Chart 1e: Aus Corpus 5-Tiếng Chart



As shown, the a-curve begins on the y-axis with the frequency for the top ranked type. Actual frequencies for the top 100 types for all corpora are also featured on the lists in Appendices N, P, and R. For Chart 1a, we see that the top type *#,* which symbolizes the use of single digits in the text, has 7,216 tokens. As discussed on section 4.3, we further see that for the subsequent lists, top types do not appear as commonly for the 2-, 3-, 4-, and 5- tiếng forms, as the top types have a much lower number of tokens. Also, we can see how rapidly the tokens decline in frequency, with a much less sharply declining curve for the 1- tiếng list compared to the 2- tiếng list, which then declines much less sharply than the 3-tiếng list and so on.

The following series of charts represents the US corpus:

Chart 2a: US Corpus 1-Tiếng Chart

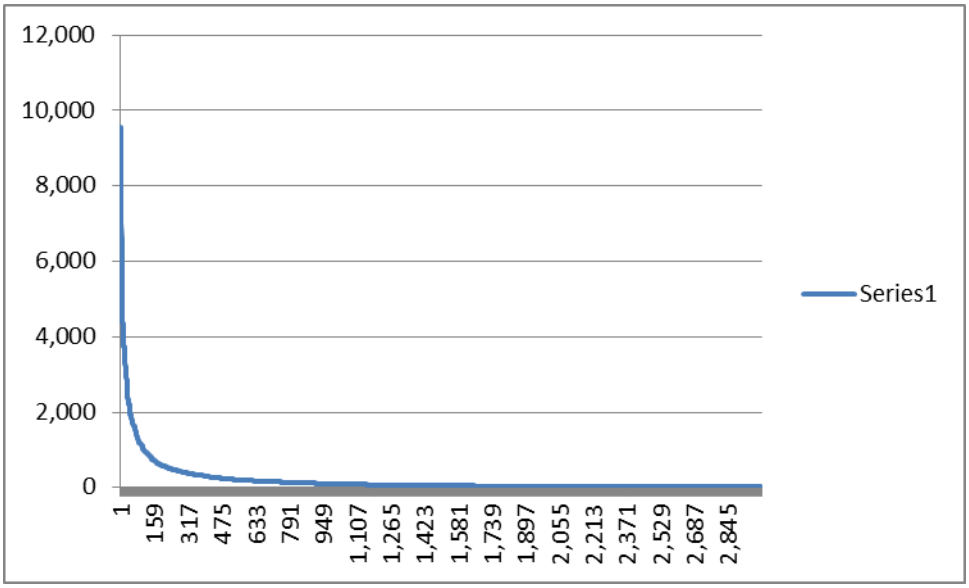


Chart 2b: US Corpus 2-Tiếng Chart

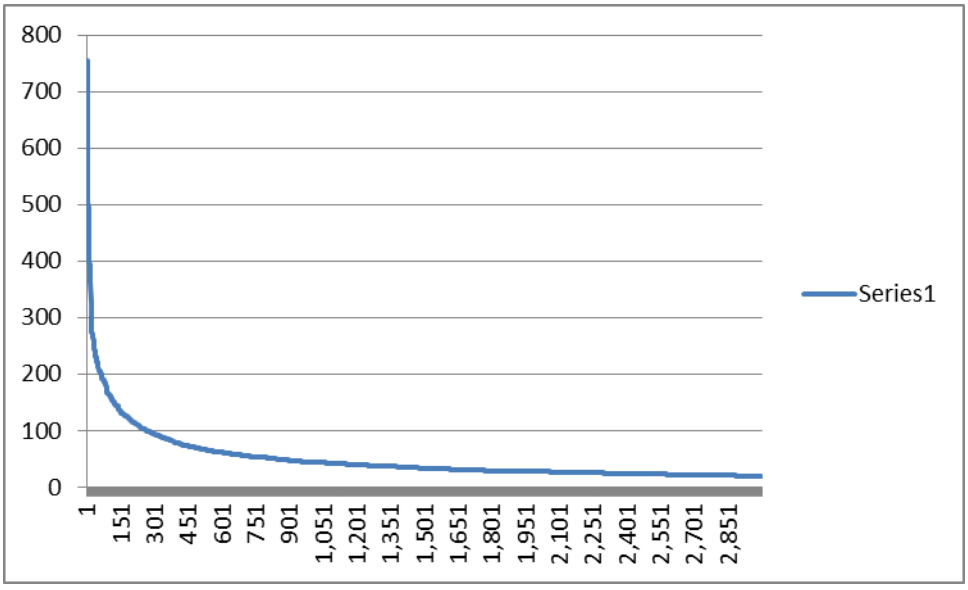


Chart 2c: US Corpus 3-Tiếng Chart

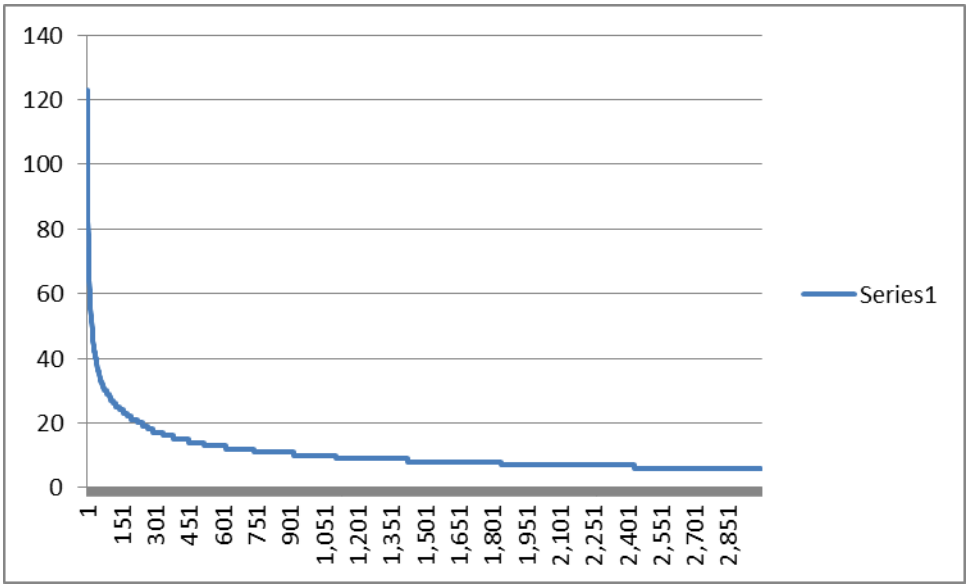


Chart 2d: US Corpus 4-Tiếng Chart

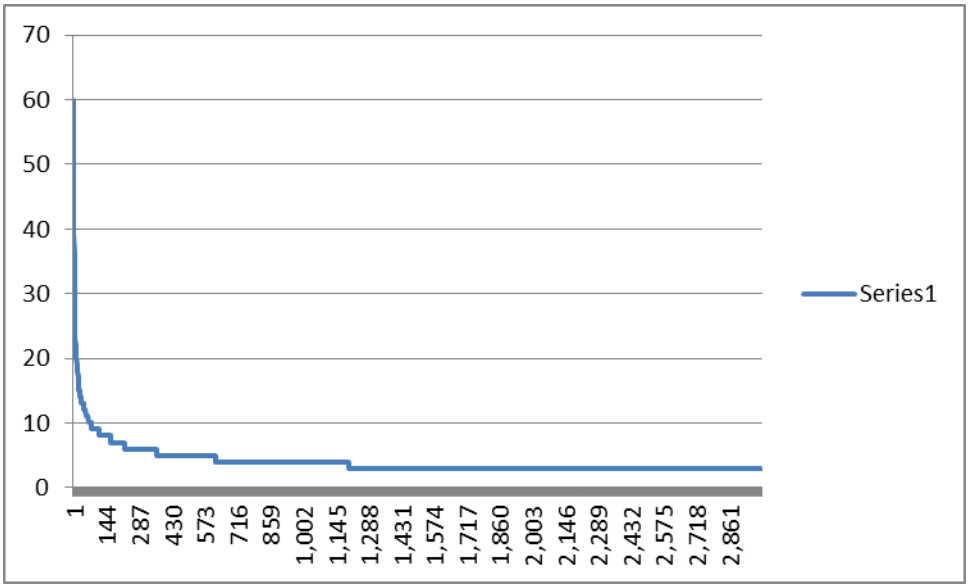
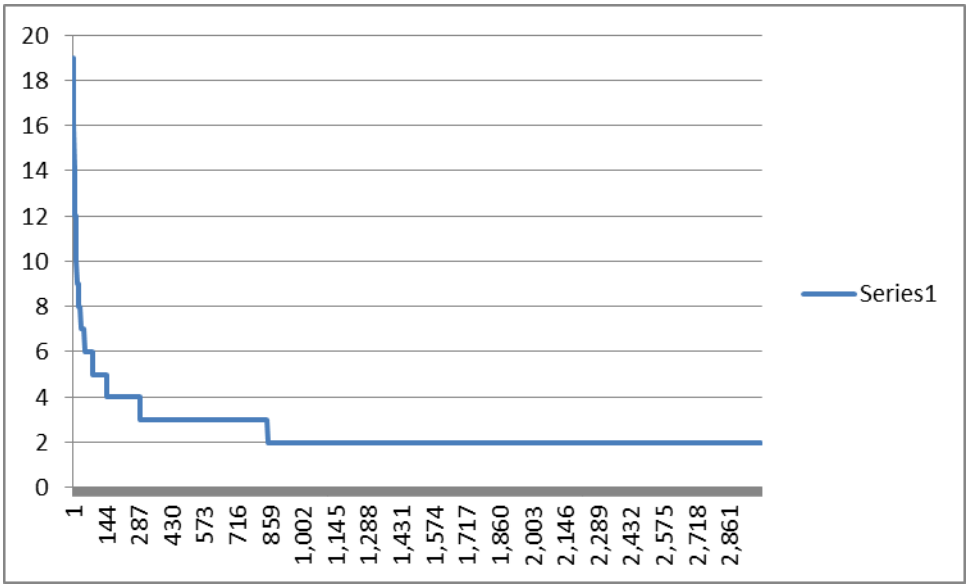


Chart 2e: US Corpus 5-Tiếng Chart



Here, we see the same pattern as for the AUS corpus lists. The a-curve begins on the y-axis with the frequency for the top ranked type. For Chart 2a, we see that the top type #, representing the use of digits in the corpus, has 9,562 tokens. As happened for the AUS corpus, we further see that for the subsequent lists, top types do not appear as commonly for the 2-, 3-, 4-, and 5- tiếng forms, as the top types have a much lower number of tokens. Also, we can see how rapidly the tokens decline in frequency, with a much less sharply declining curve for the 1- tiếng list compared to the 2- tiếng list, which then declines much less sharply than the 3-tiếng list and so on.

The following series of charts represents the VN corpus:

Chart 3a: VN Corpus 1-Tiếng Chart

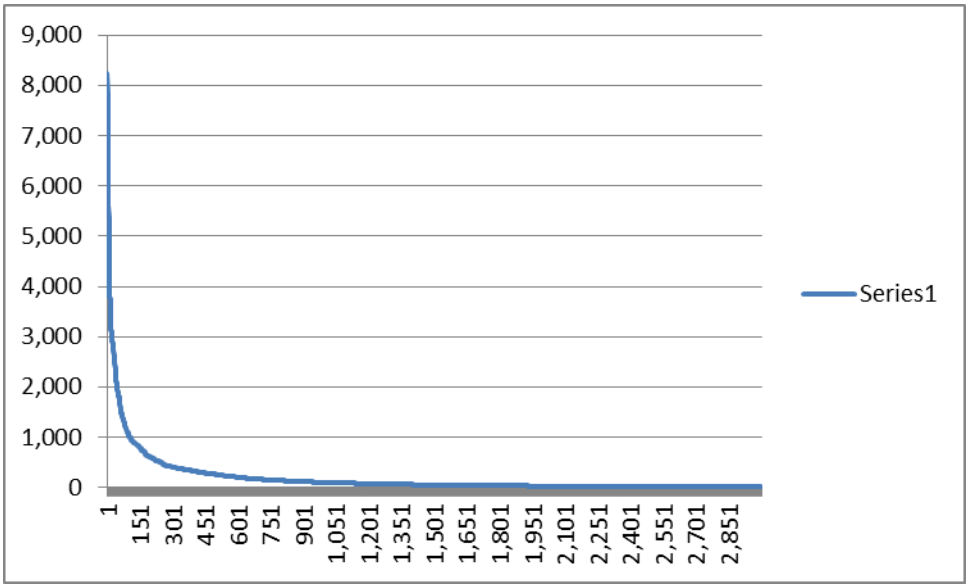


Chart 3b: VN Corpus 2-Tiếng Chart

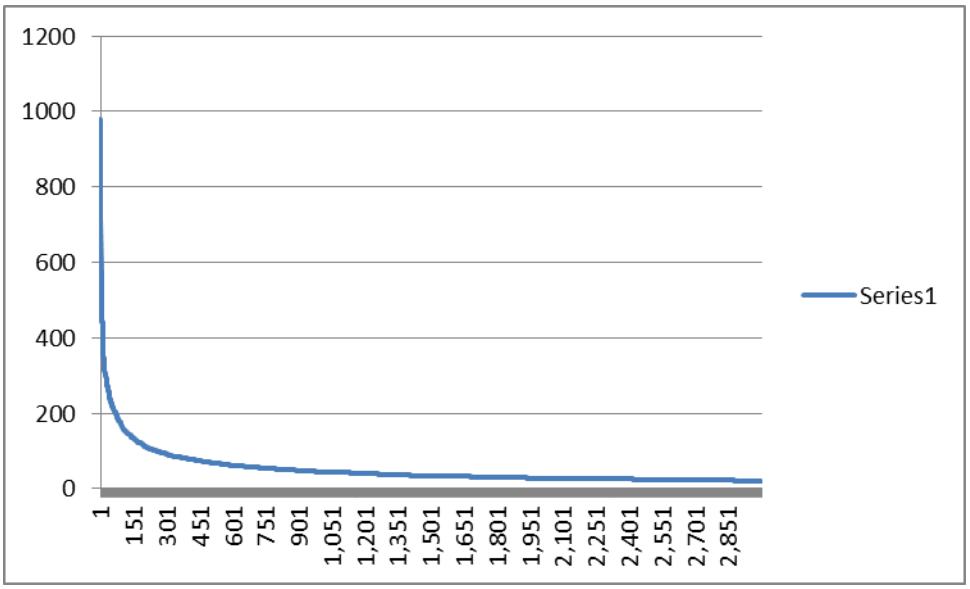


Chart 3c: VN Corpus 3-Tiếng Chart

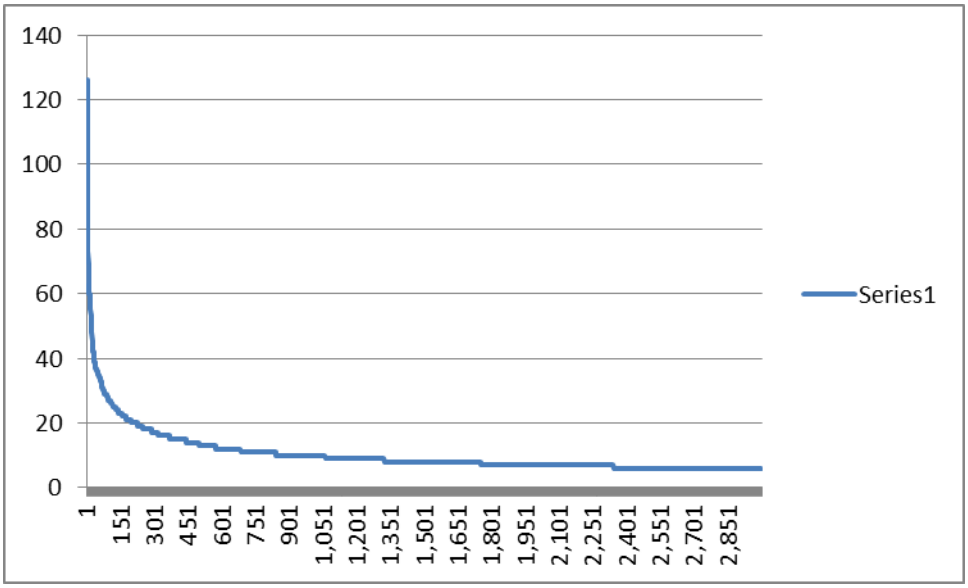


Chart 3d: VN Corpus 4-Tiếng Chart

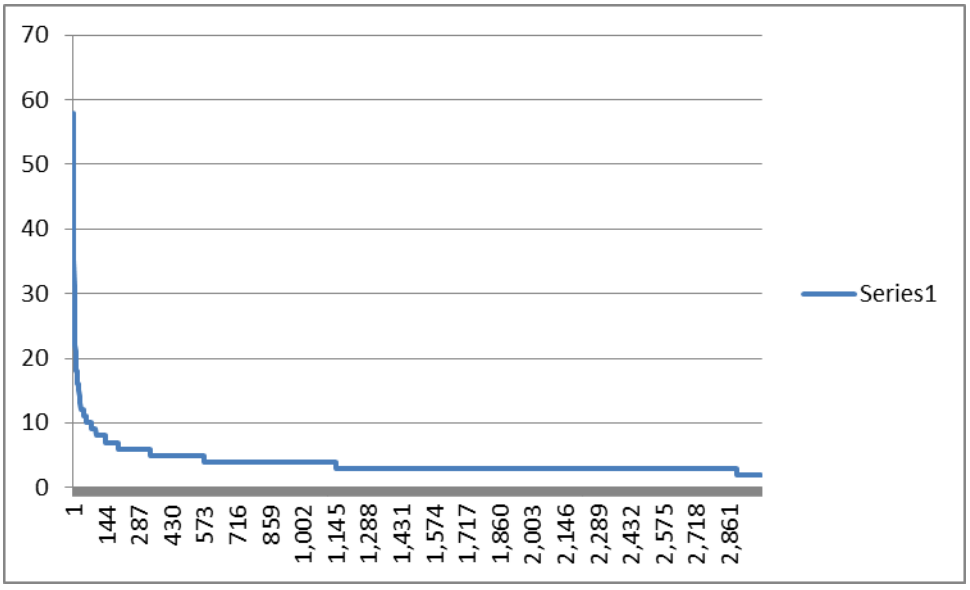
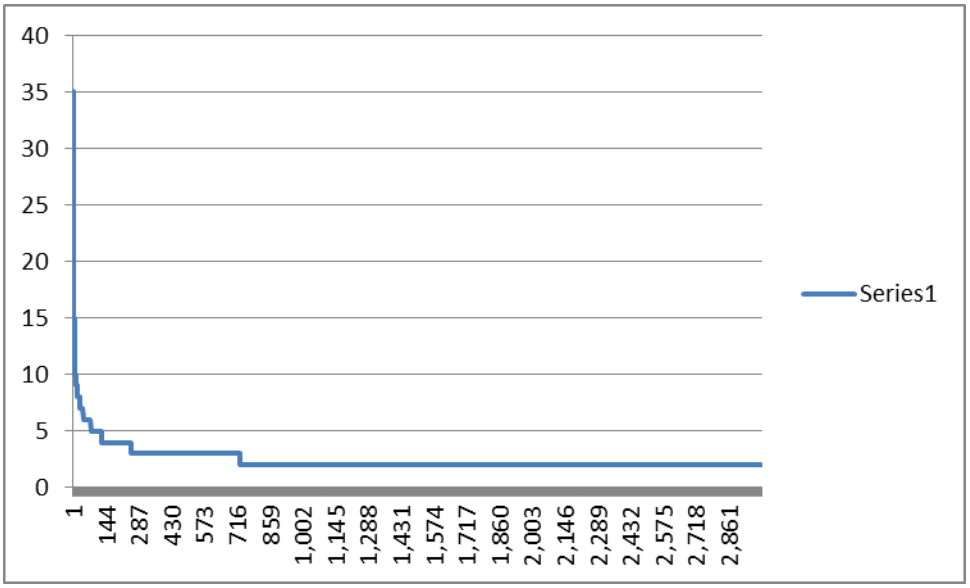


Chart 3e: VN Corpus 5-Tiếng Chart



And, again, the same pattern persists with the VN corpus as for the full and blogs corpora. The same a-curve pattern showing the inverse relationship between rank and frequency with the 1- tiếng list showing the highest initial token frequency for the type *LÀ* 'to be' at 8,248 and the least sharp decline down to the 5- tiếng list showing the lowest initial token frequency and the sharpest decline in frequency for subsequent types.

6.5 Structural Analysis

Table 3, below, describes the relevant one- through five-tiếng forms for each of the respective corpora. The first analysis column Morphemes, shows that for each list, there was almost a one-to-one ratio between tiếng and morphemes. A few tiếng were initials, such as KHKT for *khoa học kỹ thuật* 'science technology,' which explains how

there might be a higher morpheme number than tiếng in a very few cases in the 4- and 5-tiếng forms lists.

For the column Words, tiếng and combinations of tiếng were categorized and counted as words. For 1-tiếng forms, there is again almost a one-to-one ratio between tiếng, morphemes and words. Bound morphemes, such as *NHÂN* 'man, mankind, person, individual' were not considered words in themselves as these forms are considered roots. But, as the tiếng forms get larger, the number of possible words per form does not keep up with the tiếng and morpheme counts. The column Full Form Words gives a count of how many forms may be considered as one word covering the entire form. The numbers are high for the 1-tiếng lists and equal nearly half of each of the respective 2-tiếng lists. But, these numbers begin to decrease sharply with the 3-tiếng lists.

Only one potential full form word appears in the 4-tiếng list, the reduplicative form *CHUA CHUA NGỌT NGỌT* 'sour sour sweet sweet' or 'very sweet-sour.' The non-reduplicative base form *chua ngọt* 'sour sweet' is attested in three of the dictionaries (Ban Biên Soạn Chuyên Từ Điển: New Era 2001, Ban Biên Soạn Chuyên Từ Điển: New Era 2005, Ban Biên Soạn Từ Điển Ngọc-Xuân-Quỳnh 2006). As reduplication is considered a productive language process for Vietnamese, by extension the 4-tiếng reduplicative form is included in the list of full form words.

There are no 5-tiếng full form words in any of this chapter's top 100 forms lists. Several full form words for the 3- and 4-tiếng forms are digits. A list of full form words for the respective corpora appears in Appendix T.

Phrases, Clauses and Sentences are not counts, but each number represents the presence of a phrase, clause or sentence within the form, whether it takes up the entire form or not. In a result opposite of the Word column, phrases, clauses and sentences appear most often in the 2- through 5-tiếng forms lists, with no clauses or sentences in the 1-tiếng lists and virtually no clauses or sentences in the 2-tiếng lists. The patterns displayed here are similar to patterns for Chapter 5 regarding structural units of analysis, with the expanded inclusion of sentence forms containing a subject and a modal verb.

The results in Table 3 and the full forms word list in Appendix X provide more support for the conclusions given in Chapter 5 that Vietnamese cannot be considered a wholly monosyllabic language. While 1-tiếng forms predominate in these subsamples of the full corpus, there are sufficient forms in the 2-tiếng and 3-tiếng lists to contradict assertions of monosyllabicity. The range of 2-tiếng full word forms includes some forms attested in the Chapter 5 lists as well as forms that did not reach a high enough token rate to be included in those lists.

2-tiếng forms include the reduplicative and reversible form *BẠN BÈ* 'friend(s).' Reversible means that the two tiếng in the word may switch order without affecting the meaning of the word. The 2-tiếng, 1-morpheme form *CÀ PHÊ* ‘coffee,’ a borrowing from French and as such unanalyzable as mentioned in Chapter 5 is also on this list, with such classifier/measure noun collocates as *QUÁN* 'hut, inn, restaurant' and *MUỐNG* 'spoon'. The lists also includes a range of attested 2-tiếng noun, verb, adjective and adverb forms, such as *CÔNG THỨC* ‘recipe,’ *DOANH NGHIỆP* ‘trade, business,’ *ĐẦU*

TƯ ‘to invest,’ *HẠNH PHÚC* ‘happiness, to be happy,’ *HỖN HỢP* ‘joint, mixed mixture,’ and *TUY NHIÊN* ‘however.’

Table 3: Structural Analysis of Forms

Form	# of Forms	Tieng	Morphemes	Words	Full Form Words	Phrases	Clauses	Sentences
Aus								
1-Tieng	100	100	100	96	96	0	0	0
2-Tieng	100	200	200	150	50	35	1	0
3-Tieng	100	300	299	217	10	52	4	1
4-Tieng	100	400	400	281	1	63	2	4
5-Tieng	100	500	509	390	0	51	2	5
US								
1-Tieng	100	100	100	99	99	0	0	0
2-Tieng	100	200	200	160	38	44	1	0
3-Tieng	100	300	299	241	4	64	1	0
4-Tieng	100	400	400	307	1	69	0	6
5-Tieng	100	500	507	391	0	62	5	7
VN								
1-Tieng	100	100	100	96	96	0	0	0
2-Tieng	100	200	199	151	47	37	1	0
3-Tieng	100	300	299	238	2	61	2	1
4-Tieng	100	400	403	292	0	65	1	6
5-Tieng	100	500	503	377	0	50	5	7

The 2- and 3-tiếng lists also include attested forms that cross into the noun phrase classifier/measure word-main noun/verb boundary discussed in Chapter 5. These forms include *NHÀ VĂN* ’writer,’ *CUỘC ĐỜI* ‘life,’ *MÓN ĂN* ‘dish, course, as part of a meal,’ *MỐI QUAN HỆ* ‘relationship.’ The list even includes the full noun form counterpart to the verb *ĐẦU TƯ* ‘to invest,’ included in the paragraph above where the classifier *NHÀ* ‘classifier for experts, professionals’ combines with the verb to give *NHÀ ĐẦU TƯ* ‘investor.’

Interestingly, there are also two forms that could not be attested with the dictionaries consulted, but which might have been (but were not) included in the lists based on extension of other attested forms. The form *CUỘC ĐỐI THOẠI* ‘conversation, dialogue’ includes the classifier *CUỘC* ‘classifier for games, parties, meetings, actions, etc.’ (Nguyen 1967). The form *CUỐN TIỂU THUYẾT* ‘novel’ includes the classifier *CUỐN* ‘classifier for books.’ On the 2-tiếng list are the forms CUỘC ĐỜI and CUỘC SỐNG which both mean ‘life,’ but where each form takes only one part of the full 2-tiếng word *đời sống* ‘life, living, existence, livelihood.’ Both *CUỘC ĐỜI* and *CUỘC SỐNG* are attested, so it should be possible to include the above form *CUỘC ĐỐI THOẠI* in the word list. The form *CUỐN SÁCH* ‘book’ includes the classifier *CUỐN*, so it might be possible to include by extension the form *CUỐN TIỂU THUYẾT* above.

The decision not to include these three forms into the full form words list is primarily a concession to the lexicographers who wrote the dictionaries consulted. Although this dissertation uses the linguistic method, where the linguist makes decisions on whether to include a form or not, as well as the dictionary method, the fact that these forms were not attested *and* that they crossed the collocational boundary into the colligational realm was the basis for the decision. It may also be that where forms are not as common, the tendency to represent them in dictionaries with their classifier is rarer. It is proposed that the word *CUỐN SÁCH* ‘book’ would be more common across the language than the more specific form *CUỐN TIỂU THUYẾT* ‘novel.’ Also, where the 4-tiếng reduplicative form *CHUA CHUA NGỌT NGỌT* ‘very sour sweet’ was included in the list, the decision rested on knowledge of reduplication as a productive

and creative word-formation process and the understanding that as such, no dictionary would ever be able to account for all ad hoc reduplicative productions.

Beyond the word level, the patterns in the structural analysis replicate the patterns seen for the full, blogs and comments corpora in Chapter 5, where the transition from collocational to colligational structures begins with half of the forms in the 2-tiếng list to predominate in the 3-, 4-, and 5-tiếng lists. This indicates that Vietnamese lexis is primarily located in the 1- and 2-tiếng levels with larger constituents, phrases, clauses and sentences, appearing in the 2-, 3-, 4-, and 5-tiếng levels.

6.6 Content Analysis of Forms

Table 4 below describes the relevant forms from the three axes, Content and Function, Free and Bound and Open and Closed as used in Chapter 5. Although the forms in the list contain larger units than mere morphemes or words, an attempt was made to categorize each full form according to whether the content within could be defined as a 'reality' that conformed to the label as such. There are many polysemous forms in Vietnamese. Forms, especially 1-tiếng forms, can function in a variety of ways for the Vietnamese language, however, so at times forms were included in both sides of each axis. For example, some prepositions can also act as verbs, such as for the word RA which can mean 'out' as well as 'to go out,' therefore, it was included in both the Function and Content columns. Forms such as *BỊ*, which also appears in Chapter 5, can mean 'bag,' 'modal verb showing lack of fortune' and 'passive marker', and it was used in context in each of these three ways throughout the corpus. As such, it was included in all six columns.

Table 4: Content Analysis of Forms

Form	# of Forms	Content	Function	Percent: Content/ Function	Open	Closed	Percent: Open/Closed	Free	Bound	Percent: Free/Bound
Aus										
1-Tieng	100	63	53	84.13%	55	65	118.18%	93	9	9.68%
2-Tieng	100	70	35	50.00%	67	44	65.67%	96	4	4.17%
3-Tieng	100	81	20	24.69%	84	27	32.14%	98	2	2.04%
4-Tieng	100	97	3	3.09%	98	7	7.14%	99	1	1.01%
5-Tieng	100	100	0	0.00%	100	0	0.00%	100	0	0.00%
US										
1-Tieng	100	58	57	98.28%	53	68	128.30%	94	7	7.45%
2-Tieng	100	73	30	41.10%	66	46	69.70%	97	3	3.09%
3-Tieng	100	82	18	21.95%	81	33	40.74%	97	3	3.09%
4-Tieng	100	97	4	4.12%	97	9	9.28%	99	1	1.01%
5-Tieng	100	99	3	3.03%	99	4	4.04%	99	1	1.01%
VN										
1-Tieng	100	61	56	91.80%	55	65	118.18%	93	8	8.60%
2-Tieng	100	72	32	44.44%	68	45	66.18%	95	5	5.26%
3-Tieng	100	77	22	28.57%	79	33	41.77%	96	4	4.17%
4-Tieng	100	90	11	12.22%	90	18	20.00%	99	1	1.01%
5-Tieng	100	99	2	2.02%	98	4	4.08%	100	0	0.00%

For the columns Content/Function and Open/Closed, the likelihood that a form could be considered as one or the other is almost evenly split for 1-tiếng forms, but begins to decline as the forms get larger, with very few or no forms considered as function forms or closed class forms at the 5-tiếng size. For the columns Free/Bound, however, there is a low likelihood that any forms, from 1- to 5-tiếng, could be considered bound. In looking at the data, it seems that a few Sino-Vietnamese forms, such as *NHÂN* ‘human, humane’ as described above might be considered bound.

Ratio percentages for each axis are provided to give additional description. These percentages were derived by dividing the second part of each axis by the first, Function as divided by Content, Free as divided by Bound and Closed as divided by Open. Percentages above 100% indicate that the second part has a larger count than the first, such as for each corpus’s 1-tiếng, Open/Closed, cell where the number of

closed class items is larger than the number of open class items. Once again, these findings are similar in distribution and scope as the findings in Chapter 5 for the full, blogs, and comments corpora.

6.7 Units of Meaning

Table 5 below introduces counts for Units of Meaning. These are provided alongside Morpheme and Word columns for perspective. While counts for Units of Meaning (UM) are not quite as high as for Words, they do have the highest correspondence to this categorization, a pattern seen as well in Chapter 5. In all cases and for all forms, from 1- through 5-tiếng, the counts for UM are only a few less than for Words. One explanation for this as explained in Chapter 5 is due to the inclusion of some units which include a noun and their respective modifier as one unit, as for example with *NGƯỜI VIỆT* 'Vietnamese person, people.' This was labeled as two words for Word analysis, *NGƯỜI* 'person' *VIỆT* 'Vietnamese,' but one unit of meaning for UM analysis. This form also crosses the classifier-noun boundary where *NGƯỜI* would be the measure noun/classifier, but does not appear in the consulted dictionaries. As such, it is counted in different ways for each of the Word and UM categories. Also, full three-part, last, middle, and first names were labeled as three words for Word analysis, but as one unit for UM analysis.

Table 5: Units of Meaning Comparison

Form	# of Forms	Tieng	Morphemes	Words	Full Form Words	Units of Meaning
Aus						
1-Tieng	100	100	100	96	96	100
2-Tieng	100	200	200	150	50	146
3-Tieng	100	300	299	217	10	201
4-Tieng	100	400	400	281	1	249
5-Tieng	100	500	509	390	0	354
US						
1-Tieng	100	100	100	99	99	100
2-Tieng	100	200	200	160	38	156
3-Tieng	100	300	299	241	4	223
4-Tieng	100	400	400	307	1	282
5-Tieng	100	500	507	391	0	356
VN						
1-Tieng	100	100	100	96	96	100
2-Tieng	100	200	199	151	47	145
3-Tieng	100	300	299	238	2	211
4-Tieng	100	400	403	292	0	263
5-Tieng	100	500	503	377	0	346

As seen for other analyses from Chapters 4 and 5 and this chapter, similar patterns across corpora and for the range of tiếng forms holds. The overall behaviors and collocational and colligational patterns as revealed by the above analyses of form types vary primarily in relative frequency rather than in their grosser patterns. Looking at individual forms, there is also a variation in content, for example with the 4-tiếng reduplicative form *CHUA CHUA NGỌT NGỌT*, which only appears in the Top 100 list for the US blogs. Other variations in content can be seen on the Full Form Word List in

Appendix T. A deeper analysis will discuss differences between the corpora in Section 6.9.

6.8 Structure of Units of Meaning

This section provides an overview of individual corpora structure types, counts, and examples for Units of Meaning (UM). In a very similar fashion to patterns observed in Chapter 5, variation in structure types expanded as the number of tiếng in the forms grew across the corpora. As the different corpora included different forms, there are also different structural possibilities noted for each.

Table 6 gives a side by side overview of the structure types and frequencies by corpus, while Tables 7, 8, and 9 provide a few examples alongside the individual corpus-specific structural possibilities. Once again, examples provided in column Example 1 of Tables 7, 8, and 9 were those that tended to be characterized as content forms. Examples provided in column Example 2 were those that tended to be characterized as function forms. The lack of available function forms for column Example 2 indicates that there were no strictly function forms for this column within that structure. Underlined forms include initials or partial units.

As in Chapter 5, for Tables 7, 8, and 9, examples provided in column Example 1 were those that tended to be characterized as content forms. Examples provided in column Example 2 were those that tended to be characterized as function forms, despite the presence of content within. The lack of available function forms for column Example 2 indicates that there were no strictly function forms for this column within that structure. Underlined forms include initials or partial units.

Similar coding decisions were made as for Chapter 5. For content forms, units larger than a morpheme were included as single UM. For forms that represented colligational structures requiring any number of single tiếng in order to actuate the structural pattern, units above the word were included as single UMs. Also, each corpus featured tiếng as initials, as shown in the individual corpus tables where the initials are underlined.

Results as shown in Tables 6, 7, 8, and 9 indicate that while there is much similarity to the form structures across the corpora, there are structures that were more prevalent in one corpus than another and some structures that appeared in one or two of the corpora, but not all three. These results show some small variation in the sorts of UM structures and frequencies of use across the corpora subsamples. The differences as revealed, however, do not indicate any radical departure from basic structures from one corpus to another.

For example, each corpus featured some English content, each corpus included forms with initials, and each corpus included similar 1-, 2-, 3-, and 4-tiếng collocational and colligational structures. The 5-tiếng 1-unit structure for the US corpus was *TẤT CẢ MỌI NGƯỜI ĐỀU* ‘everyone, every person,’ a colligational structure explained in Chapter 5 which includes the form *ĐỀU* ‘both, every, all,’ which separates the topic from the comment in Vietnamese sentence structure. Topic-comment is a common sentence type for Vietnamese, especially spoken Vietnamese. Smaller parts of this structure appeared in each of the corpora, so this is not a unique structure overall for any corpus.

Table 6: Side by Side Comparison of UM Structures and Frequencies

AUS		US		VN	
Structure	**Freq**	**Structure**	**Freq**	**Structure**	**Freq**
1	100	1	100	1	100
2	54	2	43	2	54
1+1	46	1+1	57	1+1	46
3	20	3	11	3	14
1+2	35	1+2	38	1+2	39
2+1	24	2+1	15	2+1	21
1+1+1	21	1+1+1	36	1+1+1	26
4	5	4	2	4	5
1+3	1	1+3	10	1+3	7
2+2	35	2+2	26	1(2)+3	2
3+1	5	3+1	3	2+2	30
1+1+2	9	1+1+2	14	1+1+2	12
1+1+1(2)	1	1+2+1	15	1(3)+1+2	1
1+2+1	18	2+1+1	6	1+2+1	18
2+1+1	11	1+1+1+1	24	2+1+1	12
1+1+1+1	8			1+1+1+1	13
2+3	6	5	1	1+4	1
3+2	4	1+4	2	2+3	7
1+1+3	5	2+3	4	1+2+2	18
1+2+2	14	3+2	1	1+1+3	1
1+3+1	3	1+1+3	6	2+1+2	15
2+1+2	11	1+2+2	17	2+2+1	14
2+2+1	5	2+1+2	9	1+1+1+2	9
3+1+1	6	2+2+1	9	1+1+2+1	6
1+1+1+2	6	3+1+1	1	1+2+1+1	12
1+1+2+1	9	1+1+1+2	9	2+1+1+1	6
1+2+1+1	7	1+1+2+1	9	2+2+2+2	1
1+2+1+1(2)	1	1+2+1+1	8	1+1+1+1+1	10
2+1+1+1	4	2+1+1+1	10		
2+2+2+2	1	3+2+2+1	1		
1+1+1+1+1	18	1+1+1+1+1	16		

Table 7: AUS Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
			AUS		
1-Tieng	1	1	100	CHUYỆN	CẢ
2-Tieng	1	2	54	THỊ TRƯỜNG	VẪN CÒN
	2	1+1	46	KHOẢNG #	ĐẾN KHI
3-Tieng	1	3	20	NHÀ LÃNH ĐẠO	MỘT TRONG NHỮNG
	2	1+2	35	VỀ NHÂN QUYỀN	
	2	2+1	24	NÓI CHUYỆN VỚI	
	3	1+1+1	21	LÀ NHỮNG NGƯỜI	TRONG ĐÓ CÓ
4-Tieng	1	4	5	CHỦ NGHĨA XÃ HỘI	XÃ HỘI CHỦ NGHĨA
	2	1+3	1	CÁC CHÍNH TRỊ GIA	LÀ MỘT TRONG NHỮNG
	2	2+2	35	NGƯỜI VIỆT TỴ NẠN	CHÚNG TA CÓ THỂ
	2	3+1	5	BÀI PHÁT BIỂU CỦA	
	3	1+1+2	9	THOẠI VỀ NHÂN QUYỀN	SẼ KHÔNG BAO GIỜ
	3	1+1+1(2)	1	ỦY HỘI SÔNG MEKONG	
	3	1+2+1	18	SỐ ẤN PHẨM KHOA	LÀM THẾ NÀO ĐỂ
	3	2+1+1	11	ĐỐI THOẠI VỀ NHÂN	CÓ THỂ LÀM ĐƯỢC
	4	1+1+1+1	8	SIDE OF THE WORLD	TỪ # ĐẾN #
5-Tieng	2	2+3	6	CHÂU Á THÁI BÌNH DƯƠNG	
	2	3+2	4	NHÀ CẦM QUYỀN VIỆT NAM	
	3	1+1+3	5	THƠ CỦA JUAN RAMÓN JIMÉNEZ	
	3	1+2+2	14	NHỮNG CON SỐ THỐNG KÊ	
	3	1+3+1	3	CÁC CUỘC ĐỐI THOẠI VỀ	
	3	2+1+2	11	PHÁT BIỂU CỦA DÂN BIỂU	
	3	2+2+1	5	BÀI BÁO KHOA HỌC TRÊN	
	3	3+1+1	6	MÀU VÀNG ÚA VÀ MÀU	
	4	1+1+1+2	6	GIỮA MỸ VÀ TRUNG QUỐC	
	4	1+1+2+1	9	TRỨNG Ở NHIỆT ĐỘ PHÒNG	
	4	1+2+1+1	7	XEM PHẦN # Ở ĐÂY	CHO BÁC SĨ VÀ Y
	4	1+2+1+1(2)	1	CỦA ỦY HỘI SÔNG MEKONG	
	4	2+1+1+1	4	KHOA HỌC TRÊN CÁC TẠP	
	4	2+2+2+2	1	CỘNG ĐỒNG NVTD ÚC CHÂU	
	5	1+1+1+1+1	18	UPON A TIME IN CABRAMATTA	

Table 8: US Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
			US		
1-Tieng	1	1	100	HÌNH	LẠI
2-Tieng	1	2	43	BẠN BÈ	LÀM SAO
	2	1+1	57	# GIỜ	CŨNG KHÔNG
3-Tieng	1	3	11	TRƯỜNG ĐẠI HỌC	TẤT CẢ NHỮNG
	2	1+2	38	MÙA GIÁNG SINH	KHÔNG BAO GIỜ
	2	2+1	15	GIA ĐÌNH MÌNH	
	3	1+1+1	36	MỸ GỐC VIỆT	
4-Tieng	1	4	2		TẤT CẢ MỌI NGƯỜI
	2	1+3	10	CÁC NHÀ ĐẦU TƯ	LÀ MỘT TRONG NHỮNG
	2	2+2	26	NGƯỜI VIỆT TỴ NẠN	CHÚNG TA CÓ THỂ
	2	3+1	3	XỨ CAO BỒI NÀY	MỌI NGƯỜI ĐỀU CÓ
	3	1+1+2	14	# MUỖNG CÀ PHÊ	KHÔNG BIẾT BAO NHIÊU
	3	1+2+1	15	NĂM TRỞ LẠI ĐÂY	KHÔNG CÓ GÌ ĐỂ
	3	2+1+1	6	NƯỚC MẮM CHANH TỎI	CÁCH ĐÂY # NĂM
	4	1+1+1+1	24	NATIONAL GALLERY OF ART	
5-Tieng	1	5	1		TẤT CẢ MỌI NGƯỜI ĐỀU
	2	1+4	2		CỦA TẤT CẢ MỌI NGƯỜI
	2	2+3	4	CỜ BẠC BẤT HỢP PHÁP	
	2	3+2	1	DU HỌC SINH VIỆT NAM	
	3	1+1+3	6	THƯ CHO ÔNG GIÀ NOEL	CŨNG LÀ MỘT TRONG NHỮNG
	3	1+2+2	17	LUẬT BẢO HIỂM SỨC KHỎE	
	3	2+1+2	9	CẦU NGUYỆN CHO HÒA BÌNH	
	3	2+2+1	9	TIẾNG NÓI NGƯỜI MỸ GỐC	
	3	3+1+1	1		MỌI NGƯỜI ĐỀU CÓ QUYỀN
	4	1+1+1+2	9	KÍNH MỜI QUÝ ĐỒNG HƯƠNG	
	4	1+1+2+1	9	PHILIPPINES TRẠI TỊ NẠN BATAAN	
	4	1+2+1+1	8	CHIỀU THÀNH PHỐ MƯA BAY	
	4	2+1+1+1	10	NÊM NẾM CHO VỪA ĂN	
	4	3+2+2+1	1	CUỘC CÁCH MẠNG KHKT LẦN	
	5	1+1+1+1+1	16	UPON A TIME IN CABRAMATTA	

Table 9: VN Corpus Distribution of Units of Meaning

Form	Units	Structure Types	Counts	Example 1	Example 2
			VN		
1-Tieng	1	1	100	VĂN	VÌ
2-Tieng	1	2	54	NHÂN VẬT	SAU KHI
	2	1+1	46	MỘT CHÚT	KHÔNG CÒN
3-Tieng	1	3	14	CUỐN TIỂU THUYẾT	LÚC NÀO CŨNG
	2	1+2	39	CẤP THÂM NIÊN	
	2	2+1	21	TRỞ THÀNH MỘT	NÀO CŨNG CÓ
	3	1+1+1	26	KHÔNG CÓ GÌ	MỘT TRONG NHỮNG
4-Tieng	1	4	5	RẠCH GẦM XOÀI MÚT	TẤT CẢ NHỮNG GÌ
	2	1+3	7	NHỮNG NGƯỜI PHỤ NỮ	
	2	1(2)+3	2	BS {Last} {Middle} {First}	
	2	2+2	30	TRÁI TIM BẠC NHƯỢC	CHÚNG TA CÓ THỂ
	3	1+1+2	12	# MUỖNG CÀ PHÊ	KHÔNG BIẾT BAO NHIÊU
	3	1(3)+1+2	1	NXB HỘI NHÀ VĂN	
	3	1+2+1	18	KHÔNG NHẤT THIẾT PHẢI	MỘT CÁI GÌ ĐÓ
	3	2+1+1	12	LÀM ƠN IM ĐI	
	4	1+1+1+1	13	KHÔNG PHẢI LÀ MỘT	
5-Tieng	2	1+4	1		VỚI TẤT CẢ MỌI NGƯỜI
	2	2+3	7	GIAO TIẾP PHI NGÔN NGỮ	
	3	1+2+2	18	PHỐ BIẾN HÒA THỜI GIAN	
	3	1+1+3	1		KHÔNG PHẢI LÚC NÀO CŨNG
	3	2+1+2	15	HỘI CHỢ HÀNG THỦ CÔNG	
	3	2+2+1	14	DOANH NGHIỆP KINH DOANH XĂNG	
	4	1+1+1+2	9	BÌA COCKTAIL CHO TÌNH YÊU	
	4	1+1+2+1	6	LỜI LÀM TAN NÁT LÒNG	
	4	1+2+1+1	12	TRONG DÒNG SÔNG CỦA HERACLITUS	
	4	2+1+1+1	6	CÓ THỂ BẠN MUỐN ĐỌC	
	4	2+2+2+2	1	CÔNG TY TNHH MỘT THÀNH	
	5	1+1+1+1+1	10	KHI CẬU GẶP CẬU TA	

6.9 Analysis of Corpus Differences

In the above sections, results have revealed some difference between the three corpora, primarily in the form of content and in the small variation in the presence of UM structures and their relative frequencies. This section features additional analysis of difference between the three corpora.

Appendix U features lists of open class content forms across the three corpora by tiếng size and including their frequencies. As can be seen, many of the same forms at the 1-tiếng level appear in somewhat similar orders; however, these forms vary in their absolute orders and in their frequency of appearance. Occasionally for this list a form appears in one and not in the others, such as for *BÁNH* 'tire, cake, bread, pastry.'

Some interesting variation occurs in the 2-tiếng list beginning with the presence of the form *VIỆT NAM* ‘Vietnam’ for the two diaspora corpora, AUS and US, but this form does not appear at all for the Viet Nam corpus subsample. The two mentioned locations for the Viet Nam corpus are *HÀ NỘI* ‘Hanoi’ and *SÀI GÒN* ‘Saigon,’ the two major cities in the country. There is also variation in pronoun use for the respective corpora. The Australia corpus features *CHÚNG TA* ‘we, inclusive’ and *NGƯỜI TA* ‘people, you, we, everyone’ as the third and fourth most frequent forms and the Viet Nam corpus features *NGƯỜI TA* as the second most frequent form. For the US corpus, the form *NGƯỜI TA* is the thirteenth most common form. The form *CHÚNG TÔI* ‘we exclusive’ appears at rank number eight for the Viet Nam corpus, but at ranks nineteen and twenty for the US and AUS corpora respectively.

There is a range of differences in content and frequencies as well for the 3-, 4- and 5-tiếng lists. These include the term *VỀ VIỆT NAM* ‘return to Viet Nam,’ which appears in the AUS corpus. The broken up term *NGƯỜI MỸ GỐC VIỆT* ‘American of Vietnamese origin’ appears in the 3-tiếng list and in its full form in the 4-tiếng list for the US corpus. Also, it appears that degrees are reported in Celsius in Australia, as evidenced by the form *# ĐỘ C* ‘# degrees Celsius.’ US and Australian entities, groups and cities are named in these corpus’s lists, such as the suburb CABRAMATTA in Australia and reference to *NGƯỜI VIỆT HẢI NGOẠI* ‘overseas Vietnamese’ for the US corpus, while understandably more Vietnamese organizations and locations are mentioned for the Viet Nam corpus. Also, the presence of English is much more frequent for the two diaspora corpora than for the Viet Nam corpus.

Generally, however, once again, other than the presence of whole English forms or code-switches into English, which are more common for the AUS and US corpora, there are no gross differences in the structures of the forms between corpora. Lastly, no form stands out as syntactically wrong as an example of Vietnamese and each form found is a plausible format for a Vietnamese language form, a point of concern when analyzing language in diaspora communities. A few possible explanations are that while inclusion into this study's sample required self-identification as Vietnamese and as a person residing in the particular location required, Australia, the US or Viet Nam, no requirement was made for length of residency or generation status. It is possible that only first generation Vietnamese are likely to blog at length in Vietnamese. Also, the ubiquity of the internet and the possibility of travel between countries may make variation less likely.

To get a closer look at potential function class form and colligational variation, Appendix V includes a list of function forms and their frequencies across the three corpora. As with the other lists, however, this list shows variation in the frequency of most of the same forms across the three corpora. There are a few terms that appear in one, but not in the others, as in the form *AI* 'who' for the Viet Nam corpus and a few more forms in the 4-tiếng list for the VN corpus, but in general, this is a matter of content variation, not necessarily in how the language is structured.

One highly notable exception is noted for the form *BỊ* 'bag, passive marker, negative luck modal verb.' This form also collocates with other tiếng to form 2-tiếng clusters or words, such as *TRANG BỊ* 'to equip', *CHUẨN BỊ* 'to prepare', and *THIẾT BỊ* 'equipment,' among others, as appear in the corpora. This form appears in the top 100

lists for the AUS and US corpora, but not for the VN corpus. The form *BỊ* was found farther down on the VN corpus 1-tiếng list at rank 107. It does, however, appear at a higher token rate than the AUS corpus, but the AUS corpora has a lower tiếng count due to the low word counts for many of the AUS blogs (token rate for *BỊ* by country: 1,425 for US, 963 for VN, and 760 for AUS). As this form when used as a modal indicates negative luck or bad fortune its lower frequency in the VN corpus is very interesting. The form's positive luck modal verb corollary *ĐƯỢC* appears at a similar rank across all three corpora, which makes the variable use of *BỊ* even more interesting. This begs the question of whether this indicates that the Vietnamese population in Viet Nam is generally luckier than its diaspora counterparts.

One last look at potential corpus variation includes the notion of keyness. Using the WordSmith Tools 5.0 KeyWord feature, the key words in each 1-, 2-, 3-, 4-, and 5-tiếng list by corpus were derived through comparison with the respective word lists for the full corpus 1-, 2-, 3- ,4-, and 5-tiếng files. This determines which words in the AUS, US and VN corpora are least commonly related to the full corpus including comments, giving a score called 'keyness.' The greater the keyness score, the less commonly the word is found in the full corpus and the greater the indication of the 'aboutness' of the article, to use a term from Mike Scott's WordSmith Tools help feature (2008). These keyness forms and scores are included in Appendix X.

The keyness forms and scores are truly indicative of variation across the corpora. For content, the AUS corpus includes many terms about organizations in Australia, about places in Australia and even a plethora of cooking terms. As each corpus represents not only an informal network of Vietnamese speakers in a particular location,

but also persons potentially sharing interests, the presence of these words reveal common topics of interest for this particular localized digital social network. Top 1-tiếng cooking words include *BÁNH* ‘cake,’ *BỘT* ‘flour,’ *TRỨNG* ‘egg,’ *KEM* ‘ice cream, *ĐUN* ‘to boil, to heat’ and *HỖN* ‘a bound form meaning to mix.’ Top 1-tiếng location words in include *ÚC* ‘Australia and *BRISBANE* ‘a city in Australia.’

For the US corpus, some interesting 1-tiếng words stand out revealing difference of place and interests such as that not found in subtropical Viet Nam and warm Australia, such as *TUYẾT* ‘snow,’ *LẠNH* ‘cold,’ *LỄ* ‘holiday, celebration,’ *OBAMA*, and *NOEL* ‘Noel, Christmas.’ For the VN corpus, some top forms reveal an interest in the simple pleasures of daily life to include love, vehicles, pathways and place, with forms like *YÊU* ‘love,’ *TRUYỆN* ‘story,’ *CHIẾC* ‘classifier for cars, watches,’ *XE* ‘vehicle,’ *ĐƯỜNG* ‘road,’ *LẶNG* ‘silence,’ *QUÁN* ‘restaurant, shop,’ *ĐỀN* ‘temple,’ and *PHỐ* ‘street.’

These themes continue with the first couple of entries in the 2-tiếng list. The first two forms for the AUS blogs are *CÔNG THỨC* ‘recipe’ and *HỖN HỢP* ’mixture, joint.’ The first key word for the US blog is *HÔM NAY* ‘today’ and the second is *BẢO HIỂM* ‘insurance’ revealing a common concern and a major political topic for American these days. For the Vietnamese corpus, the top 2-tiếng forms are *HÀ NỘI* ‘Hanoi’ and *TIỂU THUYẾT* ‘novel.’ Larger 3-, 4- and 5-tiếng forms also reveal differences among the three corpora to include more forms related to the topics above as well as localized topics between the diaspora.

It should be noted that the effects of bloggers’ voices and the number of times one particular blogger mentions such topics within a country corpus will affect keyness,

regardless of how few other blogs in the corpus mention these topics. The presence of cooking terms and other topics, such as Australian locations and forms referring to photography are indicative of the corpus as a whole, but also may reflect a skewing towards subjects especially for this corpus where the number of words is half of the size of the other two corpora. As such, the forms coming from the larger sized sample blogs will predominate. Furthermore, if any one particular blogger tends to use certain larger size form clusters repeatedly, these will come to the top of the token ranks as the number of recurring forms for the larger tiếng clusters tended to be generally low as compared to the predominance of 1-tiếng forms. To compare, the top 1-tiếng form '#' for the AUS subcorpus had a token frequency of 7,216, whereas the top 5-tiếng form *NỀN CÔNG NGHIỆP VĂN HÓA* had a frequency of 25.

One 1-tiếng form that recurs across the larger sized tiếng clusters for keyness for the two diaspora blogs is *QUYỀN* 'power, authority, rights.' This form does not appear on the 1-tiếng word list at a high enough token rate to appear at the top 100 for any subcorpus, but it does recur for the larger collocational and colligational structures in the word lists and is a notable semantic node for the larger keyness clusters. In fact, the form appears with collocates and within discussed colligational structures in the 5-tiếng keyness lists several times for the AUS blog and within the first form for the US blogs. The form does not appear once, however, for any forms on the VN corpus Top 100 lists. Looking into the complete word lists for the three subcorpora, Table 10 shows the 1-tiếng token rate for *QUYỀN* and the number of collocates with *QUYỀN* on the 2-, 3-, 4-, and 5-tiếng lists. Table 11 provides a list of the forms in which *QUYỀN* appears by corpus in the Top 100 Forms lists, to include the Full, Blogs and Comments corpora.

Table 10: Token Rate and Collocates for QUYỀN

	AUS	US	VN
1-tieng	336	404	181
2-tieng	198	230	154
3-tieng	593	736	380
4-tieng	938	1139	514
5-tieng	1225	1443	604

Collocates for *QUYỀN* in Table 11 include CẦM 'to hold,' as in *NHÀ CẦM QUYỀN* 'authorities, persons holding power.' The three *HÁN-VIỆT* or Sino-Vietnamese bound forms that collocate with *QUYỀN* are *NHÂN* 'human, humanity,' for *NHÂN QUYỀN* 'human rights,' *TAM* for *TAM QUYỀN* 'the three powers, legislative, judicial, and executive,' and *CHÍNH* 'govern, administer' for *CHÍNH QUYỀN* 'political power' The verb form *CÓ* 'to have' together with *QUYỀN* gives CÓ *QUYỀN* 'to have power, to have rights.'

Forms that collocate with the above collocations include *ĐỐI THOẠI VỀ* 'conversation, dialogue about' for *ĐỐI THOẠI VỀ NHÂN QUYỀN* 'dialogue about human rights.' The country for which these rights and authorities collocate with is *VIỆT NAM* which gives us the two phrases *NHÀ CẦM QUYỀN VIỆT NAM* 'Vietnamese authorities' and *NHÂN QUYỀN Ở VIỆT NAM* 'human rights in Viet Nam.' There is also a larger sense of human rights when the form collocates with *QUỐC TẾ* 'international, universal' to give *QUỐC TẾ NHÂN QUYỀN* 'international human rights.' The order for these two 2-tiếng collocates indicates it is not a noun phrase where international modifies human rights specifically, but part of the name of the International Declaration of Human Rights *'Tuyên ngôn Quốc tế Nhân quyền,'* which appears in parts in other tiếng clusters on the lists. Lastly, we have a set of two 2-tiếng collocates for

the form *TAM QUYỀN PHÂN LẬP* ‘separation of powers (legislative, judicial, executive).’ One non-contiguous colligation shows where rights fit in with the one UM structural unit for people *NGƯỜI* gives *MỌI NGƯỜI ĐỀU* ‘every person’ and *CÓ* ‘to have’ for *MỌI NGƯỜI ĐỀU CÓ QUYỀN* ‘every person has rights.’

Collocates with the form *NHÂN QUYỀN* ‘human rights’ give a more specific sense of the associated semantic preferences for the form. *TÌNH TRẠNG NHÂN QUYỀN* and *VẤN ĐỀ NHÂN QUYỀN* both provide a negative attitude or sense for human rights issues, where *TÌNH TRẠNG* means ‘situation, condition’ and *VẤN ĐỀ* means ‘problem, issue.’ Together the reader gets a sense that the human rights situation and the problem of human rights are not seen in a positive light.

Regarding political issues in the Vietnamese diaspora community, the above collocations and semantic preferences reveal the current state for where discussion of politics, especially human rights issues and critique of certain government policies and practices in Viet Nam, may happen in the diaspora, but where such discussion is discouraged within Viet Nam. The form *VIỆT NAM* ‘Viet Nam’ is lower down on the top 100 list for the VN corpus as compared with the other two subcorpora. This difference is also noted in the keyness lists where the forms *CỘNG SẢN* ‘communist, communism,’ *VIỆT NAM* ‘Viet Nam,’ and *DÂN CHỦ* ‘democracy, democratic,’ are at the top of the negative keyness score list for 2-tiếng collocates, indicating that these are topics which the VN corpus expressly does not discuss in comparison with the full corpus. It should be noted that this sample of Vietnamese-language blogs comes only from publicly accessible blog sources. Locked blogs and blog entries and private blogs

may reveal different collocations and preferences regarding any of these discussed forms.

Table 11: *QUYỀN* and Collocates Across Corpora

CORPUS	FORM
AUS Corpus	CẦM QUYỀN VIỆT NAM
	ĐẢNG CẦM QUYỀN LAO ĐỘNG
	ĐỐI THOẠI VỀ NHÂN QUYỀN
	NHÀ CẦM QUYỀN
	NHÀ CẦM QUYỀN VIỆT
	NHÀ CẦM QUYỀN VIỆT NAM
	NHÂN QUYỀN Ở
	NHÂN QUYỀN Ở VIỆT
	NHÂN QUYỀN Ở VIỆT NAM
	QUYỀN Ở VIỆT NAM
	THOẠI VỀ NHÂN QUYỀN
	THOẠI VỀ NHÂN QUYỀN VỚI
	TÌNH TRẠNG NHÂN QUYỀN
	TÌNH TRẠNG NHÂN QUYỀN Ở
	TRẠNG NHÂN QUYỀN Ở
	TRẠNG NHÂN QUYỀN Ở VIỆT
	VẤN ĐỀ NHÂN QUYỀN
	VỀ NHÂN QUYỀN
	VỀ NHÂN QUYỀN VỚI
	VỀ TÌNH TRẠNG NHÂN QUYỀN
BLOG	CẦM QUYỀN VIỆT NAM
	ĐỐI THOẠI VỀ NHÂN QUYỀN
	MỌI NGƯỜI ĐỀU CÓ QUYỀN
	NHÀ CẦM QUYỀN VIỆT NAM
	NHÂN QUYỀN Ở VIỆT NAM
COMMENTS	CHÍNH QUYỀN
	TAM QUYỀN PHÂN LẬP
FULL	NHÀ CẦM QUYỀN VIỆT NAM
	NHÂN QUYỀN Ở VIỆT NAM
	TAM QUYỀN PHÂN LẬP
US	MỌI NGƯỜI ĐỀU CÓ QUYỀN
	NGÔN QUỐC TẾ NHÂN QUYỀN
	NGƯỜI ĐỀU CÓ QUYỀN
	NGƯỜI ĐỀU CÓ QUYỀN ĐƯỢC
	NHÂN QUYỀN TẠI VIỆT NAM
	QUỐC TẾ NHÂN QUYỀN

6.10 Conclusion

In comparison with analyses in Chapters 4 and 5, this analysis provided a description of the full corpus for this study as divided by the three originating geographic sources, Viet Nam, the US and Australia. As seen in Chapters 4 and 5, results indicate that much of the content between the three subcorpora are similar types. Results also indicate the much higher frequency rate for unique forms for the 1- tiếng lists, with lower initial token rates and more sharply declining frequency rates for the subsequent form iterations. The prevalence of 1-frequency forms is highest as the number of tiếng in the forms increases, with the 3-, 4- and 5- tiếng forms lists being comprised of up to and over 90% 1-frequency forms. This indicates that 1- and 2- tiếng are relatively more commonly used and represent more typical language patterns than 3-, 4- and 5- tiếng forms.

A-curve charts further confirm the ubiquity of the inverse relationship between rank and frequency of forms, no matter the size of the tiếng cluster, and no matter the geographic origin of the blog, supporting assertions of this sample of Vietnamese blogs as representing “speech as a complex system” (Kretzschmar, 2009, p. 159).

Analysis according to the three dimensions explored above indicate many differences depending on the size of the form and the forms therein. In the structural analysis, findings show that there is an almost one-to-one ration between tiếng and morpheme, with the exception of a few bound forms. Also, while the smaller tiếng forms include primarily one- and two-tiếng words, with a few words appearing in the 3-tiếng forms as well, larger forms, to include two-tiếng forms, feature words within larger phrases, clauses and sentences. It is firmly at the 3-tiếng form level where the structure of the forms switches from words to these larger constituents, giving support for

Vietnamese as more than a monosyllabic language, while also showing that word forms primarily tend to range in the 1- and 2-tiếng size for these particular subsamples of the larger corpus.

In the form content analysis, we again see that most form content is free, with a few forms considered to be bound types. For open/closed class forms and content/lexical class forms, patterns vary by form size. For 1-tiếng forms, numbers are closer to even between categorization by class, but as the tiếng size increases the types of forms found within increase to include more content and open class forms. There are very few entirely closed and lexical class 4- and 5-tiếng forms.

For the units of meaning analysis, we see that by virtue of defining units of meaning above the level of the morpheme, there is then a count for UM that corresponds most closely to counts for words for collocational analysis. For colligational analysis, forms extended beyond individual words as one colligational structure of note for Vietnamese may include several words which surround content morphemes and must include the range of forms to enact the structural form.

Distributional analysis of UM reveals a variety of possible structures, with more structural forms appearing as the number of tiếng increases. 1-tiếng forms in the list naturally remain at the one UM stage, while differing clusters of UM within larger forms create many different structural distribution patterns. Also, depending on size and type of form, different patterns emerge, with most categorization and description of units as larger than a morpheme resulting in a primarily word and phrase level analysis. Other than the example just given, there are virtually no 4- and 5-tiếng full form UM, which mirrors patterns for words, but only when including phrases and idioms.

Analysis of difference between the three corpora reveals that while content and the frequency of content varies, larger structural differences do not seem to exist. The most notable exception is the absence of the negative luck modal verb form *BỊ* as noted by the function form lists in Appendix V. Keyness lists show the aboutness particular to the three corpora, with mentions of local places, organizations, political concerns and the particulars of daily life in the three locations. A deeper look at one particular 1-tiếng form *QUYỀN*
'authority, rights,' gives insight into the current state of some aspects of political discussion in the sampled blogosphere, especially as regards topics that are and aren't discussed for the particular country corpora.

CHAPTER 7

CONCLUSION

Chapters 1, 2, and 3 in this dissertation discuss demographics, language trends and migration issues regarding the Vietnamese population in Viet Nam and of Vietnamese-speaking people in the diaspora. Despite disparate total population numbers for their countries of residence, Vietnamese-Americans and Vietnamese-Australians who claim Vietnamese ethnic ancestry alone represent 0.6 percent of their respective nations' populations. Continuous immigration, social and digital network participation and attitudes, and behavior in the home and community have supported the vibrancy of the Vietnamese language in the diaspora, while Vietnamese continues to grow and change in Viet Nam. Research regarding heritage language maintenance indicates that despite positive attitudes towards language maintenance, language behavior is not always conducive for such maintenance. Despite research proposing changes in language patterns and language behaviors among Vietnamese language speakers in-country and in the diaspora, there is little research using language in use data and using corpus analysis methods that investigate varieties of Vietnamese as used today.

This dissertation began in Chapter 1 seeking to answer the following questions:

1. *What are the most common syllable forms and collocational, colligational, and topical patterns as revealed in a corpus of in-country and US and Australian Vietnamese language blogs?*

2. *What are the most common syllable forms and collocational, colligational and topical patterns as revealed in a comparison between in-country and US and Australian Vietnamese language blogs?*

3. *In corpus analysis of Vietnamese language varieties, what are the implications for how we analyze data? What are the implications for existing theory concerning segmentation into meaningful units in Vietnamese? How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?*

Questions 1 and 2 above were explored through analysis of a corpus of informal web-based Vietnamese divided in Chapters 4 and 5 between the full corpus and the blogs and comments corpora and in Chapter 6 for the three by-country corpora. These chapters found that patterns for these varieties of the Vietnamese language adhered to patterns for other studied languages. By using a corpus analytic approach and using techniques for assessing the relationship between rank and frequency for Vietnamese, A-curve patterns reveal this sample of informal internet Vietnamese to be an example of a complex speech system. When parsing the full blogs corpus by country as in Chapter 6, the A-curve pattern obtained in a similar manner, showing that “frequency of

frequencies" (Kretzschmar 2009) holds in samples for the Vietnamese, Australian, and US blog varieties collected.

Specifically, Chapter 4 discusses patterns for the Vietnamese language variety as found in a series of intersecting social media networks. Word lists and A-curve charts show that for the corpus as a whole and as divided into subcorpora, i.e., blogs and comments, Zipf's Law regarding the inverse relationship between a language type's rank and its frequency in a list of types holds for these sampled varieties of Vietnamese. This finding connects research into Vietnamese with analyses of many languages of the world where the above A-curve shows that inverse relationship. Restating Kretzschmar, Zipf's law "extends to experimental data from survey research as well as to words in texts, and thus it stands as a primary characteristic of speech as a complex system" (2009, p. 159).

In addition, for this corpus, analysis shows that for this sample of the Vietnamese language, the distribution of forms according to the A-curve holds not only for 1- tiếng lists, but also for each the 2-, 3-, 4- and 5- tiếng lists. Moreover, when considering 1-tiếng forms lists, it is important to remember that the high token rates for these forms includes both 1-tiếng forms as single meaningful units in themselves, but also counts their use as part of larger 2- through 5-tiếng forms. As such, the distribution of 1-tiếng forms as part of larger clusters also shows this inverse relationship between rank and frequency and further reinforces the notion of the Vietnamese language as a complex system. For example, when looking at Appendix B, the form *CÓ* appears in the Full Corpus 1-tiếng list as the second most common form with a token rate of 84,069. This form also appears in the 2-tiếng list as part of the collocation *CÓ THỂ* 'could, possibly,'

as part of the colligation *KHÔNG CÓ* ‘not have,’ and as part of the colligation *CÓ MỘT* ‘have one.’ These three 2-tiếng examples occur in descending rank according to their token rate (8,151; 6,809; and 3,679 respectively), rather than appearing at the same rank for each 2-tiếng form, showing that 1-tiếng forms themselves and 1-tiếng forms as they appear as part of larger clusters also adhere to Zipf’s law. This pattern obtains for all a-curve distributions no matter how the corpus is subdivided or how small or large the clusters.

Chapter 5 analyzes the full corpus and the two subcorpora, blogs and comments, using various approaches, by structure and by content and/or function, open class and/or closed class and free and/or bound forms approaches. Patterns emerged relating to the number of tiếng in the form, polysemous forms and that while most forms could be described as firmly on one side of the three axes, content/function, free/bound and open/closed, a small number of forms, like classifiers, nouns, and pronouns could be described from several different perspectives. Exploration of the pre-defined notion of content units of meaning larger than the morpheme level indicated alignment with the notion of word, but also extending into the phrasal level. When considered from a colligational standpoint, there were several units of meaning that included a series of contiguous and non-contiguous tiếng that together combined to create grammatical structures unique for Vietnamese.

Further in-depth study of the distribution of units according to structure, content analysis, and units of meaning analyses revealed that the full corpus, the blogs and comments corpora, and the three corpora as divided by country showed similar patterning across analyses. Discussion in Chapters 5 and 6 indicate that while there is

almost a one-to-one relationship between tiếng and morphemes, full form words are most common for the 1- and 2-tiếng forms, with few full words for the 3-tiếng category, and virtually none for the 4- and 5-tiếng categories. Despite the fact that the literature attests such forms, only one appeared in the top 100 lists that were the focus of this dissertation (Nhan 1984, Schiering & Bickel 2007, Thompson 1963). These results confirm assertions that Vietnamese is not a true monosyllabic language, despite the way the language is written and despite the predominance of 1-tiếng word forms for the sampled corpus.

Larger forms especially for the 3-, 4-, and 5-tiếng form levels included a variety of structural forms above the level of the word, to include phrases, clauses, and sentences. At the 2-tiếng level, there was an equal mix between 2-tiếng full word forms and clauses. At the 3-tiếng level, there are only a few attested full form words and colligational structures predominate. At the 4- and 5-tiếng level, content is interspersed with grammar patterns within the discussed phrases, clauses, and sentences.

Content level analysis by the three axes free/bound, content/lexical, and open/closed class also showed similar patterns for each of the corpora, the full corpus or its various subcorpora, no matter how they were subdivided. For 1-tiếng forms, there was almost equal distribution between content and lexical forms and open and closed class forms, with more forms being characterized as content and open class the larger the forms got. This reflects the fact that larger forms tend to be larger discursive units with content forms appearing throughout, which precluded characterization for most large forms as either lexical or closed class. Very few forms were found to be bound

forms regardless of form size, with most being forms of Sino-Vietnamese origin, which are restricted in their use as well as one or two of the available pluralizing forms.

For analysis based on the concept of units of meaning (UM), the starting definition of ‘units larger than the morpheme’ aligned content UM units of meaning determinations as closest in pattern to the syntactic word. In a few instances, phrases and idioms were included as one unit of meaning. And, again, despite characterization of a few phrases and multiple digit numbers as one unit of meaning, very few 3, 4-, and 5-tiếng full form units of meaning were found.

For each of the discussed analyses, the similarity in patterns and distributions across corpora reveal that differences between the corpora include some variation and some similarity. Similarities include those mentioned above, in that results for each of the subcorpora include many of the same unit of meaning (UM) structures, even though these structures contain different content elements and appear at varying frequencies across the respective corpora. In addition, many of the same content forms and lexical patterns appear, albeit at differing frequency.

The main differences in the corpora occur in the content areas. Topics discussed vary based on corpus country and preferred blogger topics within the blogs as social networks. Common topics across the corpora are the self, family, activities and hobbies like reading and photography, work, shopping, the weather, poetry, music, cooking, holidays-most notably Tết and Christmas, and the Vietnamese and English languages.

Blog content from Viet Nam tended to be more specific with local places names, perhaps because of an expectation that all readers capable of reading Vietnamese

would be familiar with the places mentioned, to include specific mention of street names and stores, whereas for diaspora readers, specific localized place names tended not to appear in the top 100 lists. A few other notable differences do occur between the diaspora and Viet Nam corpora, as for the absence of the bad fortune modal verb *BỊ* for the Viet Nam corpus (top 100 list) and regarding the 1-tiếng form *QUYỀN* and its collocates and semantic preferences. Also, while political topics and discussion about returning to Viet Nam and Vietnamese identity in particular are mentioned in the US and AUS corpora, virtually no mention of these topics occurs in the Viet Nam corpus. Explanations for this are possible pressure within Viet Nam against speaking about politics as well as the need for persons outside Viet Nam living in multicultural contexts to understand and discuss 'Vietnamese-ness' and what it means to be Vietnamese and to prepare for travel back home to visit family and friends.

More generally, however, when considering function forms, a few numbers may differ and the frequency of a few items may be higher or lower, but in general, the top forms tend to be more similar than less. This conforms with Kretzschmar's (2009) assertion that differences in language might be considered as a variation in degree of use rather than an overall absence of forms in one variety and their sole existence in another. When considering content, there is more wide variation as the topics discussed, the locations mentioned, and the social network practices and topics show a lot of difference.

7.2 Question 3

Using data and analyses from Chapters 4 through 6, this section will discuss the various questions raised in Question 3. The methods used for this dissertation included using a multi-tiếng unit approach. Any study that seeks to analyze Vietnamese must take into account the variable size of potential forms, regardless of whether the focus is on grammar or lexis. As the literature reviewed in Chapters 2 and 3 suggest, lexical forms can be from 1- up to several tiếng.

The correspondence between morphemes and words do lend support to convictions that Vietnamese is a monosyllabic language, however the wide variety of 2-, 3-, and 4-tiếng forms as presented lend strong evidence against that conclusion. 2-tiếng forms include 2-tiếng full form verb and noun compounds, 2-tiếng adverbs and adjectives, unanalyzable 2-tiếng/1-morpheme borrowings, 2- and 4-tiếng reduplicatives, and attested forms which include the classifier/measure noun with the main noun/verb form for 2- and 3-tiếng words. Based on this evidence, Vietnamese should not be considered a monosyllabic language.

Grammatical forms and patterns also stretch beyond the single unit, to include 4- and 5-tiếng forms, despite the fact that grammatical forms as such are not usually included in dictionary head entries. And, depending on the chosen focus, whether structural or by a units of meaning approach, researchers should be clear about what exactly they mean when they talk about how the language works overall. A focus on tiếng, morphemes, or words alone will not reveal the larger topical patterns and various collocates that individual or even larger tiếng forms keep and the colligational patterns that individual tiếng come together to form.

One of the biggest implications as mentioned in the previous paragraph is to be certain as to the focus of a study and which size of unit or, indeed, range of size of units is appropriate. As is clearly shown, no study of actual Vietnamese discourse should limit itself to single tiếng units unless that is the unit for analysis and the limitations regarding discourse are taken into account. Furthermore, study that limits itself to 1- and 2-tiếng units would also miss common lexical and grammatical patterns above that size. Of course, depending on how the term 'units of meaning' is defined will lead the choice of unit size. Any study looking at the Vietnamese phrase would necessarily have to look at units larger than 1- or 2-tiếng forms, despite the fact that some smaller phrases may be only 2-tiếng in size. Table 1 below shows some of the range of possible collocational types and patterns, including both monomorphemic and polymorphemic types.

One issue of segmentation is whether the classifier should be included as part of a meaningful unit for Vietnamese. As seen in Chapters 5 and 6, popular use and even dictionaries tend to pair the classifier/measure noun in with the main form that the classifier/measure noun indexes. Semi-affixes called bán phụ tố also occupy the classifier/measure noun slot in creating new forms, but dictionaries as yet do not include the full range of possible forms that these combinations could obtain.

Table 1: Tiếng and Collocational Patterns

Tiếng	Monomorphemic	Polymorphemic
1-Tiếng	*là* 'be' *có* 'have' *sớm* 'early', *đi* 'go'	
2-Tiếng	*Sài Gòn* 'Saigon' *cà phê* 'coffee' *cao bồi* 'cowboy' *vui vẻ* 'happy'	*bao giờ* 'when' *chế độ* 'regime' *Giáng Sinh* 'Christmas' *trở thành* 'become' *hì hì* 'laughter'
3-Tiếng		*ngôn ngữ học* 'linguistics' *chính trị gia* 'politician' *Hồ Chí Minh* 'Ho Chi Minh'
4-Tiếng		*chua chua ngọt ngọt* 'very sweet sour' *trung tâm thành phố* 'city center'
5-Tiếng		*Á châu Thái Bình Dương* 'Asia Pacific Ocean' *Chủ tịch Hồ Chí Minh* 'Chairman Ho Chi Minh'

For a reminder example, the term *CÂU TRẢ LỜI* 'answer' includes the 2-tiếng form *TRẢ LỜI* 'to answer', which is a verb consisting of two morphemes *TRẢ* 'to return, give back' and *LỜI* 'spoken utterance'. The 1-tiếng form *CÂU* 'sentence, line' acts as a classifying affix here; however, *CÂU* is itself a full 1-tiếng noun form. It is clear to native speakers of Vietnamese and even elementary learners of Vietnamese that the full 3-tiếng form *CÂU TRẢ LỜI* means '(an) answer' in spoken discourse. It also extends even to written discourse as evidenced from its inclusion in this corpus of informal written internet Vietnamese. Dictionaries, though, vary in their inclusion of this full form. The verb exists as a 2-tiếng form in each of the consulted dictionaries (Nguyen 1963, Ban Biên Soạn Chuyên Từ Điển: New Era 2001, Bùi Phụng 2003, Ban Biên Soạn Chuyên Từ Điển: New Era 2005, Ban Biên Soạn Từ Điển Ngọc-Xuân-Quỳnh 2006, Viện Khoa Học Xã Hội Việt Nam: Viện Ngôn Ngữ Học 2009). The 3-tiếng noun form for

‘answer’, *CÂU TRẢ LỜI*, however appears only in two of the referenced dictionaries, once in the Bùi Phụng (2003) dictionary as a head word and once in the example given for the English entry ‘answer’ (Viện Khoa Học Xã Hội Việt Nam: Viện Ngôn Ngữ Học 2009). This leads to the conclusion that affirmations of the entire 3-tiếng form as one definite word is not a given for all dictionaries and for the language in general.

Interestingly, in the second example above, the head definition for ‘answer’ as a noun is given as *sự trả lời* ‘answer’ where *sự* ‘classifier for events, things or actions’ is the semi-affix that derives nouns from verbs. The head definition is followed in the example with *câu trả lời* as in "*Câu trả lời nó đưa ra thật đáng kinh ngạc*" “‘The answer he gave was quite surprising’” (2009, p. 123). Here, because the main form is a verb, in order to derive a noun, the classifier/measure noun slot must be filled with a form that indexes nouns. While it may be possible once context has been established to refer to a poem *bài thơ* as a *bài* ‘classifier for written papers’ or even possibly as a *thơ* ‘poem, poetry,’ one cannot refer to an answer as a thing with only the verb form. As such, the classifier/measure noun is required for specificity in the case of *CÂU TRẢ LỜI* ‘answer.’

This study did not require that Vietnamese conform to English rules of morpho-syntax, where affixes are necessarily bound and any form acting as an affix would also be considered to be bound. From the above examples, *sự* is considered a bound form and so if it had appeared as a collocate with *trả lời* to render *sự trả lời*, then assertions of the form being one structural word have more support. In this context; however, the 3-tiếng form appeared as c*âu trả lời*, where *câu* is not a bound form. As such, while the function of the form is obvious as a noun, it is not required to view this as a one-word form.

On the other hand, however, the consulted dictionaries did differ in just this respect. Some, the Bui dictionary especially (2003), included many of these collocations into colligational structural combinations. The Bui dictionary is a Vietnamese to English dictionary, however, so perhaps there is a tendency for maximal representation of common forms in the language regardless of the prescriptive syntactic rules for Vietnamese. This becomes especially important when adding numbers to the collocation, as Vietnamese prescriptively requires the classifier/measure noun slot be filled in noun phrases when counting forms, as in *hai câu trả lời* 'two answers.'

If affixes and classifiers can create innumerable combinations of content forms, then perhaps no dictionary, even an electronic one created from corpora of language in use data, would ever be able to include all possible forms. However, one must take these issues into account when working with the language. Vietnamese is simply distributed differently than English. For unit of meaning analysis, though, the full form is clearly one concept and as such was considered to be one unit.

The last question asks, “How does the pattern of segmentation as used confirm or challenge existing research and theory regarding the units of meaning for Vietnamese discourse generally, especially when taking varieties in-country as well as in the diaspora into account?” Existing research is very well focused on just this question, especially for the Corpus Linguistics (CL) and Natural Language Processing (NLP) fields. This study used corpus language processing software, namely WordSmith Tools, version 5 (Scott, 2008) to analyze the data. Individual tiếng forms lists and clusters lists from 2- to 5-tiếng were chosen as the units of analysis because the literature as mentioned above attested to the fact that single units of meaning, here

words, existed in these tiếng-sized ranges. This follows Nguyen (1984) as regards the different ways that the Vietnamese syllable may be compared to the concept of word, with various classifications, such as the orthographic word, the dictionary word, the grammatical word, and the phonetic or prosodic word. While orthography supports the notion of the orthographic word, dictionaries and grammatical conventions vary widely. Considering the language from a 'units of meaning' perspective allowed exploration of meaning apart from whether and how orthography or other perceptual and psycholinguistic characterizations of the ways syllables collocate and colligate are applied.

It should be noted, however, that this study was not focused on translation, per se. In the NLP literature, discussion centers around automatic processing and automatic translation software which attempts to make one-to-one unit translations. For this analysis, no attempt was made to force Vietnamese units into a one-to-one correspondence with any other language forms. It is linguists working with Vietnamese and in Viet Nam and in the diaspora themselves who have most addressed this question. This study has attempted to add to this discussion, especially as regards discussion of units larger than the proposed monosyllabic forms attested in some Western venues.

When it came time to determine exactly how big a unit 'should' be for this study, an arbitrary decision was made to focus on certain predefined categories, such as words, phrases, full form words, free and bound classes, etc. As such, forms were assessed as to whether they might be used at the varying form sizes chosen. This allowed much freedom in understanding exactly how the sampled language varieties

work and how the 1- through 5-tiếng forms may be understood from the relevant structural, content analysis, and units of meaning approaches. As noted before, as units increased in size, the varying types of forms contained within them tended to prevent inclusion into the more restricted bound, lexical, and closed categories. Table 2 below shows some of the range of colligational types and patterns, including classifier/main form combinations as well as structural and other grammatical forms.

Table 2: Colligational Patterns

Tiếng	**Colligations: Units as Words?**	**Colligations**
1-Tiếng		
2-Tiếng	*bài thơ* ‘poem,’ *bài viết* ‘essay,’ *cuốn sách* ‘book’ *nhà văn* 'writer‘ *rất là* 'very (stative verb/adj)'	*ai cũng* ‘everyone’ *chụp hình* 'take a photo‘ *của mình* 'of me/us‘ *không phải* 'must not' *càng…càng* ‘more…more'
3-Tiếng	*bài phát biểu* ‘(a) speech’ *cuốn tiểu thuyết* 'novel' *nền kinh tế* 'economy‘ *người đàn ông* 'man'	*công thức này* 'this recipe‘ *đã trở thành* 'had become‘ *# quả trứng* ‘# eggs’ *lúc nào cũng* 'whenever' *xe food truck* 'food truck'
4-Tiếng	*xã hội chủ nghĩa* 'socialism‘ *cha truyền con nối* 'hereditary'	*các cuộc đối thoại* 'all/dialogues‘ *Chúng ta có thể* ‘We can.‘ *Đây là lần đầu* ‘This is first time.‘ *hơn bao giờ hết* 'more than ever'
5-Tiếng		*cái xấu của hình đẹp* 'the ugliness of a beautiful picture‘ *mọi người đều có quyền* 'everyone has rights'

In order to understand the data as used in context as well as from a more prescriptive tradition, dictionaries were used as mentioned in Chapter 3. The variable methods used for each of the dictionaries did not necessarily assist in final

determinations along pre-defined structural and content approaches, so the remaining unit of meaning analysis allowed much more relaxed assessment of units.

Determinations for other forms were also variable. Full names were considered to be three words, with the traditional Vietnamese word order of last name, middle name and first name. As discussed in Chapters 5 and 6, despite the fact that names represent one 'thing', the two raters varied in their assessment of the units of meaning for names. For the second rater, a name is a special case of three words or units which cohere to make one unit of meaning with three individual parts (1+1+1), rather than one unit of meaning with three parts as in (3), as would be noted for the 3-tiếng, full form noun, one unit of meaning *NGHIÊN CỨU SINH* 'student researcher' and as the first rater categorized them.

Other novel forms, such as interrogatives and discourse markers such as laughter forms, *HA HA HA, HE HE HE, HI HI HI, HÍ HÍ HÍ*, and *HÌ HÌ HÌ*, did not occur in any dictionaries, but are obvious from context and experience to be what they are. Each of these 3-tiếng forms was considered to be made up of three morphemes each, but to be one word and one unit of meaning. It is through reduplication that the repeating series of sounds become what they are; however, each individual part was considered to mean a laugh morpheme singly as well by both raters.

Lastly, it should be noted that the use of an unannotated, untagged raw corpus, full of the variety of alternate and unique forms as noted in Chapters 3 and 6, includes much data that has yet to be addressed in this study. This dissertation focused on overall descriptive statistics for the corpora and only on top forms for detailed analysis.

Future studies of this corpus will be able to home in on particular forms for analysis and description, but these are out of the scope of possibility for this study.

7.3 Implications and Discussion

As a study of Vietnamese language variation in in-country and diasporic contexts and an exploration of the appropriate unit of analysis for linguistic study of the Vietnamese language, this new research provides insight into several key issues in contemporary language corpus study. Despite the somewhat limited prevalence of English language research into the Vietnamese language, there are few studies that actually examine the language using in-use data. This study sought to fill these gaps.

The expanding body of work that is beginning to appear on the web and in regional publications tends to focus on corpus linguistics (CL) and natural language processing (NLP) related issues. Studies in these areas use corpus methods to investigate linguistic phenomena, describe methods for corpus creation, and examine the best natural language processing techniques for Vietnamese, although most of them are focused on the word as the proper unit of segmentation for Vietnamese. This dissertation departed from this method in order to assess how units of meaning may be understood from varying perspectives, namely structural, content, and a somewhat lightly defined general 'units of meaning' perspective.

The tiếng and subsequent iterations of the form up to 5-tiếng were the initial forms used for corpus processing, but subsequent analysis expanded that notion to include smaller units such as the morpheme and bigger units up to the sentence level. While spaces are used to separate syllables, it should not be assumed, therefore, that

the tiếng is also the base unit of meaning for any and every analysis of the language. Linguistic studies of the Vietnamese language should take careful note of the questions they are trying to answer and seek to use the appropriate unit of analysis to answer that question. The use of corpus data for language research and especially for teaching should also take this concept into account.

One caveat here recognizes the limitations inherent in studying unedited forms of language and with such a small sample. Six million words may sound like a lot of language, but when one considers the billion words that a person will speak and encounter in a lifetime (Kretzschmar 2009), six million is quite small. Also, this sample should not be considered representative of the language as a whole, but rather one small subset of a variety of Vietnamese as found in an intersecting series of internet blogs.

Lastly, the full range of variation that may be found in the corpus as a whole was likely not represented by choosing to concentrate on only the top forms. The hundreds of thousands of forms in 1-frequency tails of the data may include forms revealing more difference between the corpora as well as innovations that have not made their way into the more commonly used language patterns. Future studies using this data may discover these patterns.

7.4 Conclusion

There is a growing body of research concerning corpus techniques as applied to the Vietnamese language. Corpus techniques that segment the language according to cluster size and number of syllables affirm Vietnamese distribution patterns according to

Zipf's law and the frequency of frequencies. No matter how large the cluster size and no matter whether the clusters consist of collocations or colligations, the inverse relationship between rank and frequency for all forms shown holds.

This study of language in use data also showed that while there is variation among varieties of Vietnamese, this variation is primarily confined to varying degrees of use for function forms and colligational patterns, while the bulk of variation occurs for content forms. Variation is then both a matter of degrees of use and in the topics discussed. The expectation that the varieties explored would have begun to qualitatively differentiate, with widely varying grammar and in other substantive ways for lexis, at least for the top forms examined is not supported.

While seminal studies of Vietnamese began with the question of how and whether tiếng corresponded to *từ* 'word' even for 1-tiếng forms, this study explores units of meaning and whether units larger than 1-tiếng might also be considered words. Based on the evidence, this study affirms the Vietnamese syllable, tiếng, and union of syllables, *tổ hợp tiếng (*Nguyen, 1984) as the primary distributional units for the Vietnamese language. Despite the conclusion that Vietnamese is clearly not a strictly monosyllabic language, the union of tiếng as described throughout this dissertation, however, may not always fit applied definitions of word, especially as regards the boundaries between collocation and colligation and the inherent flexibility in such unions where reversibility, interruptibility and orthographic conventions permit patterns less common to languages where the word is the primary distributional unit.

Use of the term ‘considered’ above regarding the difference between syllables and words is particularly relevant. Specifically regarding discussion of the relationship between *từ* and tiếng in Vietnamese, Nguyen Thien Giap (1984) writes:

> *“Trước hết cần lưu ý rằng 'từ' không phải chỉ là đơn vị ngôn ngữ học thuần tuý mà còn là đơn vị tâm lý-ngôn ngữ học. Những quan niệm truyền thống về từ rõ ràng đã dựa vào tính hiển nhiên về mặt tâm lý của nó. F.de Saussure viết, "...từ là một đơn vị luôn luôn ám ảnh tư tưởng chúng ta như một cái gì đó trung tâm trong toàn bộ cơ cấu ngôn ngữ học mặc dù khai niệm này khó định nghĩa"” (p. 61).*

> “First of all it should be noted that 'word' is not only a purely linguistic unit but also a psycho-linguistic unit. Traditional concepts about word are clearly based on the psychological face of the evidence. F.de Saussure writes, "... word is a unit that always haunts our thoughts as something central to the whole of linguistic structure, although the concept is difficult to define"” (p. 61).

The question of meaningful units for the Vietnamese language remains a complicated one. And now, it is another linguist’s turn to be haunted.

REFERENCES

Androutsopoulos, J. (2011). From variation to heteroglossia in the study of computer-mediated disourse. In C. Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the New Media* (pp. 277-298). New York: Oxford University Press.

Australian Bureau of Statistics. (2013). *Australian demographic statistics*, Sep 2012. Accessed April 8, 2013 at http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0

Ban Biên Soạn Chuyên Từ Điển: New Era. (2001). *Từ Điển Anh-Việt Việt-Anh: English-Vietnamese Vietnamese-English Dictionary*. Hà Nội: Nhà Xuất Bản Văn Hoá Thông Tin.

Ban Biên Soạn Chuyên Từ Điển: New Era. (2005). *Từ Điển Tiếng Việt*. Hà Nội: Nhà Xuất Bản Văn Hoá Thông Tin.

Ban Biên Soạn Từ Điển Ngọc-Xuân-Quỳnh. (2006). *Từ điển tiếng việt*. Hà Nôi: Nhà Xuất Bản Từ Điển Bách Khoa.

Ben-Moshe, D., & Pyke, J. (2012). The vietnamese diaspora in australia: Current and potential links with the homeland. In *Report of an Australian Research Council Linkage Project*. Deakin University, Australia: Centre for Citizenship and

Globalisation. Accessed January 12, 2013 at http://hdl.handle.net/10536/DRO/DU:30047381

Binh, N.N. (2003). *Elementary vietnamese: Revised edition*. Singapore: Tuttle Publishing.

Bùi Phụng. (2003). *Từ Điển việt anh: Vietnamese-english dictionary*. Hà Nội: Nhà Xuất Bản Thế Giới.

Cao, X.H. (1985). Về cương vị ngôn ngữ học của <tiếng>. *Ngôn Ngữ*, 2, 25-69.

Collins, J., Slembrouk, S, & Baynham, M. (2009). Introduction: scale, migration and communicative practice. In J. Collins, S. Slembrouk & M. Baynham, (Eds.), *Globalization and Language in Contact: Scale, Migration and Communicative Practices* (pp.1-18). London: Continuum International Publishing Group.

Dao, H. T. (2011). Vấn đề phân tích tự động thuật ngữ trong khối liệu ngôn ngữ tiếng việt. *Từ điển học và Bách khoa thư*, 4 (12), 34-38.

de Saussure, F. (1983). *Course in general linguistics*. Illinois: Open Court Publishing.

Dinh, D. & Hoang, K. (2003). POS-Tagger for english-vietnamese bilingual corpus. Proceedings from *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond* (pp. 88-95). Edmonton, Alberta.

Dinh, D., Hoang, K., & Nguyen, V.T. (2001). Vietnamese word segmentation. Proceedings from *The 6th NLPRS* (pp. 749-756). Tokyo, Japan.

Dinh, Q.T., Le, H.P., Nguyen, T.M.H., Nguyen, C.T., Rossignol, M., Vu, X.L. (2008). Word segmentation of vietnamese texts: A comparison of approaches. Proceedings from *LREC* (pp 245-265).

Hieu, L. T., Vu, L. A., Kien. L. T. (2010). An unsupervised learning and s statistical approach for vietnamese word recognition and segmentation. Proceedings from *ACIIDS*, 2, (pp. 195-204).

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). Corpus linguistics and the web. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 1-6). New York: Rodopi.

IBM Corp. (2012). *IBM SPSS Statistics for Windows*, *Version 21.0*. Armonk, NY: IBM Corp.

Kretzschmar, W.A. (2009). *The linguistics of speech*. Cambridge: Cambridge University Press.

Le, H.A. (2003). A method for word segmentation in vietnamese. Proceedings from *The International Conference on Corpus Linguistics.* Lancaster, UK.

Le, H. T & Tran, V.N. (2002). *Từ điển từ hán việt: Sách giao khoa phổ thông: Tiếng việt và văn học*. Thành phố Hồ Chí Minh, VN: Nhà Xuất Bản Đại Học Quốc Gia:.

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

McEnery, T. & Ostler, N. (2000). A new agenda for corpus linguistics-Working with all of the world's languages. *Literary and Linguistic Computing*, 15 (4), 403-419.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge.

Milroy, L., & Li, W. (1995). A social network approach to code-switching: The example of a bilingual community in britain. In L. Milroy & P Muyksen (Eds.), *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching* (pp. 136-157). Cambridge: Cambridge University Press.

Ngo, B.N. & Hoai, B.T. (2001). The vietnamese language learning framework. *Journal of Southeast Asian Language Teaching*, X, 1-24.

Nguyen, C.T., Nguyen, T.K., Phan, X.H., Nguyen, L.M., Ha, Q.T. (2009). Vietnamese word segmentation with crfs and svms: An investigation. Proceedings from *20th PACLIC* (pp. 215-222). Wuhan, China.

Nguyen, D. (2009). Using search engine to construct a scalable corpus for vietnamese lexical development for word segmentation. Proceedings from *7th Wordshop on Asian Language Resources* (pp. 171-178). Singapore: Suntec.

Nguyen, D.H. (1967). *Vietnamese-English Student's Dictionary*. Saigon, VN: Vietnamese American Association.

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

Nguyen, T.G. (1984). Về mối quan hệ giữa từ và tiếng trong việt ngữ. *Ngôn ngữ*, 3, 60-69.

Nguyen thi P.T. (2010). *Personal Interview*. July 15, 2010.

Nguyen, T.M.H, Romary, L., Rossignol, M., & Vu, X.L. (2006). A lexicon for vietnamese language processing. *Language Resources & Evaluation*, 40, 291-309.

Nguyen, T.V., Tran, H.K., Nguyen, T.T.T., & Nguyen, H. (2006). Word segmentation for vietnamese text categorization: An online corpus approach. Proceedings from *Research, Innovation and Vision for the Future: The 4th International Conference on Computer Sciences*.

Nhàn, Ngo Thanh (1984). *The syllabeme and patterns of word formation in Vietnamese*. (Doctoral Dissertation). Retrieved from Dissertation Abstracts International.

O'Leary, C.F. (1989). *Language maintenance and shift in a vietnamese refugee community: A study of attitudes and behaviors*, Vol I, II. (Doctoral Dissertation). Retrieved from Dissertation Abstracts International.

Pham, D. D., Tran, G. B., & Pham, S. B. (2009). A hybrid approach to vietnamese word segmentation using part of speech tags. Proceedings from *2009 International Conference of Knowledge and Systems Engineering (IEEE)*. Washington, DC.

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

Pham, H. A. (2002). Gender in addressing and self-reference in vietnamese: Variation and change. In M. Hellinger &, H. Bussmann (Eds.), *Gender across Languages: The Linguistic Representation of Women and Men*, Vol. 2 (pp. 281-312). Amsterdam: John Benjamins.

Pham, H. A. (2008). The non-issue of dialect in teaching vietnamese. *Journal of Southeast Asian Language Teaching*, 14, 22-39.

Schiering, R. & Bickel, B. (2007). Does vietnamese have prosodic words? A mon-khmer development and its typological significance. Proceedings from *Austroasiatic Workshop,* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Scott, M. (2008). *WordSmith tools version 5*. Liverpool: Lexical Analysis Software.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Thomason, S. (2008). Social and linguistic factors as predictors of contact-induced change. *Journal of Language Contact-THEMA2.* Accessed January 23, 2013 at http://www.jlc-journal.org/.

Thompson, Laurence, C. (1965). *A vietnamese grammar*. Seattle: University of Washington Press.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam and Philadelphia: John Benjamins.

Tuc, H. D. (2003). *Vietnamese-english bilingualism: Patterns of code-switching*. London: RoutledgeCurzon.

US Census Bureau. (2000). *American community survey*. Accessed November 10, 2011 at http://www.census.gov/acs/www

US Census Bureau. (2010). *American community survey*. Accessed November 10, 2011 at http://www.census.gov/acs/www/.

Viện Khoa Học Xã Hội Việt Nam: Viện Ngôn Ngữ Học. (2009). *Từ Điển Anh-Việt: English-Vietnamese Dictionary*. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

Vigdor, J. L. (2008). Measuring immigrant assimilation in the united states. *Civic*

Report, 53. Accessed on June 15, 2011 at http://www.manhattan-institute.org/html/cr_53.htm.

Vu, H.Q., Pham, N.T., Nguyen, D.H.H., Huynh, B.T., Le, H.B., & Hoang, K. (2003). Towards a multi-objective corpus for vietnamese language. Proceedings from *COCOSDA.* Singapore.

Weinreich, U. (1963). *Languages in contact: Findings and problems*. The Hague, Mouton.

APPENDIX A

PRE-ANALYSIS CORPUS COUNTS

Table 1: Word, Blog and Comments Totals

Blogger	Total Words	Blogs Words	Comments Words	Percent: Comments/Blogs
Aus-1	23,891	19,426	4,465	22.98%
Aus-2	352,694	340,461	12,233	3.59%
Aus-3	68,598	48,132	20,466	42.52%
Aus-4	23,947	21,956	1,991	9.07%
Aus-5	68,167	14,648	53,519	365.37%
Aus-6	22,050	21,682	368	1.70%
Aus-7	1,427,203	288,500	1,138,703	394.70%
Aus-8	54,534	53,451	1,083	2.03%
Aus-9	23,301	5,012	18,289	364.90%
Aus-10	31,891	17,118	14,773	86.30%
Aus-11	1,178	669	509	76.08%
Aus-12	23,527	12,252	11,275	92.03%
Aus-13	2,943	1,825	1,118	61.26%
Aus-14	1,762	713	1,049	147.12%
Aus-15	1,959	1,285	674	52.45%
Aus-16	12,442	3,977	8,465	212.85%
Aus-17	2,296	1,306	990	75.80%
Aus-18	4,157	2,089	2,068	98.99%
Aus-19	2,101	1,594	507	31.81%
Aus-20	1,815	1,522	293	19.25%
Aus-21	18,441	18,119	322	1.78%
Aus-22	27,441	27,136	305	1.12%
Aus-23	36,560	31,639	4,921	15.55%
Aus-24	59,186	38,613	20,573	53.28%
Aus-25	1,748	1,748	0	0.00%
US-1	29,837	29,837	0	0.00%
US-2	67,483	67,483	0	0.00%
US-3	8,402	2,944	5,458	185.39%
US-4	15,238	11,245	3,993	35.51%

US-5	44,988	44,988	0	0.00%
US-6	22,810	19,069	3,741	19.62%
US-7	53,669	45,355	8,314	18.33%
US-8	67,647	35,244	32,403	91.94%
US-9	87,656	73,465	14,191	19.32%
US-10	64,600	34,661	29,939	86.38%
US-11	165,048	102,534	62,514	60.97%
US-12	34,499	19,546	14,953	76.50%
US-13	155,519	85,661	69,858	81.55%
US-14	40,819	26,324	14,495	55.06%
US-15	17,108	9,517	7,591	79.76%
US-16	101,250	48,003	53,247	110.92%
US-17	71,708	42,149	29,559	70.13%
US-18	128,085	50,689	77,396	152.69%
US-19	100,691	32,638	68,053	208.51%
US-20	78,715	39,898	38,817	97.29%
US-21	179,622	88,495	91,127	102.97%
US-22	330,248	179,592	150,656	83.89%
US-23	156,844	79,359	77,485	97.64%
US-24	3,117	3,085	32	1.04%
US-25	11,609	7,083	4,526	63.90%
VN-1	76,079	26,631	49,448	185.68%
VN-2	69,539	47,097	22,442	47.65%
VN-3	297,857	57,526	240,331	417.78%
VN-4	82,046	53,951	28,095	52.08%
VN-5	20,020	18,435	1,585	8.60%
VN-6	29,556	23,300	6,256	26.85%
VN-7	13,796	13,632	164	1.20%
VN-8	52,082	40,431	11,651	28.82%
VN-9	19,077	14,626	4,451	30.43%
VN-10	28,626	28,203	423	1.50%
VN-11	127,749	58,956	68,793	116.69%
VN-12	5,483	3,817	1,666	43.65%
VN-13	9,110	7,877	1,233	15.65%
VN-14	58,062	26,877	31,185	116.03%
VN-15	129,527	70,304	59,223	84.24%
VN-16	78,489	77,629	860	1.11%
VN-17	74,556	74,556	0	0.00%
VN-18	75,632	75,474	158	0.21%
VN-19	24,884	21,123	3,761	17.81%

VN-20	137,593	69,997	67,596	96.57%
VN-21	12,659	11,901	758	6.37%
VN-22	29,324	22,096	7,228	32.71%
VN-23	99,985	66,587	33,398	50.16%
VN-24	116,938	111,729	5,209	4.66%
VN-25	91,233	88,694	2,539	2.86%

APPENDIX B

FULL CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
LÀ	87,789	146
CÓ	84,069	146
#	79,496	143
KHÔNG	67,581	145
VÀ	60,261	144
CỦA	57,949	143
MỘT	51,757	144
CHO	46,439	143
NGƯỜI	46,112	142
THÌ	40,114	145
MÀ	38,333	144
CŨNG	37,940	143
ĐƯỢC	35,400	146
NHỮNG	33,910	143
TRONG	32,376	141
NÀY	31,434	146
NHƯ	30,798	141
TÔI	30,559	128
MÌNH	30,377	142
EM	30,308	136
LÀM	29,833	140
ĐI	29,738	141
VỚI	28,715	142
CÁI	28,599	142

ANH	28,573	138
CON	28,020	141
ĐỂ	27,746	142
ĐÓ	27,746	142
Ở	27,598	141
ĐÃ	26,831	142
LẠI	26,291	142
RA	26,101	143
PHẢI	24,651	142
CÁC	24,392	135
VỀ	23,966	142
CÒN	22,532	141
NHƯNG	22,278	145
NHÀ	22,238	140
NÓI	21,952	138
RỒI	21,798	140
CHỈ	21,366	143
KHI	21,035	142
ĐẾN	19,946	142
NHIỀU	19,461	142
BẠN	19,408	142
THẾ	18,525	138
CHỊ	18,329	134
BIẾT	18,200	142
NÀO	18,169	143
GÌ	18,165	143
NƯỚC	17,808	130
NĂM	17,547	135
THẤY	17,387	142
ÔNG	17,203	134
TỪ	17,004	137

HAY	16,858	142
TA	16,640	134
VÀO	16,590	140
MỚI	16,497	140
NGÀY	16,345	138
HỌC	16,261	131
NÊN	16,055	142
DÂN	15,929	119
VÌ	15,800	141
ĂN	15,646	133
CẢ	15,402	137
VẬY	14,859	138
QUÁ	14,442	140
TRÊN	14,344	133
SẼ	14,299	141
THỂ	14,274	136
HƠN	13,842	141
QUA	13,697	141
RẤT	13,337	140
SỰ	13,271	130
VIỆT	13,112	124
NÓ	13,054	135
BỊ	12,894	135
SAO	12,847	140
MẸ	12,588	130
GIỜ	12,399	138
CÔNG	12,265	133
LÊN	12,222	137
CHÍNH	12,047	124
ĐẦU	11,911	130
HỌ	11,867	123

HAI	11,829	133
AI	11,797	133
NAM	11,698	124
NAY	11,585	138
ĐÂY	11,298	137
SAU	11,293	137
MẤY	10,981	134
QUỐC	10,902	101
BÀI	10,891	135
THÀNH	10,770	134
THEO	10,578	137
THÔI	10,510	136
NỮA	10,238	136
VẪN	10,175	132

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
CÓ THỂ	8,151	130
VIỆT NAM	7,030	108
KHÔNG CÓ	6,809	131
LÀ MỘT	6,245	121
NGƯỜI TA	4,778	121
KHÔNG PHẢI	4,453	126
NHỮNG NGƯỜI	4,057	114
BÂY GIỜ	3,795	124
CÓ MỘT	3,679	122
ĐÓ LÀ	3,666	121
CẢM ƠN	3,534	99
NĂM #	3,531	110
NHƯ THẾ	3,488	116

KHÔNG BIẾT	3,455	126
CHỈ CÓ	3,394	125
GIA ĐÌNH	3,388	117
CŨNG CÓ	3,327	119
NHƯ VẬY	3,307	119
# NĂM	3,249	118
CỦA MÌNH	3,237	122
THẾ GIỚI	3,232	106
CHÚNG TA	3,079	109
THỜI GIAN	3,042	124
TRUNG QUỐC	3,038	51
XÃ HỘI	2,993	87
CÁM ƠN	2,812	98
NÀO CŨNG	2,810	123
BAO GIỜ	2,780	119
# #	2,779	99
CÁC BẠN	2,767	103
TẤT CẢ	2,689	116
KHOA HỌC	2,608	61
NGHIÊN CỨU	2,603	73
KHÔNG THỂ	2,593	113
CŨNG KHÔNG	2,578	110
VẤN ĐỀ	2,573	95
HÔM NAY	2,450	123
NHẤT LÀ	2,395	122
MỌI NGƯỜI	2,298	112
LÀ NGƯỜI	2,259	115
LÀ NHỮNG	2,250	107
CON NGƯỜI	2,214	112
HẠNH PHÚC	2,199	117
CÓ LẼ	2,194	115

CÓ NHIỀU	2,184	116
CHÍNH TRỊ	2,179	47
KINH TẾ	2,178	75
DÂN CHỦ	2,166	36
CHỈ LÀ	2,163	120
MÀ KHÔNG	2,160	120
CÓ NHỮNG	2,137	119
CÓ #	2,125	113
VĂN HÓA	2,041	73
ĐÂY LÀ	2,033	114
CỦA NGƯỜI	2,012	114
PHẢI LÀ	2,005	118
THẾ NÀO	1,999	121
MỘT NGƯỜI	1,997	108
CŨNG LÀ	1,986	116
BAO NHIÊU	1,967	111
ĐẤT NƯỚC	1,922	66
GỌI LÀ	1,916	114
CHÚNG TÔI	1,899	71
NGƯỜI VIỆT	1,874	84
Ở ĐÂY	1,855	113
MỘT CÁCH	1,829	99
NĂM NAY	1,828	107
TỰ DO	1,825	83
TRẢ LỜI	1,813	112
NHÀ NƯỚC	1,810	56
RẤT NHIỀU	1,808	112
MỘT SỐ	1,791	94
LÃNH ĐẠO	1,777	50
HÀ NỘI	1,772	85
CỦA CÁC	1,763	92

CHỨ KHÔNG	1,755	110
DÂN TỘC	1,750	60
TÁC GIẢ	1,747	90
CÁI GÌ	1,745	110
CUỘC SỐNG	1,730	101
QUỐC GIA	1,718	62
LÀM SAO	1,716	111
TẠI SAO	1,689	112
ĐẦU TIÊN	1,676	110
ĐÚNG LÀ	1,664	108
TÔI KHÔNG	1,663	85
BẮT ĐẦU	1,646	110
THÁNG #	1,642	104
AI CŨNG	1,623	110
VẪN CÒN	1,620	114
ĐÃ CÓ	1,616	117
CỦA ANH	1,610	105
NHIỀU NGƯỜI	1,606	106
NHÂN DÂN	1,598	41
LÀM GÌ	1,595	107
CỦA ÔNG	1,591	92
KHÔNG CÒN	1,580	109
LÀ #	1,565	113
NGƯỜI DÂN	1,562	69
SAU KHI	1,549	109

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
KHÔNG PHẢI LÀ	1,345	107
Ở VIỆT NAM	1,002	75

LÚC NÀO CŨNG	880	101
KHÔNG BAO GIỜ	852	96
CHỨ KHÔNG PHẢI	745	86
NHƯ THẾ NÀO	744	92
TRÊN THẾ GIỚI	715	63
# # #	689	26
CẢM ƠN EM	592	26
CŨNG CÓ THỂ	589	87
HI HI HI	586	25
MỘT TRONG NHỮNG	584	78
# THÁNG #	527	65
HA HA HA	525	38
TẤT CẢ CÁC	522	72
NHƯ THẾ NÀY	521	82
LIÊN QUAN ĐẾN	518	54
CHƯA BAO GIỜ	514	82
THÂN ÁI GỬI	512	2
BẠN DÂN CHỦ	501	2
CẢM ƠN ANH	489	38
CÓ NGHĨA LÀ	477	73
LẦN ĐẦU TIÊN	477	87
CÁC BẠN DÂN	469	3
KHÔNG CÓ GÌ	468	80
NÀO CŨNG CÓ	451	84
TẤT CẢ NHỮNG	451	80
NHÀ KHOA HỌC	449	21
CỦA VIỆT NAM	445	43
CŨNG LÀ MỘT	434	75
CÓ THỂ NÓI	432	56
LÀ MỘT TRONG	428	72
CHỈ LÀ MỘT	426	76

ĐÓ LÀ MỘT	422	58
VÀ GIA ĐÌNH	403	61
LÀ NHỮNG NGƯỜI	401	65
NGÀY # THÁNG	400	50
NGƯỜI VIỆT NAM	397	53
CHỈ CÓ #	396	74
CÂU TRẢ LỜI	395	69
BÀI VIẾT CỦA	388	55
CÓ KHẢ NĂNG	379	60
# MẸ CON	367	33
CÀNG NGÀY CÀNG	367	58
HỒ CHÍ MINH	365	27
TA CÓ THỂ	365	67
TRONG ĐÓ CÓ	353	72
CÓ THỂ LÀ	350	72
CỦA NHỮNG NGƯỜI	350	67
NỀN KINH TẾ	350	27
CỦA TRUNG QUỐC	346	22
ĐÂY LÀ MỘT	340	61
ĐẶC BIỆT LÀ	328	70
CÓ THỜI GIAN	319	77
VÀ HẠNH PHÚC	311	56
CHỈ CÓ MỘT	308	68
NÓI CHUYỆN VỚI	307	74
# ĐẾN #	305	48
HƠN # NĂM	303	68
MỘT THỜI GIAN	302	69
QUAN TÂM ĐẾN	302	59
TỪ NĂM #	296	49
CÓ RẤT NHIỀU	295	75
ĐỒNG Ý VỚI	295	56

CÁM ƠN BẠN	294	42
NGƯỜI TA KHÔNG	293	60
# NĂM #	291	55
CÓ CƠ HỘI	290	74
TRỞ THÀNH MỘT	290	43
HÌ HÌ HÌ	287	20
NGƯỜI ĐÀN ÔNG	287	59
BẠN CÓ THỂ	285	61
NGƯỜI PHỤ NỮ	282	57
VÀO NĂM #	282	36
AI CŨNG BIẾT	280	58
CÁM ƠN ANH	280	45
BAO GIỜ CŨNG	276	53
CÓ THỂ LÀM	274	74
CHO ĐẾN KHI	273	71
CỦA CHÚNG TA	273	56
TRƯỜNG ĐẠI HỌC	270	47
ĐÃ GHÉ THĂM	268	10
NGHIÊN CỨU KHOA	268	11
CỨU KHOA HỌC	267	11
CÁM ƠN CHỊ	264	41
ĐẢNG CỘNG SẢN	264	14
THÁNG # NĂM	264	58
LÀ MỘT NGƯỜI	263	64
CÓ Ý NGHĨA	261	61
NẾU KHÔNG CÓ	261	64
TẤT CẢ MỌI	260	62
BIẾT BAO NHIÊU	259	60
CÁI GÌ CŨNG	259	72
{Last} {Middle} {First}	258	3
CÁI GỌI LÀ	257	35

NGƯỜI TA CÓ	257	65
CỦA CON NGƯỜI	256	58
TIẾNG ĐỒNG HỒ	256	65
BÂY GIỜ THÌ	255	62
ANH EM {Name}	253	9

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
CÁC BẠN DÂN CHỦ	465	2
# # # #	413	11
NGÀY # THÁNG #	354	47
LÀ MỘT TRONG NHỮNG	313	68
NGHIÊN CỨU KHOA HỌC	267	11
XÃ HỘI CHỦ NGHĨA	230	15
ANH {Last} {Middle} {First}	215	2
THÁNG # NĂM #	214	44
BÀI BÁO KHOA HỌC	207	3
CẢM ƠN EM ĐÃ	177	5
CÔNG TRÌNH NGHIÊN CỨU	162	10
TẤT CẢ MỌI NGƯỜI	155	52
CÁC NHÀ KHOA HỌC	143	16
TỪ # ĐẾN #	138	38
NGƯỜI TA CÓ THỂ	136	43
CHỨ KHÔNG PHẢI LÀ	135	40
CỘNG SẢN VIỆT NAM	131	10
CHÚNG TA CÓ THỂ	128	28
KHÔNG PHẢI LÀ MỘT	126	41
CHÚC MỪNG SINH NHẬT	125	35
# THÁNG # NĂM	124	34
# TIẾNG ĐỒNG HỒ	123	43

TRÊN CÁC TẬP SAN	121	5
ĐẢNG VÀ NHÀ NƯỚC	117	9
THẾ LỰC THÙ ĐỊCH	116	10
TỰ DO DÂN CHỦ	114	10
CHẾ ĐỘ ĐỘC TÀI	112	10
CHẾ ĐỘ CỘNG SẢN	109	8
SẼ KHÔNG BAO GIỜ	108	47
KHÔNG BAO GIỜ CÓ	105	35
ẤN PHẨM KHOA HỌC	104	2
CHÚC MỪNG NĂM MỚI	104	41
CẢM ƠN ANH ĐÃ	103	10
VIỆT NAM HIỆN NAY	102	15
CÓ THỂ NÓI LÀ	101	26
ĐỀ CƯƠNG NGHIÊN CỨU	101	3
CỘNG ĐỒNG NGƯỜI VIỆT	95	17
GỬI ANH {Last} {Middle}	95	2
ĐÃ GHÉ THĂM VÀ	93	4
XÃ HỘI DÂN SỰ	93	8
CHÚC EM LUÔN VUI	92	3
LÀ LẦN ĐẦU TIÊN	92	45
BÀI VIẾT CỦA ANH	91	14
TẬP SAN KHOA HỌC	89	3
KHÔNG PHẢI LÀ NGƯỜI	86	27
ANH {Last} {Middle} {First}	85	2
KHÔNG AI CÓ THỂ	85	32
NHÂN DÂN VIỆT NAM	85	10
CHỦ NGHĨA XÃ HỘI	83	8
GỬI ANH {Last} {Middle}	83	2
KHÔNG CÓ NGHĨA LÀ	83	34
TẤT CẢ NHỮNG GÌ	83	37
ỦNG HỘ TINH THẦN	83	10

DÂN TỘC VIỆT NAM	82	11
LÀM THẾ NÀO ĐỂ	81	33
TỰ DO NGÔN LUẬN	80	13
TRONG VÀ NGOÀI NƯỚC	79	20
MẸ CON NHÀ {Name}	78	5
KHỎE VÀ HẠNH PHÚC	77	10
KHÔNG BIẾT BAO NHIÊU	77	38
CÔNG BỐ QUỐC TẾ	76	2
ĐỂ HÔN EM LẦN	75	2
HÔN EM LẦN NỮA	75	2
KINH TẾ THỊ TRƯỜNG	75	11
BIỂU TÌNH CHỐNG TRUNG	74	9
ĐẢNG CỘNG SẢN VIỆT	74	8
KHÔNG CÓ THỜI GIAN	73	40
PHÁT TRIỂN KINH TẾ	73	16
TRONG VÒNG # NĂM	73	26
PHÊ BÌNH VĂN HỌC	72	6
TRẢ LỜI CÂU HỎI	72	23
BUỔI TỐI VUI VẺ	71	2
CÁCH ĐÂY # NĂM	71	35
KHÔNG PHẢI AI CŨNG	71	35
MỘT NGƯỜI ĐÀN ÔNG	71	33
CHIA SẺ THÂN ÁI	70	1
GỬI {Last} {Middle} {First}	70	1
TAM QUYỀN PHÂN LẬP	70	3
CHỦ NGHĨA TƯ BẢN	69	11
CÓ LIÊN QUAN ĐẾN	69	16
LÚC NÀO CŨNG CÓ	69	34
ĐÓ KHÔNG PHẢI LÀ	68	20
EM ĐÃ CHIA SẺ	68	4
SỰ PHÁT TRIỂN CỦA	68	17

THÂN ÁI GỬI ANH	68	2
THÂN ÁI GỬI {Name}	68	1
NƯỚC ÚC PHÁT THÈM	67	1
ƠN EM ĐÃ CHIA	67	4
TỪ NĂM # ĐẾN	67	15
GỬI {Last} {Middle} {First}	66	1
HOÀN TOÀN KHÔNG CÓ	66	18
CẢM ƠN ANH NHIỀU	65	10
CHỦ NGHĨA CỘNG SẢN	65	10
ĐỘC LẬP DÂN TỘC	65	4
ĐỘC LẬP TỰ DO	65	11
Á HI HI HI	64	6
CŨNG KHÔNG PHẢI LÀ	64	27
QUAN TRỌNG NHẤT LÀ	64	21
CÁC CUỘC BIỂU TÌNH	63	8

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
# # # # #	287	6
# THÁNG # NĂM #	118	33
NGÀY # THÁNG # NĂM	108	31
GỬI ANH {Last} {Middle} {First}	93	2
GỬI ANH {Last} {Middle} {First}	81	2
ĐỂ HÔN EM LẦN NỮA	75	2
ĐẢNG CỘNG SẢN VIỆT NAM	70	8
ƠN EM ĐÃ CHIA SẺ	67	4
BIỂU TÌNH CHỐNG TRUNG QUỐC	63	8
ỦNG HỘ TINH THẦN CHO	62	4
CẢM ƠN EM ĐÃ CHIA	61	1
CỦA CÁC BẠN DÂN CHỦ	60	2

CÁC BẠN DÂN CHỦ THÌ	59	2
SỐ ẤN PHẨM KHOA HỌC	58	2
CHÚC EM LUÔN VUI KHỎE	57	1
CẢM ƠN EM CHÚC EM	53	2
ĐÃ CHIA SẺ THÂN ÁI	53	1
CÁC TẬP SAN QUỐC TẾ	52	3
TRÊN CÁC TẬP SAN QUỐC	52	3
XÃ HỘI CHỦ NGHĨA VIỆT	52	7
GỬI CHÚ {Last} {Middle} {First}	50	1
THÀNH PHỐ HỒ CHÍ MINH	49	15
CÁC THẾ LỰC THÙ ĐỊCH	48	9
ĐÂY LÀ LẦN ĐẦU TIÊN	48	31
GỬI {Name} {Name} {Name} {Name}	48	1
HỘI CHỦ NGHĨA VIỆT NAM	48	6
HOÀNG SA VÀ TRƯỜNG SA	47	7
NHÀ CẦM QUYỀN VIỆT NAM	47	4
CẢM ƠN EM ĐÃ GHÉ	46	2
NẾU KHÔNG MUỐN NÓI LÀ	46	9
ƠN EM ĐÃ GHÉ THĂM	46	3
KHẮP NƠI TRÊN THẾ GIỚI	44	18
VUI VẺ VÀ HẠNH PHÚC	44	8
# BÀI BÁO KHOA HỌC	43	2
GHÉ THĂM VÀ CHIA SẺ	43	2
HÒA XÃ HỘI CHỦ NGHĨA	43	4
CÁM ƠN DÌ NHIỀU NHIỀU	41	1
LUÔN ỦNG HỘ TINH THẦN	41	4
VIỆT NAM VÀ TRUNG QUỐC	41	5
HỘ TINH THẦN CHO {Name}	40	2
VÀ HẠNH PHÚC THÂN ÁI	40	1
BAN CHẤP HÀNH TRUNG ƯƠNG	39	3
CỘNG HÒA XÃ HỘI CHỦ	39	4

MẠNH KHỎE VÀ HẠNH PHÚC	39	4
SỐ BÀI BÁO KHOA HỌC	39	2
VÀO NGÀY # THÁNG #	39	16
BÀ AUNG SAN SUU KYI	38	7
CÁC TẬP SAN KHOA HỌC	38	3
ĐÃ GHÉ THĂM VÀ CHIA	37	1
HỌC TRÊN CÁC TẬP SAN	37	3
KHOA HỌC TRÊN CÁC TẬP	36	3
KHÔNG PHẢI LÚC NÀO CŨNG	36	25
MỘT CÔNG TRÌNH NGHIÊN CỨU	36	6
ƠN BẠN ĐÃ GHÉ THĂM	36	3
THỦ TƯỚNG NGUYỄN TẤN DŨNG	36	4
CẢM ƠN EM ĐÃ ĐỒNG	35	1
CHIA SẺ THÂN ÁI GỬI	35	1
HỘI NHÀ VĂN VIỆT NAM	35	5
TỰ DO CÁI CON C	35	3
VIẾT ĐỀ CƯƠNG NGHIÊN CỨU	35	2
ANH {Name} VÀ CHỊ {Name}	34	2
CHÚC ANH VÀ GIA ĐÌNH	34	11
ƠN EM ĐÃ ĐỒNG CẢM	34	1
TÀI TRỢ CHO NGHIÊN CỨU	34	2
VÀ GIA ĐÌNH NĂM MỚI	34	12
CÁC PHƯƠNG TIỆN TRUYỀN THÔNG	33	10
LÀ MỘT TRONG NHỮNG NGƯỜI	33	23
VẤN ĐỀ LIÊN QUAN ĐẾN	33	12
CÁC BẠN DÂN CHỦ KHÔNG	32	2
CHỦ NGHĨA TƯ BẢN THÂN	32	2
CHÚC CHỊ VÀ GIA ĐÌNH	32	11
NGHĨA TƯ BẢN THÂN HỮU	32	2
CHỈ BIẾT CÒN ĐẢNG CÒN	31	2
CHỈ CÓ TRUNG QUỐC LÀ	31	2

CHÚC ANH BUỔI TỐI VUI	31	2
MẠNH KHỎE VÀ VUI VẺ	31	3
NHÂN QUYỀN Ở VIỆT NAM	31	3
XÂY DỰNG ĐƯỢC GÌ TRÊN	31	2
BÀI VIẾT CỦA ÔNG QUỐC	30	2
BLOG FOR YOU GREAT HTTP	30	16
CÓ TRUNG QUỐC LÀ TỐT	30	1
CỦA ĐẢNG VÀ NHÀ NƯỚC	30	7
GỬI ANH {Last} {Middle} {First}	30	1
NGÀY MỚI NHIỀU NIỀM VUI	30	2
THANK BLOG FOR YOU GREAT	30	16
BÀI BÁO KHOA HỌC TRÊN	29	3
CÁI NƯỚC MÌNH NÓ THẾ	29	5
CHÂU Á THÁI BÌNH DƯƠNG	29	8
{Name} VÀ CHỊ {Name} {Name}	29	2
HOÀN TOÀN ĐỒNG Ý VỚI	29	14
HƠN ĐỘC LẬP TỰ DO	29	5
KHÔNG CÓ GÌ QUÝ HƠN	29	3
MUỐN LÀM GÌ THÌ LÀM	29	12
NƯỚC CỘNG HÒA XÃ HỘI	29	3
SINH RA VÀ LỚN LÊN	29	16
THẦY TRÒ CÁC BẠN DÂN	29	2
TRÒ CÁC BẠN DÂN CHỦ	29	2
# NĂM TRỞ LẠI ĐÂY	28	8
ANH BUỔI TỐI VUI VẺ	28	2

APPENDIX C

FULL CORPUS A-CURVE CHARTS

Chart 1a: Full Corpus 1-Tiếng Chart

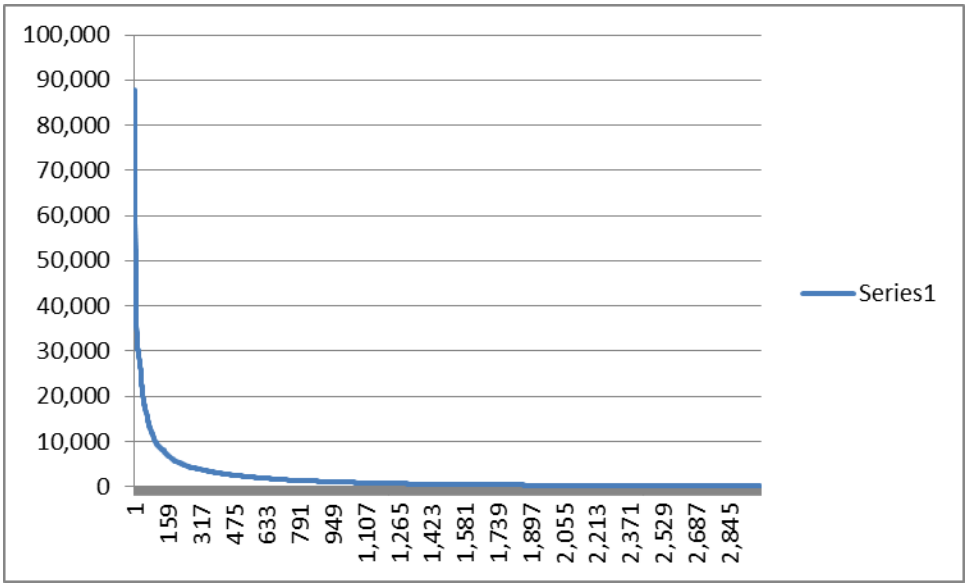


Chart 1b: Full Corpus 2-Tiếng Chart

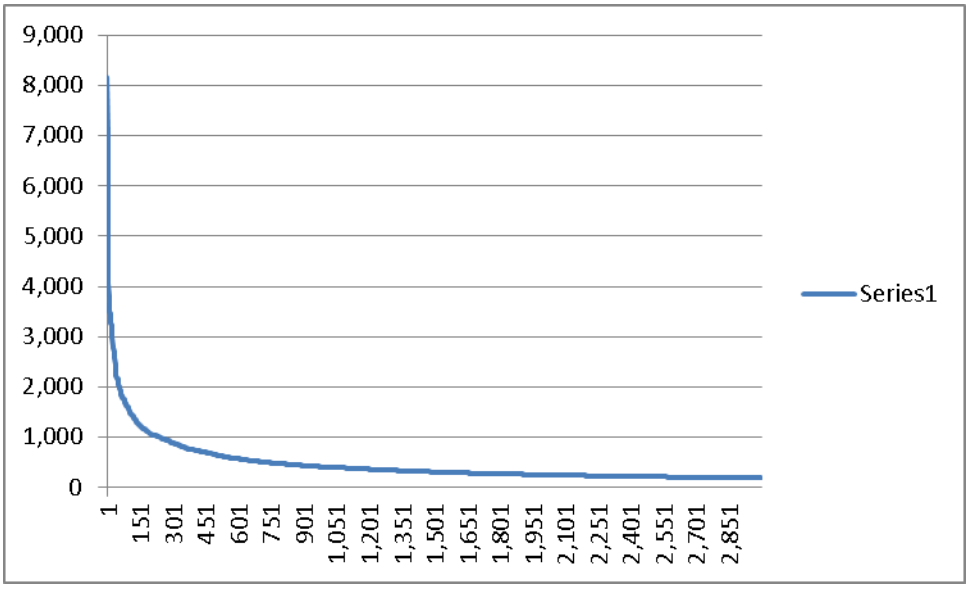


Chart 1c: Full Corpus 3-Tiếng Chart

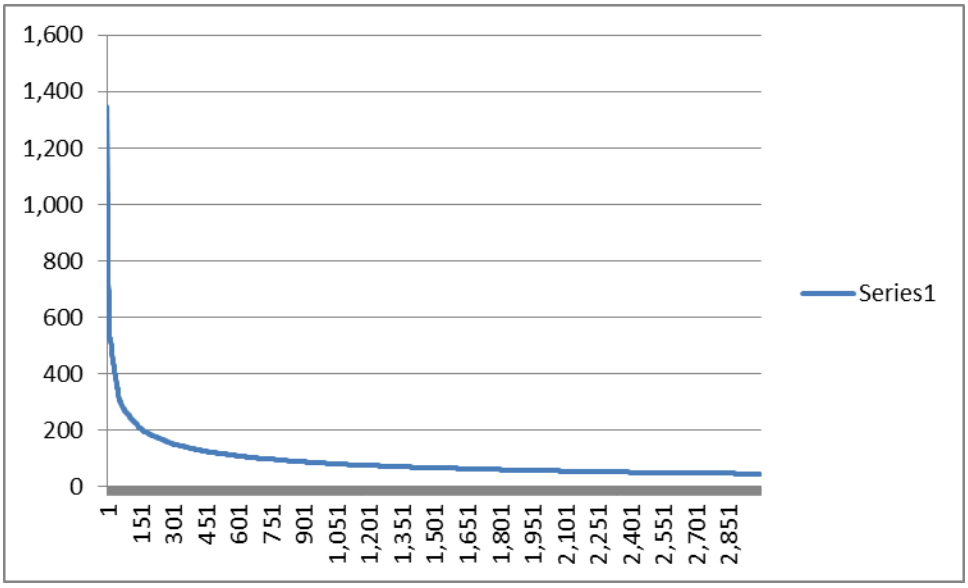


Chart 1d: Full Corpus 4-Tiếng Chart

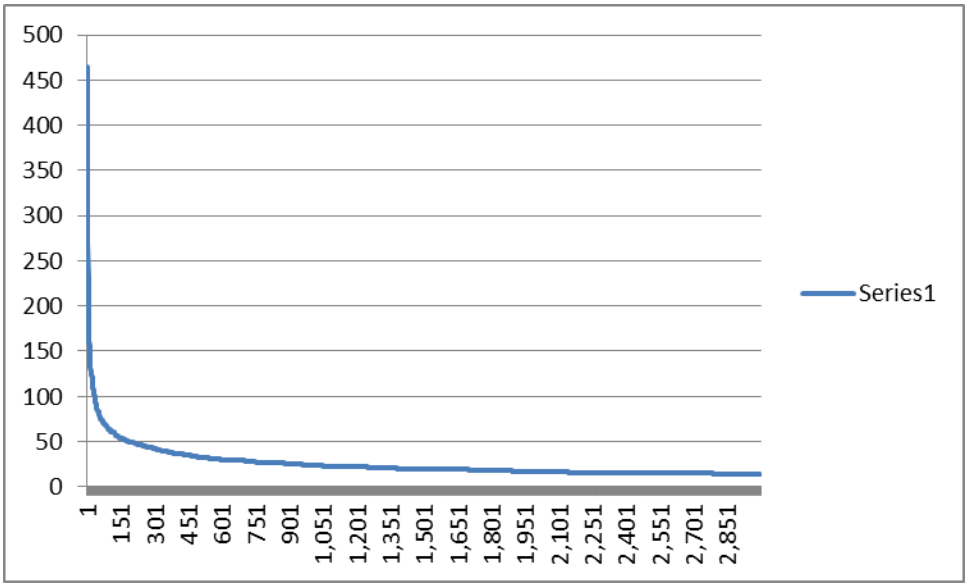
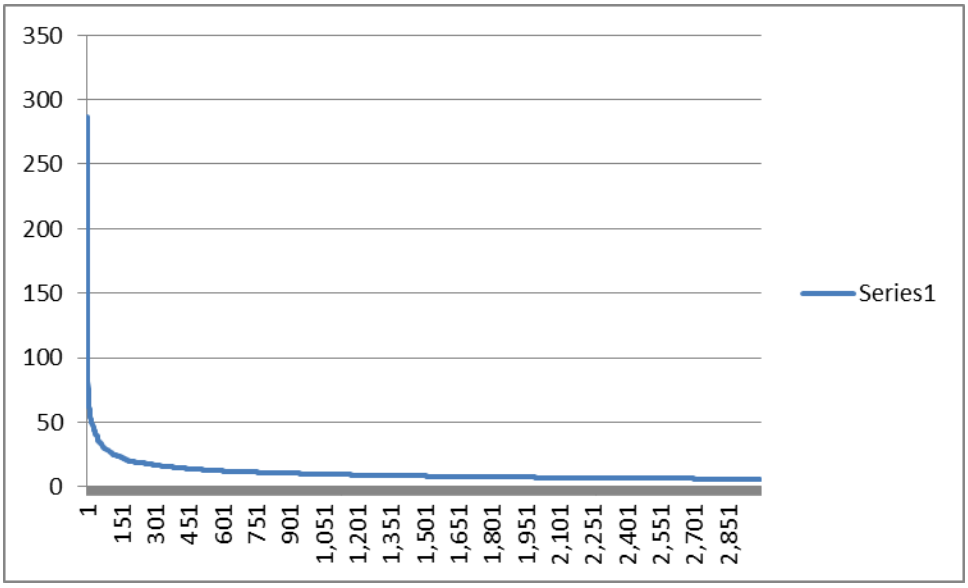


Chart 1e: Full Corpus 5-Tiếng Chart



APPENDIX D

BLOG CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
#	51,019	75
LÀ	45,851	76
CÓ	45,040	76
VÀ	37,400	76
MỘT	36,455	76
KHÔNG	35,554	76
CỦA	32,972	76
NGƯỜI	25,535	76
TÔI	24,851	71
CHO	24,716	76
NHỮNG	22,274	76
TRONG	21,403	75
ĐƯỢC	19,843	76
CŨNG	19,285	76
THÌ	19,104	76
MÌNH	18,706	75
VỚI	17,485	76
Ở	17,024	76
NHƯ	17,008	76
NÀY	16,949	76
ĐI	16,934	76
MÀ	16,606	76
LÀM	15,730	75
ĐỂ	15,616	76
ĐÃ	15,578	75

LẠI	15,431	75
RA	15,306	76
CÁC	15,302	75
VỀ	14,881	76
CON	14,787	75
CÁI	13,990	76
ĐÓ	13,923	75
ĐẾN	13,222	76
NHÀ	12,994	75
NHƯNG	12,785	76
PHẢI	12,704	76
KHI	12,682	76
CHỈ	11,815	75
ANH	11,633	73
CÒN	11,400	75
NĂM	11,367	75
RỒI	10,946	76
VÀO	10,912	76
NÓI	10,846	73
HỌC	10,825	74
TỪ	10,779	75
EM	10,131	71
NHIỀU	10,070	75
NGÀY	10,050	73
BẠN	9,449	75
THẾ	9,323	76
NÀO	9,302	76
THỂ	9,276	76
TRÊN	9,199	75
CẢ	9,035	73
THẤY	8,782	76

BIẾT	8,671	76
ĂN	8,619	73
NƯỚC	8,587	73
HAI	8,568	73
TA	8,534	74
VÌ	8,454	76
GÌ	8,425	75
ÔNG	8,415	75
NÊN	8,216	76
MỚI	8,161	75
HAY	8,111	76
SẼ	8,023	74
HƠN	7,956	76
QUA	7,797	76
ĐẦU	7,740	72
MẸ	7,740	69
SỰ	7,632	73
SAU	7,490	75
NÓ	7,389	74
LÊN	7,373	75
RẤT	7,302	74
CÔNG	7,179	73
VIỆT	7,151	71
ĐÂY	6,836	75
VẬY	6,672	74
GIỜ	6,666	73
BỊ	6,629	76
SỐ	6,574	72
THÀNH	6,400	75
CÔ	6,327	74
NAM	6,225	72

CHÍNH	6,150	70
CÁCH	6,112	76
NHẤT	6,107	73
KHÁC	6,095	74
THEO	6,060	75
NAY	5,945	73
TRƯỚC	5,914	76
VẪN	5,812	73
VIỆC	5,734	71
HỌ	5,715	71
MẤY	5,653	73
CHUYỆN	5,620	75
CHÚNG	5,562	74

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
CÓ THỂ	5,682	74
VIỆT NAM	4,507	63
LÀ MỘT	4,441	72
KHÔNG CÓ	3,524	73
NGƯỜI TA	2,815	68
NĂM #	2,585	65
CÓ MỘT	2,503	71
NHỮNG NGƯỜI	2,481	68
KHÔNG PHẢI	2,478	68
# #	2,444	57
KHOA HỌC	2,379	39
NGHIÊN CỨU	2,345	45
ĐÓ LÀ	2,125	73

CỦA MÌNH	2,001	70
THỜI GIAN	1,995	70
CHÚNG TA	1,941	61
# NĂM	1,897	69
NHƯ THẾ	1,867	68
BÂY GIỜ	1,830	69
GIA ĐÌNH	1,824	65
CHỈ CÓ	1,758	71
CŨNG CÓ	1,750	69
TẤT CẢ	1,720	67
KHÔNG BIẾT	1,702	70
VẤN ĐỀ	1,600	54
NHƯ VẬY	1,576	70
TRUNG QUỐC	1,571	35
HÔM NAY	1,552	68
THẾ GIỚI	1,547	64
BAO GIỜ	1,487	67
CHÚNG TÔI	1,487	51
KHÔNG THỂ	1,470	64
CŨNG KHÔNG	1,469	63
CÓ LẼ	1,457	65
MỘT NGƯỜI	1,454	65
NÀO CŨNG	1,453	68
LÀ NHỮNG	1,414	65
XÃ HỘI	1,395	51
NHẤT LÀ	1,376	71
MỘT SỐ	1,363	62
MỌI NGƯỜI	1,337	66
MỘT CÁCH	1,325	64
ĐÂY LÀ	1,318	69
CÓ NHỮNG	1,296	66

TÔI KHÔNG	1,294	55
ĐẦU TIÊN	1,286	66
KHOẢNG #	1,261	58
CÁC BẠN	1,255	61
BẮT ĐẦU	1,250	67
NĂM NAY	1,240	66
Ở ĐÂY	1,240	66
TÁC GIẢ	1,220	53
ĐẠI HỌC	1,203	55
LÀ NGƯỜI	1,194	68
THÁNG #	1,192	62
CỦA TÔI	1,189	61
CÓ NHIỀU	1,148	69
CỦA NGƯỜI	1,129	69
CÓ #	1,124	62
CŨNG LÀ	1,107	67
QUAN TRỌNG	1,089	62
SAU KHI	1,086	65
CỦA CÁC	1,084	57
CON NGƯỜI	1,081	67
CHỈ LÀ	1,077	67
KINH TẾ	1,072	45
TRẢ LỜI	1,070	68
PHẢI LÀ	1,065	66
ĐẶC BIỆT	1,063	65
BAO NHIÊU	1,062	66
HÀ NỘI	1,056	49
THẾ NÀO	1,054	68
ĐÃ CÓ	1,049	69
MỘT CÁI	1,042	63
RẤT NHIỀU	1,039	65

MÀ KHÔNG	1,027	68
CUỐI CÙNG	1,012	68
SAU ĐÓ	1,010	68
THÀNH PHỐ	1,010	63
VỚI NHỮNG	1,003	63
CỦA MỘT	995	65
GỌI LÀ	985	65
TÔI ĐÃ	982	58
CŨNG NHƯ	971	59
VẪN CÒN	960	67
TRONG NHỮNG	956	66
CUỘC SỐNG	950	64
LÀM VIỆC	945	64
CHỨ KHÔNG	933	63
LÀ #	930	62
CHÍNH TRỊ	913	31
NHIỀU NGƯỜI	909	64
KHÔNG CÒN	905	67
Ở NHÀ	897	64
TÔI CÓ	888	60
MỘT CHÚT	878	67
QUỐC GIA	877	41
TẠI SAO	877	65
NGƯỜI VIỆT	873	50
TRƯỚC KHI	870	69

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
Ở VIỆT NAM	825	51

KHÔNG PHẢI LÀ	744	61
# # #	659	17
MỘT TRONG NHỮNG	469	53
LÚC NÀO CŨNG	460	58
NHÀ KHOA HỌC	436	15
# THÁNG #	433	39
KHÔNG BAO GIỜ	433	59
TRÊN THẾ GIỚI	424	43
LIÊN QUAN ĐẾN	412	34
NHƯ THẾ NÀO	401	53
CŨNG CÓ THỂ	399	55
CHỨ KHÔNG PHẢI	391	51
TẤT CẢ CÁC	388	46
LẦN ĐẦU TIÊN	353	57
LÀ MỘT TRONG	342	49
ĐÓ LÀ MỘT	339	42
NGÀY # THÁNG	338	33
CÓ THỂ NÓI	333	38
CỦA VIỆT NAM	332	35
NHƯ THẾ NÀY	319	48
TẤT CẢ NHỮNG	307	48
CŨNG LÀ MỘT	300	49
CHƯA BAO GIỜ	294	51
CHỈ LÀ MỘT	276	48
CHỈ CÓ #	272	47
TA CÓ THỂ	272	48
# ĐẾN #	260	35
CÓ NGHĨA LÀ	260	47
CỨU KHOA HỌC	253	8
ĐÂY LÀ MỘT	253	45
NGHIÊN CỨU KHOA	253	8

CÂU TRẢ LỜI	250	46
LÀ NHỮNG NGƯỜI	243	43
KHÔNG CÓ GÌ	239	46
CÓ KHẢ NĂNG	238	39
CÓ THỂ LÀ	238	44
VÀO NĂM #	236	28
# NĂM #	230	39
ĐẶC BIỆT LÀ	230	43
BẠN CÓ THỂ	228	45
MỘT THỜI GIAN	225	48
NÀO CŨNG CÓ	225	50
TỪ NĂM #	224	34
TRỞ THÀNH MỘT	223	37
CỦA NHỮNG NGƯỜI	219	45
TRONG ĐÓ CÓ	218	45
CỦA TRUNG QUỐC	217	16
NÓI CHUYỆN VỚI	217	48
NGƯỜI VIỆT NAM	215	38
QUAN TÂM ĐẾN	214	39
TIẾNG ĐỒNG HỒ	214	44
THÁNG # NĂM	209	39
BÁO KHOA HỌC	208	4
CÀNG NGÀY CÀNG	207	29
TRƯỜNG ĐẠI HỌC	207	34
BÀI BÁO KHOA	206	2
ĐƯỢC XEM LÀ	204	19
CHO ĐẾN KHI	200	51
TRONG THỜI GIAN	199	39
NGƯỜI ĐÀN ÔNG	198	41
CÓ RẤT NHIỀU	192	46
CÓ THỂ LÀM	192	49

CHỈ CÓ MỘT	191	47
MỘT NGƯỜI BẠN	190	52
QUAN TRỌNG NHẤT	185	31
CỦA CHÚNG TA	184	40
BAO GIỜ CŨNG	183	28
CÓ Ý NGHĨA	183	37
CHÚNG TA CÓ	182	31
LÀ MỘT NGƯỜI	182	43
CÁC TẬP SAN	181	4
NGƯỜI PHỤ NỮ	181	42
TRÌNH NGHIÊN CỨU	181	8
NHỮNG VẤN ĐỀ	178	23
ANH EM {Name}	173	1
TÔI CÓ THỂ	168	40
NGƯỜI TA CÓ	167	43
NỀN KINH TẾ	165	18
HƠN # NĂM	164	41
CÓ THỜI GIAN	163	43
TẤT CẢ MỌI	163	38
VIỆT NAM VÀ	163	19
CÓ CƠ HỘI	162	43
AI CŨNG BIẾT	161	41
ĐÓ LÀ NHỮNG	161	37
CHỦ YẾU LÀ	159	33
ĐÔNG NAM Á	159	18
CÁC QUỐC GIA	158	17
KHÔNG PHẢI CHỈ	158	23
TRONG KHI ĐÓ	158	36
CÔNG TRÌNH NGHIÊN	157	7
HAI VỢ CHỒNG	157	29
TẤT CẢ ĐỀU	157	34

CHẲNG HẠN NHƯ	154	16
THẾ KỶ #	154	24
TRONG TRƯỜNG HỢP	154	29
CÓ LẼ LÀ	151	34
ĐẦU TIÊN CỦA	151	46
MỘT SỐ NGƯỜI	151	31

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
# # # #	401	4
NGÀY # THÁNG #	302	31
NGHIÊN CỨU KHOA HỌC	253	8
LÀ MỘT TRONG NHỮNG	252	48
BÀI BÁO KHOA HỌC	206	2
THÁNG # NĂM #	169	30
CÔNG TRÌNH NGHIÊN CỨU	157	7
CÁC NHÀ KHOA HỌC	132	11
TRÊN CÁC TẬP SAN	120	4
TỪ # ĐẾN #	116	28
CHÚNG TA CÓ THỂ	109	22
# TIẾNG ĐỒNG HỒ	104	32
ẤN PHẨM KHOA HỌC	104	2
TẤT CẢ MỌI NGƯỜI	104	33
ĐỀ CƯƠNG NGHIÊN CỨU	100	2
# THÁNG # NĂM	99	22

NGƯỜI TA CÓ THỂ	98	32
KHÔNG PHẢI LÀ MỘT	90	28
TẬP SAN KHOA HỌC	89	3
LÀ LẦN ĐẦU TIÊN	79	36
CÓ THỂ NÓI LÀ	77	19
CÔNG BỐ QUỐC TẾ	76	2
XÃ HỘI CHỦ NGHĨA	69	10
VIỆT NAM HIỆN NAY	67	10
KHÔNG AI CÓ THỂ	65	25
CỘNG ĐỒNG NGƯỜI VIỆT	64	10
CHỨ KHÔNG PHẢI LÀ	63	20
PHÊ BÌNH VĂN HỌC	63	5
LÀM THẾ NÀO ĐỂ	62	21
SẼ KHÔNG BAO GIỜ	62	32
TRONG VÒNG # NĂM	62	20
MỘT NGƯỜI ĐÀN ÔNG	61	26
TẬP SAN QUỐC TẾ	59	2
ĐỂ HÔN EM LẦN	58	1
HÔN EM LẦN NỮA	58	1
SỐ ẤN PHẨM KHOA	58	2
TẤT CẢ NHỮNG GÌ	57	25
CÓ LIÊN QUAN ĐẾN	56	10
XÃ HỘI DÂN SỰ	56	5
CÂU HỎI NGHIÊN CỨU	55	2
ĐẠI HỌC QUỐC GIA	55	5
KHOA HỌC XÃ HỘI	55	8
CHỈ SỐ TRÍCH DẪN	54	2
CÁC NHÀ NGHIÊN CỨU	53	11
ĐÂY LÀ LẦN ĐẦU	53	28
KHOA HỌC VIỆT NAM	53	3
TỪ NĂM # ĐẾN	52	12

CÁC TẬP SAN QUỐC	51	2
QUAN TRỌNG NHẤT LÀ	51	17
MỘT NHÀ KHOA HỌC	50	3
NHƯNG TRONG THỰC TẾ	50	4
VĂN HỌC NGHỆ THUẬT	50	10
ANH EM NHÀ {Name}	49	1
CÁC ANH CHỊ EM	49	11
LẦN ĐẦU TIÊN TÔI	49	18
NGƯỜI MỸ GỐC VIỆT	49	8
TRONG THỜI GIAN #	49	3
{Name} VÀ {Name} {Name}	48	1
TRẢ LỜI CÂU HỎI	48	15
KẾT QUẢ NGHIÊN CỨU	47	4
LÚC NÀO CŨNG CÓ	47	22
ỦY HỘI SÔNG MEKONG	47	2
CÓ THỂ NÓI RẰNG	46	6
ĐÓ KHÔNG PHẢI LÀ	46	12
HỆ THỐNG NGÂN HÀNG	46	4
NẾU KHÔNG MUỐN NÓI	46	10
SỰ PHÁT TRIỂN CỦA	46	10
CÁC CUỘC BIỂU TÌNH	45	6
DOANH NGHIỆP NHÀ NƯỚC	45	3
KHÔNG BIẾT BAO NHIÊU	45	24
LUẬN ÁN TIẾN SĨ	45	4
CHỦ NGHĨA XÃ HỘI	44	4
KỶ NIỆM # NĂM	44	16
MỘT LÚC NÀO ĐÓ	44	10
TÀI LIỆU THAM KHẢO	44	7
# BÀI BÁO KHOA	43	2
CÓ THỂ XEM LÀ	43	4
KHÔNG HIỂU TẠI SAO	43	21

KHÔNG PHẢI CHỈ CÓ	43	8
MỘT CÁI GÌ ĐÓ	43	18
MỘT THỜI GIAN DÀI	43	21
# NĂM VỀ TRƯỚC	42	17
BẤT CỨ LÚC NÀO	42	22
CHẾ ĐỘ ĐỘC TÀI	42	6
CHIẾN TRANH VIỆT NAM	42	6
THU NHẬP BÌNH QUÂN	42	4
CẦM QUYỀN VIỆT NAM	41	3
ĐÓ LÀ CHƯA KỂ	41	13
PHÁT TRIỂN KINH TẾ	41	11
VĂN HỌC VIỆT NAM	41	5
CHẤT LƯỢNG NGHIÊN CỨU	40	2
CHỦ NGHĨA TƯ BẢN	40	7
ĐÃ TRỞ THÀNH MỘT	40	17
ĐƯỢC XEM LÀ MỘT	40	8
KHÔNG CÓ THỜI GIAN	40	22
MỘT NGƯỜI PHỤ NỮ	40	21
NHÀ VĂN VIỆT NAM	40	6
TẦN SỐ TRÍCH DẪN	40	2
CÁCH ĐÂY # NĂM	39	19

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
# # # # #	282	3
# THÁNG # NĂM #	95	21
NGÀY # THÁNG # NĂM	88	19
ĐỂ HÔN EM LẦN NỮA	58	1
SỐ ẤN PHẨM KHOA HỌC	58	2
CÁC TẬP SAN QUỐC TẾ	51	2

TRÊN CÁC TẬP SAN QUỐC	51	2
# BÀI BÁO KHOA HỌC	43	2
ĐÂY LÀ LẦN ĐẦU TIÊN	41	25
SỐ BÀI BÁO KHOA HỌC	39	2
CÁC TẬP SAN KHOA HỌC	38	3
NHÀ CẦM QUYỀN VIỆT NAM	37	3
BIỂU TÌNH CHỐNG TRUNG QUỐC	36	4
HỌC TRÊN CÁC TẬP SAN	36	2
NẾU KHÔNG MUỐN NÓI LÀ	36	7
KHOA HỌC TRÊN CÁC TẬP	35	2
MỘT CÔNG TRÌNH NGHIÊN CỨU	35	5
THÀNH PHỐ HỒ CHÍ MINH	35	12
VÀO NGÀY # THÁNG #	35	13
VIẾT ĐỀ CƯƠNG NGHIÊN CỨU	35	2
TÀI TRỢ CHO NGHIÊN CỨU	34	2
HỘI NHÀ VĂN VIỆT NAM	33	4
BAN CHẤP HÀNH TRUNG ƯƠNG	31	2
VIỆT NAM VÀ TRUNG QUỐC	31	3
NHÂN QUYỀN Ở VIỆT NAM	30	2
ANH {Name} VÀ CHỊ {Name}	29	1
KHẮP NƠI TRÊN THẾ GIỚI	29	16
BÀI BÁO KHOA HỌC TRÊN	28	2
CHỦ NGHĨA TƯ BẢN THÂN	28	1
NGHĨA TƯ BẢN THÂN HỮU	28	1
BỐ TRÊN CÁC TẬP SAN	27	2
CÂU HỎI ĐẶT RA LÀ	27	4
CÔNG BỐ TRÊN CÁC TẬP	27	2
VẤN ĐỀ LIÊN QUAN ĐẾN	27	9
BÁO KHOA HỌC TRÊN CÁC	26	2
CÁC PHƯƠNG TIỆN TRUYỀN THÔNG	26	8
CÔNG BỐ ĐƯỢC # BÀI	26	2

HOÀNG SA VÀ TRƯỜNG SA	26	4
KHOA HỌC VÀ CÔNG NGHỆ	26	3
NỀN CÔNG NGHIỆP VĂN HÓA	26	2
# NĂM TRỞ LẠI ĐÂY	25	5
CÚN VÀ CHỊ CHUỘT NHẮT	25	1
ĐẢNG CỘNG SẢN VIỆT NAM	25	5
VIẾT BÀI BÁO KHOA HỌC	25	2
BÀI DỰ THI SỐ #	24	1
KHÔNG PHẢI LÚC NÀO CŨNG	24	17
LÀ MỘT TRONG NHỮNG NGƯỜI	24	19
SÁNG KIẾN HẠ LƯU MEKONG	24	1
TỪ NĂM # ĐẾN #	24	7
{Name} {Name} VÀ {Name} {Name}	23	1
GIỚI LÃNH ĐẠO VIỆT NAM	23	1
LÀM NGHIÊN CỨU KHOA HỌC	22	2
MỌI NGƯỜI ĐỀU CÓ QUYỀN	21	2
TRONG THỜI GIAN GẦN ĐÂY	21	7
ẤN PHẨM KHOA HỌC CỦA	20	2
CHÂU Á THÁI BÌNH DƯƠNG	20	5
CHO NGHIÊN CỨU KHOA HỌC	20	2
CŨNG LÀ MỘT TRONG NHỮNG	20	9
GIÁO SƯ VÀ PHÓ GIÁO	20	2
KHOA HỌC CỦA VIỆT NAM	20	3
KHÔNG BIẾT BAO NHIÊU LẦN	20	17
NHƯ MỘT NGÔN NGỮ THỨ	20	1
SƯ VÀ PHÓ GIÁO SƯ	20	2
TẤT CẢ MỌI NGƯỜI ĐỀU	20	13
TRONG THỜI GIAN # NĂM	20	2
BÂY GIỜ LÀ THÁNG #	19	2
CÁC THẾ LỰC THÙ ĐỊCH	19	5
ĐÂY LÀ MỘT TRONG NHỮNG	19	11

ĐỐI THOẠI VỀ NHÂN QUYỀN	19	1
EM LÀM ƠN IM ĐI	19	6
LẦN ĐẦU TIÊN TRONG ĐỜI	19	12
NHƯNG TRONG THỰC TẾ THÌ	19	2
NHỮNG VẤN ĐỀ LIÊN QUAN	19	4
THÁI LAN VÀ MÃ LAI	19	2
TRÊN CÁC TẬP SAN KHOA	19	3
TRONG NGHIÊN CỨU KHOA HỌC	19	2
ANH {Name} VÀ {Name} {Name}	18	1
BỐ ĐƯỢC # BÀI BÁO	18	2
CÓ CÔNG BỐ QUỐC TẾ	18	2
ĐIẾU CÀY NGUYỄN VĂN HẢI	18	2
GIÁO DỤC VÀ ĐÀO TẠO	18	7
HƠN # TIẾNG ĐỒNG HỒ	18	9
MỖI NGÀY MỘT TẤM HÌNH	18	1
MỘT CÂU HỎI NGHIÊN CỨU	18	2
NGÀY # THÁNG # VỪA	18	4
NHỮNG CÔNG TRÌNH NGHIÊN CỨU	18	3
Ở VIỆT NAM HIỆN NAY	18	5
TIẾNG VIỆT NHƯ MỘT NGÔN	18	1
VIỆT NHƯ MỘT NGÔN NGỮ	18	1
CÁC CHUYÊN GIA BÌNH DUYỆT	17	2
CÁC CÔNG TRÌNH NGHIÊN CỨU	17	3
CHẤP HÀNH TRUNG ƯƠNG ĐẢNG	17	1
CÓ CHUYỆN GÌ XẢY RA	17	12
CÓ THỂ CHẤP NHẬN ĐƯỢC	17	7
CỦA NGHIÊN CỨU KHOA HỌC	17	2
ĐƯỢC # BÀI BÁO KHOA	17	2
GIỮA VIỆT NAM VÀ TRUNG	17	2
LÀ LẦN ĐẦU TIÊN TÔI	17	11
MỘT BỘ PHẬN KHÔNG NHỎ	17	6

APPENDIX E

BLOG CORPUS A-CURVE CHARTS

Chart 1a: Blog Corpus 1-Tiếng Chart

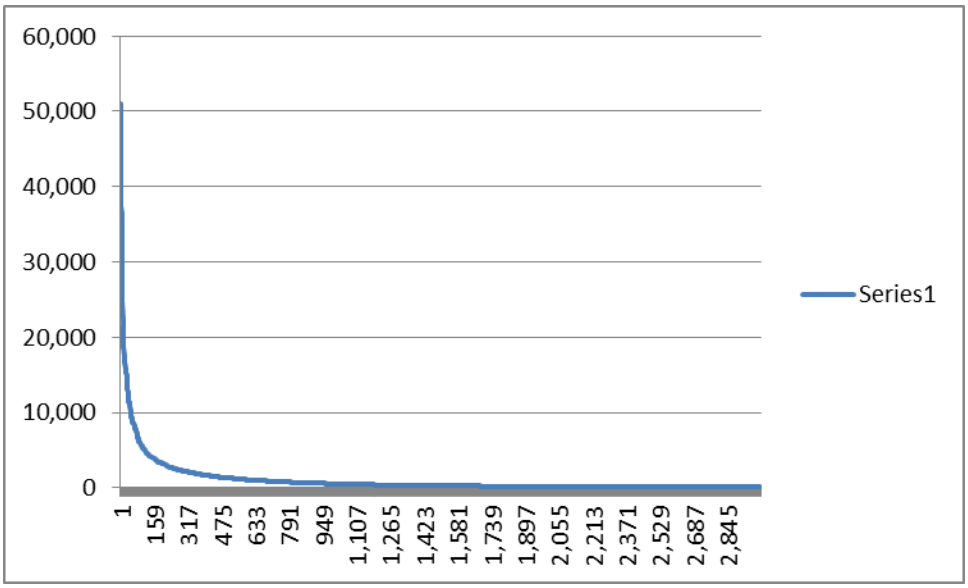


Chart 1b: Blog Corpus 2-Tiếng Chart

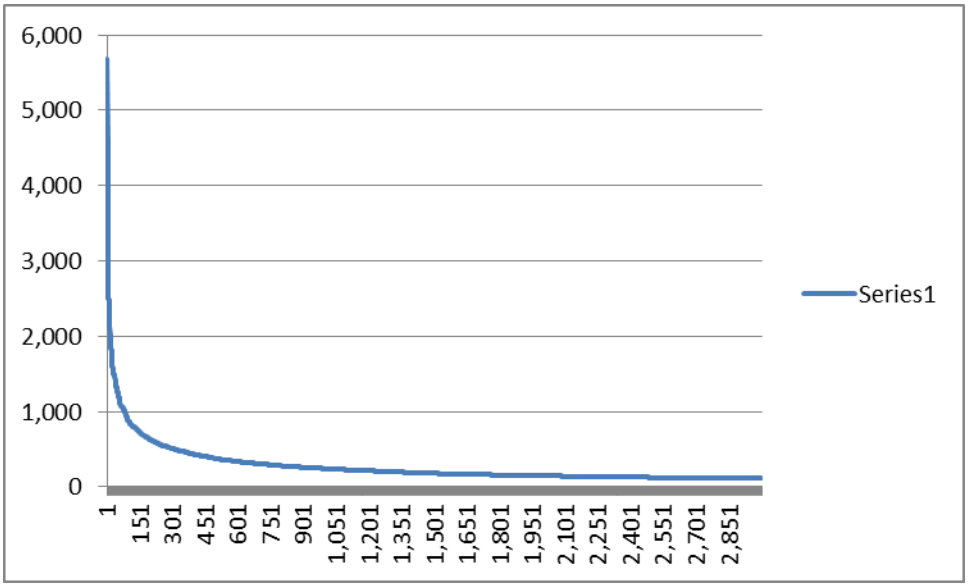


Chart 1c: Blog Corpus 3-Tiếng Chart

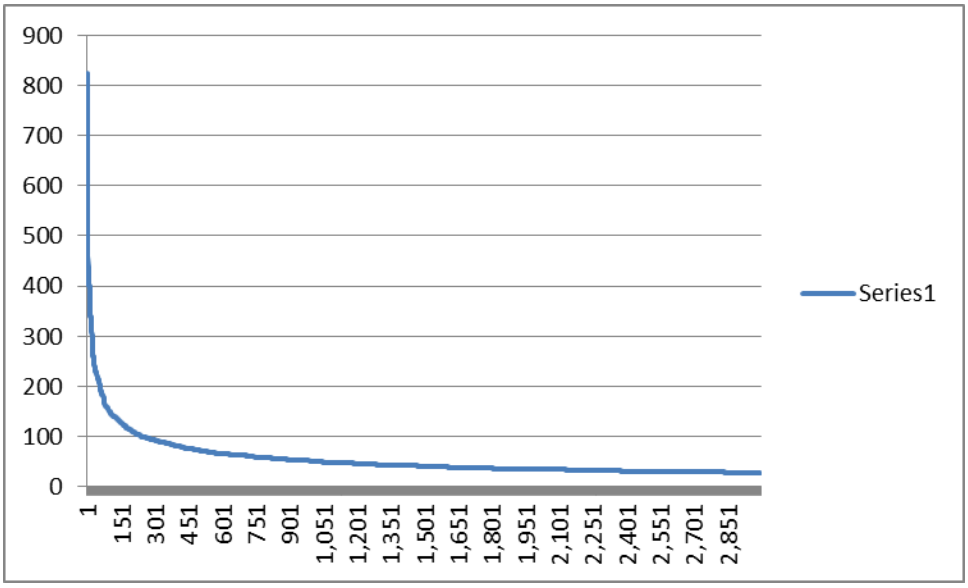


Chart 1d: Blog Corpus 4-Tiếng Chart

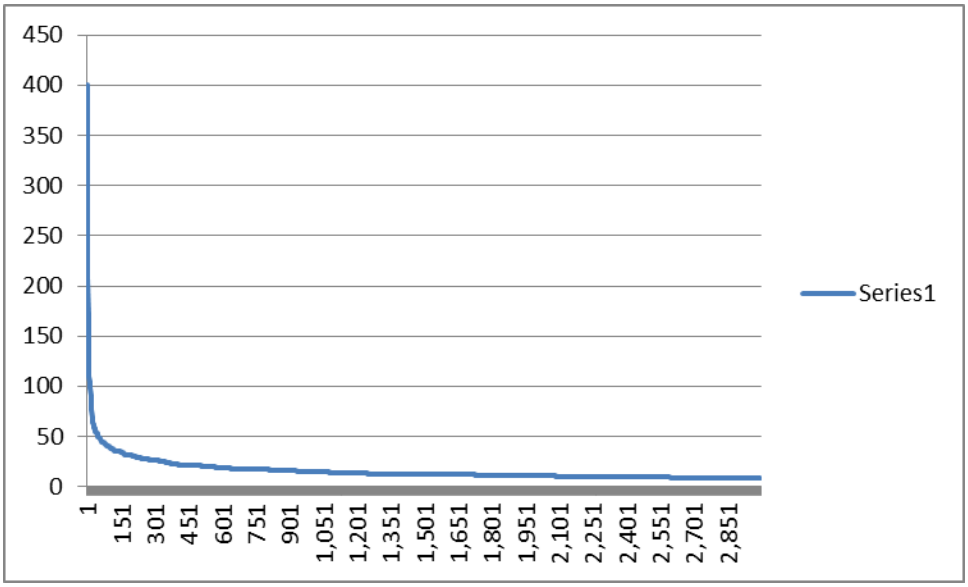
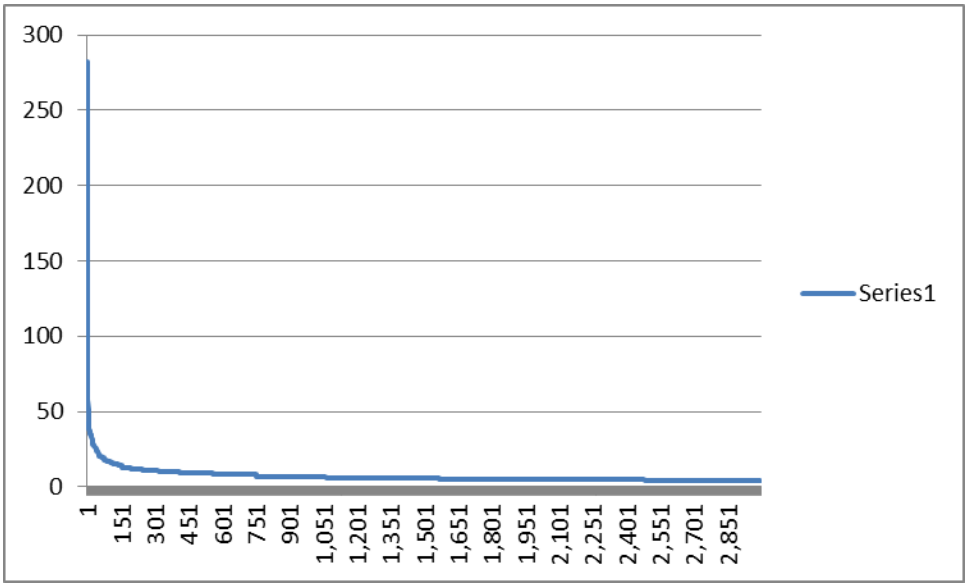


Chart 1e: Blog Corpus 5-Tiếng Chart



APPENDIX F

COMMENTS CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
LÀ	41,938	70
CÓ	39,029	70
KHÔNG	32,027	69
#	28,477	68
CỦA	24,977	67
VÀ	22,861	68
MÀ	21,727	68
CHO	21,723	67
THÌ	21,010	69
NGƯỜI	20,577	66
EM	20,177	65
CŨNG	18,655	67
ANH	16,940	65
ĐƯỢC	15,557	70
MỘT	15,302	68
CÁI	14,609	66
NÀY	14,485	70
CHỊ	14,291	65
LÀM	14,103	65
ĐÓ	13,823	67
NHƯ	13,790	65
CON	13,233	66
ĐI	12,804	65
ĐỂ	12,130	66

PHẢI	11,947	66
MÌNH	11,671	67
NHỮNG	11,636	67
DÂN	11,284	51
ĐÃ	11,253	67
VỚI	11,230	66
CÒN	11,132	66
NÓI	11,106	65
TRONG	10,973	66
LẠI	10,860	67
RỒI	10,852	64
RA	10,795	67
Ở	10,574	65
BẠN	9,959	67
GÌ	9,740	68
CHỈ	9,551	68
BIẾT	9,529	66
NHƯNG	9,493	69
NHIỀU	9,391	67
QUÁ	9,306	65
NHÀ	9,244	65
NƯỚC	9,221	57
THẾ	9,202	62
CÁC	9,090	60
VỀ	9,085	66
NÀO	8,867	67
ÔNG	8,788	59
HAY	8,747	66
THẤY	8,605	66
KHI	8,353	66
MỚI	8,336	65

VẬY	8,187	64
TA	8,106	60
VN	7,972	55
NÊN	7,839	66
SAO	7,418	66
VÌ	7,346	65
ĂN	7,027	60
ĐẾN	6,724	66
AI	6,379	58
CẢ	6,367	64
NGÀY	6,295	65
QUỐC	6,281	41
SẼ	6,276	67
BỊ	6,265	59
TỪ	6,225	62
NĂM	6,180	60
THÔI	6,178	65
HỌ	6,152	52
RẤT	6,035	66
ĐÂU	5,984	61
VIỆT	5,961	53
ƠN	5,939	62
QUA	5,900	65
CHÍNH	5,897	54
HƠN	5,886	65
CẢM	5,838	60
GIỜ	5,733	65
TÔI	5,708	57
VÀO	5,678	64
NÓ	5,665	61
NAY	5,640	65

SỰ	5,639	57
NAM	5,473	52
BÀI	5,448	63
HỌC	5,436	57
MẤY	5,328	61
VUI	5,317	60
LUÔN	5,295	60
NỮA	5,229	63
LẮM	5,153	60
TRÊN	5,145	58
ĐỌC	5,141	66
CÔNG	5,086	60
THỂ	4,998	60
NHÂN	4,965	54

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
KHÔNG CÓ	3,285	58
CẢM ƠN	3,201	54
VIỆT NAM	2,523	45
CÓ THỂ	2,469	56
CÁM ƠN	2,334	53
KHÔNG PHẢI	1,975	58
BÂY GIỜ	1,965	55
NGƯỜI TA	1,963	53
LÀ MỘT	1,804	49
KHÔNG BIẾT	1,753	56
NHƯ VẬY	1,731	49
DÂN CHỦ	1,685	12

THẾ GIỚI	1,685	42
CHỈ CÓ	1,636	54
NHƯ THẾ	1,621	48
XÃ HỘI	1,598	36
CŨNG CÓ	1,577	50
NHỮNG NGƯỜI	1,576	46
GIA ĐÌNH	1,564	52
ĐÓ LÀ	1,541	48
CÁC BẠN	1,512	42
TRUNG QUỐC	1,467	16
HẠNH PHÚC	1,423	52
DÂN TỘC	1,395	25
ĐẤT NƯỚC	1,374	24
NÀO CŨNG	1,357	55
# NĂM	1,352	49
HI HI	1,321	42
BAO GIỜ	1,293	52
NHÂN DÂN	1,289	16
ĐÚNG LÀ	1,276	50
CHÍNH TRỊ	1,266	16
TỰ DO	1,253	38
CỦA MÌNH	1,236	52
VĂN HÓA	1,196	30
CÓ MỘT	1,176	51
CỘNG SẢN	1,174	13
LÃNH ĐẠO	1,148	18
BÀI VIẾT	1,147	50
CHÚNG TA	1,138	48
CON NGƯỜI	1,133	45
MÀ KHÔNG	1,133	52
KHÔNG THỂ	1,123	49

NHÀ NƯỚC	1,119	26
CŨNG KHÔNG	1,109	47
KINH TẾ	1,106	30
HA HA	1,101	36
CHẾ ĐỘ	1,092	16
CHỈ LÀ	1,086	53
EM CŨNG	1,084	52
LÀ NGƯỜI	1,065	47
CHÚC MỪNG	1,059	50
NGƯỜI DÂN	1,052	20
THỜI GIAN	1,047	54
THÂN ÁI	1,043	10
CÓ NHIỀU	1,036	47
NHẤT LÀ	1,019	51
CÓ #	1,001	51
NGƯỜI VIỆT	1,001	34
CÁI GÌ	974	47
VẤN ĐỀ	973	41
TẤT CẢ	969	49
MỌI NGƯỜI	961	46
CỦA ANH	947	48
NĂM #	946	45
THẾ NÀO	945	53
LÀM SAO	941	49
PHẢI LÀ	940	52
GỌI LÀ	931	49
VUI VẺ	923	41
BAO NHIÊU	905	45
HÔM NAY	898	55
THAM NHŨNG	890	6
CỦA NGƯỜI	883	45

LÀM GÌ	881	43
CŨNG LÀ	879	49
CÓ NHỮNG	841	53
QUỐC GIA	841	21
LÀ NHỮNG	836	42
AI CŨNG	825	44
CHÍNH QUYỀN	825	15
CHỨ KHÔNG	822	47
TẠI SAO	812	47
CHIA SẺ	810	43
CHÚC ANH	809	28
BÀI THƠ	795	35
THÌ KHÔNG	783	49
CUỘC SỐNG	780	37
CỦA ÔNG	775	37
HÌ HÌ	770	42
RẤT NHIỀU	769	47
CÓ NGƯỜI	754	48
CHIẾN TRANH	751	20
ÔNG {Name}	751	2
HOA KỲ	749	7
TRẢ LỜI	743	44
ƠN EM	742	33
CÓ LẼ	737	50
CÓ GÌ	733	49
Ở VN	731	46

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
KHÔNG PHẢI LÀ	601	46
CẢM ƠN EM	579	18
HI HI HI	577	22
THÂN ÁI GỬI	512	2
BẠN DÂN CHỦ	501	2
HA HA HA	493	27
CẢM ƠN ANH	470	27
CÁC BẠN DÂN	468	2
LÚC NÀO CŨNG	420	43
KHÔNG BAO GIỜ	419	37
CHỨ KHÔNG PHẢI	354	35
NHƯ THẾ NÀO	343	39
BÀI VIẾT CỦA	332	32
VÀ GIA ĐÌNH	324	33
TRÊN THẾ GIỚI	291	20
CÁM ƠN BẠN	278	29
HÌ HÌ HÌ	268	19
ĐÃ GHÉ THĂM	266	8
{Last} {Middle} {First}	255	2
VÀ HẠNH PHÚC	252	29
CÁM ƠN ANH	247	29
CÁM ƠN CHỊ	242	30
HỒ CHÍ MINH	236	7
ĐỒNG Ý VỚI	235	35
# MẸ CON	234	17
KHÔNG CÓ GÌ	229	34
NÀO CŨNG CÓ	226	34
ĐÃ CHIA SẺ	224	19
CHƯA BAO GIỜ	220	31

ANH {Name} {Name}	218	3
CÓ NGHĨA LÀ	217	26
CẢM ƠN BẠN	208	27
CHÙA BÀ ĐANH	205	6
NHƯ THẾ NÀY	202	34
ƠN EM ĐÃ	195	11
CŨNG CÓ THỂ	190	32
NỀN KINH TẾ	185	9
NGƯỜI VIỆT NAM	182	15
Ở VIỆT NAM	177	24
DÂN VIỆT NAM	176	10
CẢM ƠN CHỊ	174	24
ĐẢNG CỘNG SẢN	174	6
HE HE HE	169	23
XÃ HỘI CHỦ	168	5
CỦA DÂN TỘC	167	11
HỘI CHỦ NGHĨA	166	5
CÁM ƠN DÌ	164	3
CÁI GÌ CŨNG	163	33
CÀNG NGÀY CÀNG	160	29
BÀI VIẾT NÀY	158	30
LÀ NHỮNG NGƯỜI	158	22
CÓ THỜI GIAN	156	34
Ý KIẾN CỦA	156	18
DÂN TỘC VN	155	5
CHÚC EM LUÔN	154	6
HAPPY NEW YEAR	153	30
CẢM ƠN BÁC	152	16
CÁM ƠN EM	152	19
THÌ LÀM SAO	152	25
VUI VẺ VÀ	151	21

CHỈ LÀ MỘT	150	28
NHÂN DÂN VN	148	3
CÓ PHẢI LÀ	147	32
CÂU TRẢ LỜI	145	23
CHO MỌI NGƯỜI	144	30
TẤT CẢ NHỮNG	144	32
NGƯỜI TA KHÔNG	143	24
LÀM GÌ CÓ	142	26
CÓ KHẢ NĂNG	141	21
CỦA CÁC BẠN	141	18
GỬI ANH {Name}	139	3
HƠN # NĂM	139	27
TRÊN DIỄN ĐÀN	138	6
CHÚC MỪNG ANH	136	16
DỄ THƯƠNG QUÁ	135	28
TRONG ĐÓ CÓ	135	27
CŨNG LÀ MỘT	134	26
TẤT CẢ CÁC	134	26
CHỐNG THAM NHŨNG	133	4
MỪNG SINH NHẬT	133	25
{Last} {Middle} {First}	133	3
CỦA NHỮNG NGƯỜI	131	22
HÍ HÍ HÍ	131	16
NHÔ HOA KỲ	131	2
ƠN ANH ĐÃ	131	16
CÁI GỌI LÀ	129	10
CỦA TRUNG QUỐC	129	6
NHIỀU NIỀM VUI	129	22
SOUTH CHINA SEA	129	3
CÓ CƠ HỘI	128	31
NHƯ VẬY THÌ	128	21

BIỂU TÌNH CHỐNG	126	5
DÂN TỘC VIỆT	126	5
MÀ KHÔNG CÓ	126	27
CÓ NHIỀU NGƯỜI	125	30
CHỈ CÓ #	124	27
LẦN ĐẦU TIÊN	124	30
{Last} {Middle} {First}	124	2
CỦA NHÀ NƯỚC	123	11
CỦA ÔNG {Name}	123	2

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
CÁC BẠN DÂN CHỦ	465	2
ANH {Last} {Middle} {First}	215	2
CẢM ƠN EM ĐÃ	177	5
XÃ HỘI CHỦ NGHĨA	161	5
CHÚC MỪNG SINH NHẬT	108	23
CỘNG SẢN VIỆT NAM	102	4
TỰ DO DÂN CHỦ	102	5
CẢM ƠN ANH ĐÃ	99	7
GỬI ANH {Last} {Middle}	95	2
ĐÃ GHÉ THĂM VÀ	93	4
CHÚC EM LUÔN VUI	92	3
THẾ LỰC THÙ ĐỊCH	91	5
BÀI VIẾT CỦA ANH	88	12
ĐẢNG VÀ NHÀ NƯỚC	87	5
ANH {Last} {Middle} {First}	85	2
CHÚC MỪNG NĂM MỚI	83	27
GỬI ANH {Last} {Middle}	83	2

NHÂN DÂN VIỆT NAM	77	4
CHẾ ĐỘ CỘNG SẢN	74	3
KHỎE VÀ HẠNH PHÚC	73	6
CHỨ KHÔNG PHẢI LÀ	72	20
BUỔI TỐI VUI VẺ	71	2
CHẾ ĐỘ ĐỘC TÀI	70	4
CHIA SẺ THÂN ÁI	70	1
GỬI {Last} {Middle} {First}	70	1
EM ĐÃ CHIA SẺ	68	4
THÂN ÁI GỬI ANH	68	2
THÂN ÁI GỬI {Last}	68	1
DÂN TỘC VIỆT NAM	67	4
NƯỚC ÚC PHÁT THÈM	67	1
ƠN EM ĐÃ CHIA	67	4
TAM QUYỀN PHÂN LẬP	67	2
GỬI {Last} {Middle} {First}	66	1
KHÔNG BAO GIỜ CÓ	66	15
CẢM ƠN ANH NHIỀU	65	10
Á HI HI HI	64	6
BẠN DÂN CHỦ THÌ	61	2
CỦA CÁC BẠN DÂN	61	2
ĐỘC LẬP DÂN TỘC	61	3
HẠNH PHÚC THÂN ÁI	61	1
LÀ MỘT TRONG NHỮNG	61	20
ĐỘC LẬP TỰ DO	59	5
GỬI {Last} {Middle} {First}	59	1
CẢM ƠN EM NHIỀU	58	5
ỦNG HỘ TINH THẦN	58	6
BÀI VIẾT CỦA ÔNG	57	3
CẢM ƠN EM CHÚC	57	2
EM LUÔN VUI KHỎE	57	1

HUGS # MẸ CON	57	7
TRƯNG CẦU DÂN Ý	56	4
CẢM ƠN BẠN ĐÃ	55	16
CHÚC ANH NGÀY MỚI	55	2
CẢI CÁCH RUỘNG ĐẤT	54	4
ANH VÀ GIA ĐÌNH	53	12
CHỦ NGHĨA CỘNG SẢN	53	6
ĐÃ CHIA SẺ THÂN	53	1
LÀM TAY SAI CHO	53	3
ƠN EM CHÚC EM	53	2
CÁM ƠN DÌ NHIỀU	52	3
CHIA SẺ CÙNG ANH	52	3
CHÚ {Last} {Middle} {First}	52	1
GỬI {Last} {Middle} {First}	52	1
KINH TẾ THỊ TRƯỜNG	52	5
NGÀY # THÁNG #	52	16
TỰ DO NGÔN LUẬN	52	6
GỬI CHÚ {Last} {Middle}	51	1
GỬI NGƯỜI THÁI BÌNH	51	1
TẤT CẢ MỌI NGƯỜI	51	19
{Name} {Name} {Name} {Name}	50	1
CUỐI TUẦN VUI VẺ	50	13
VUI VẺ THÂN ÁI	50	1
BÀI THƠ CỦA ANH	49	3
ĐẢNG CỘNG SẢN VIỆT	49	3
GỬI {Name} {Name} {Name}	49	1
CÁM ƠN BẠN ĐÃ	48	9
EM ĐÃ GHÉ THĂM	48	4
GỬI {Last} {Middle} {First}	48	1
MỚI NHIỀU NIỀM VUI	48	10
ƠN EM ĐÃ GHÉ	48	3

THĂM VÀ CHIA SẺ	48	2
CHỊ VÀ GIA ĐÌNH	47	16
HỘ TINH THẦN CHO	47	3
KHÔNG CÓ NGHĨA LÀ	47	14
KHÔNG PHẢI LÀ NGƯỜI	47	8
ƠN BẠN ĐÃ GHÉ	47	8
CHA TRUYỀN CON NỐI	46	5
SẼ KHÔNG BAO GIỜ	46	15
TRÊN DIỄN ĐÀN NÀY	46	3
CHÚC ANH BUỔI TỐI	45	2
CHÚC ANH LUÔN VUI	45	6
THÁNG # NĂM #	45	14
TRONG VÀ NGOÀI NƯỚC	45	4
VÀ GIA ĐÌNH MỘT	45	23
MẠNH KHỎE VÀ VUI	44	3
GHÉ THĂM VÀ CHIA	43	2
MẸ CON NHÀ {Name}	43	4
NÈ HI HI HI	43	6
NHỮNG BÀI VIẾT CỦA	43	10
GỬI NGUYỄN NGỌC CHIẾN	42	1

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
GỬI ANH {Last} {Middle} {First}	93	2
GỬI ANH {Last} {Middle} {First}	81	2
ƠN EM ĐÃ CHIA SẺ	67	4
CẢM ƠN EM ĐÃ CHIA	61	1
CỦA CÁC BẠN DÂN CHỦ	60	2
CÁC BẠN DÂN CHỦ THÌ	59	2
CHÚC EM LUÔN VUI KHỎE	57	1

CẢM ƠN EM CHÚC EM	53	2
ĐÃ CHIA SẺ THÂN ÁI	53	1
GỬI CHÚ {Last} {Middle} {First}	50	1
GỬI CỬA TÙNG DẤU YÊU	48	1
ỦNG HỘ TINH THẦN CHO	47	3
CẢM ƠN EM ĐÃ GHÉ	46	2
ƠN EM ĐÃ GHÉ THĂM	46	3
ĐẢNG CỘNG SẢN VIỆT NAM	45	3
GHÉ THĂM VÀ CHIA SẺ	43	2
CÁM ƠN DÌ NHIỀU NHIỀU	41	1
VÀ HẠNH PHÚC THÂN ÁI	40	1
VUI VẺ VÀ HẠNH PHÚC	40	5
HÒA XÃ HỘI CHỦ NGHĨA	38	2
MẠNH KHỎE VÀ HẠNH PHÚC	38	3
ĐÃ GHÉ THĂM VÀ CHIA	37	1
LUÔN ỦNG HỘ TINH THẦN	37	3
XÃ HỘI CHỦ NGHĨA VIỆT	37	3
ƠN BẠN ĐÃ GHÉ THĂM	36	3
CẢM ƠN EM ĐÃ ĐỒNG	35	1
CHIA SẺ THÂN ÁI GỬI	35	1
CHÚC ANH VÀ GIA ĐÌNH	34	11
CỘNG HÒA XÃ HỘI CHỦ	34	2
ƠN EM ĐÃ ĐỒNG CẢM	34	1
VÀ GIA ĐÌNH NĂM MỚI	34	12
HỘI CHỦ NGHĨA VIỆT NAM	33	2
CÁC BẠN DÂN CHỦ KHÔNG	32	2
CHÚC CHỊ VÀ GIA ĐÌNH	32	11
TỰ DO CÁI CON C	32	1
CHỈ BIẾT CÒN ĐẢNG CÒN	31	2
CHÚC ANH BUỔI TỐI VUI	31	2
MẠNH KHỎE VÀ VUI VẺ	31	3

XÂY DỰNG ĐƯỢC GÌ TRÊN	31	2
BÀI VIẾT CỦA ÔNG QUỐC	30	2
BLOG FOR YOU GREAT HTTP	30	16
CHỈ CÓ TRUNG QUỐC LÀ	30	1
CÓ TRUNG QUỐC LÀ TỐT	30	1
GỬI ANH {Last} {Middle} {First}	30	1
NGÀY MỚI NHIỀU NIỀM VUI	30	2
THANK BLOG FOR YOU GREAT	30	16
CÁC THẾ LỰC THÙ ĐỊCH	29	4
HỘ TINH THẦN CHO CÚN	29	1
KHÔNG CÓ GÌ QUÝ HƠN	29	3
THẦY TRÒ CÁC BẠN DÂN	29	2
TRÒ CÁC BẠN DÂN CHỦ	29	2
ANH BUỔI TỐI VUI VẺ	28	2
HƠN ĐỘC LẬP TỰ DO	28	4
ĂN VẠ VÀ VU VẠ	27	2
BIỂU TÌNH CHỐNG TRUNG QUỐC	27	4
CẢM ƠN ANH ĐÃ GHÉ	27	2
GÌ TRÊN SỰ DỐI TRÁ	27	2
NƯỚC CỘNG HÒA XÃ HỘI	27	2
CẢM ƠN EM RẤT NHIỀU	26	3
GHÉ THĂM VÀ ĐỒNG CẢM	26	1
HOÀN TOÀN ĐỒNG Ý VỚI	26	11
ƠN ANH ĐÃ GHÉ THĂM	26	3
CHÚC EM MỘT NGÀY MỚI	25	2
DỰNG ĐƯỢC GÌ TRÊN SỰ	25	2
KHỎE VÀ HẠNH PHÚC THÂN	25	2
ƠN EM CHÚC EM LUÔN	25	1
ANH NGÀY MỚI NHIỀU NIỀM	24	2
BÀ AUNG SAN SUU KYI	24	3
CÁM ƠN BẠN ĐÃ GHÉ	24	5

CẢM ƠN EM THÂN ÁI	24	1
CHÚC ANH NGÀY MỚI NHIỀU	24	2
CÓ GÌ QUÝ HƠN ĐỘC	24	3
ĐÃ GHÉ THĂM VÀ ĐỒNG	24	1
EM LUÔN VUI KHỎE THÂN	24	1
GÌ QUÝ HƠN ĐỘC LẬP	24	3
LUÔN VUI KHỎE THÂN ÁI	24	1
NGHỀ GÌ KHÁC ĐỂ SỐNG	24	2
# THÁNG # NĂM #	23	12
CẢM ƠN ĐỒNG HƯƠNG CHÚC	23	1
ĐÃ ĐỒNG CẢM THÂN ÁI	23	1
ĐƯỢC GÌ TRÊN SỰ DỐI	23	2
HÈN VỚI GIẶC ÁC VỚI	23	2
THỦ TƯỚNG NGUYỄN TẤN DŨNG	23	3
VÀ GIA ĐÌNH MỘT NĂM	23	17
VỚI GIẶC ÁC VỚI DÂN	23	2
XIN CHIA SẺ CÙNG ANH	23	2
CẢM ƠN BẠN ĐÃ GHÉ	22	4
CHÚC EM VÀ GIA ĐÌNH	22	7
DƯỚI SỰ LÃNH ĐẠO CỦA	22	3
GỬI ANH {Last} {Middle} {First}	22	1
BIẾT KIẾM NGHỀ GÌ KHÁC	21	2
CẢM ƠN ANH TỚI THĂM	21	1
GẬY ÔNG ĐẬP LƯNG ÔNG	21	4
HOÀNG SA VÀ TRƯỜNG SA	21	3
NGÀY NGHỈ CUỐI TUẦN VUI	21	2
THÂN ÁI GỬI ANH {Last}	21	2
VUI VẺ THÂN ÁI GỬI	21	1
BIẾT CÒN ĐẢNG CÒN TIỀN	20	2
CHỈ CÓ TQ LÀ TỐT	20	3
CHÚ {Last} {Middle} {First} CON	20	1

APPENDIX G

COMMENTS CORPUS A-CURVE CHARTS

Chart 1a: Comment Corpus 1-Tiếng Chart

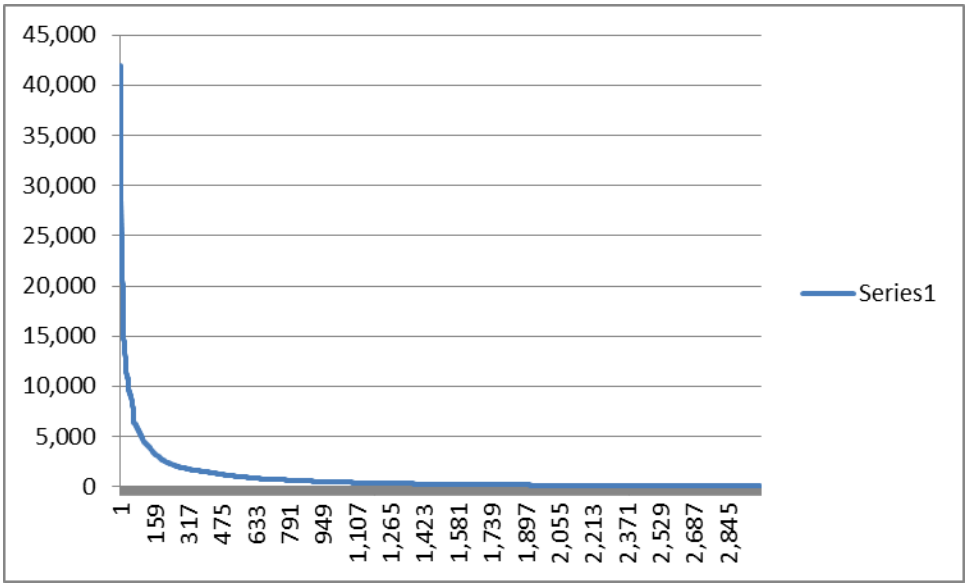


Chart 1b: Comment Corpus 2-Tiếng Chart

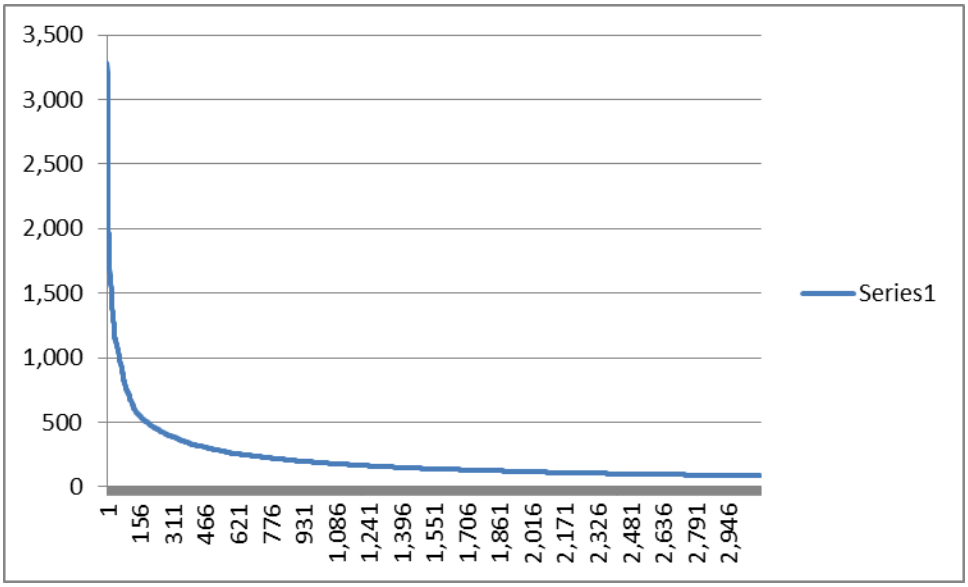


Chart 1c: Comment Corpus 3-Tiếng Chart

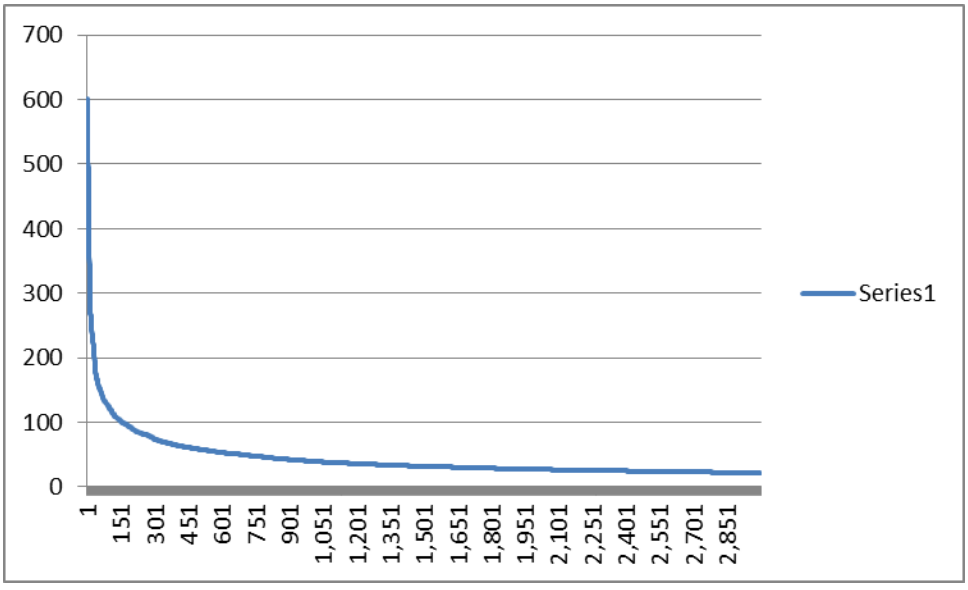


Chart 1d: Comment Corpus 4-Tiếng Chart

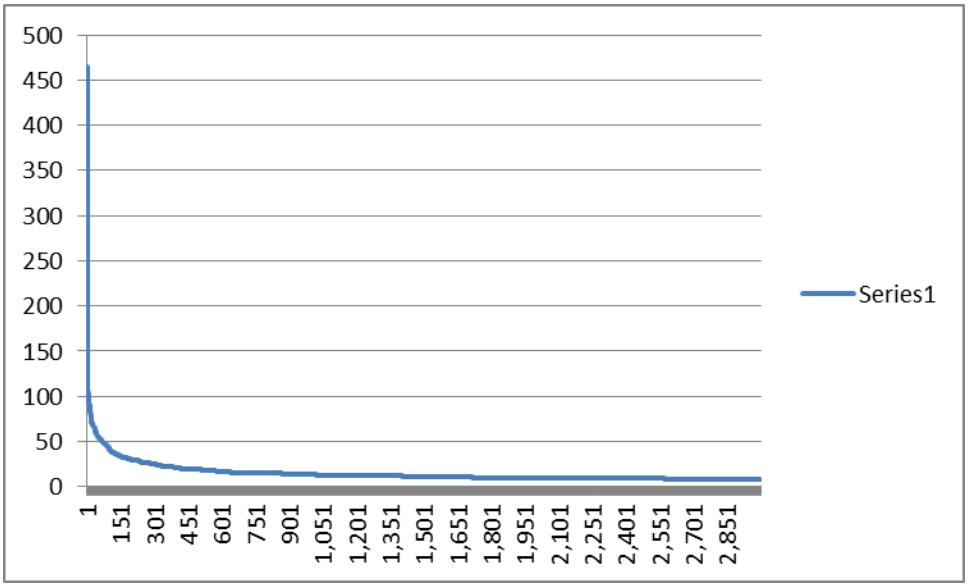
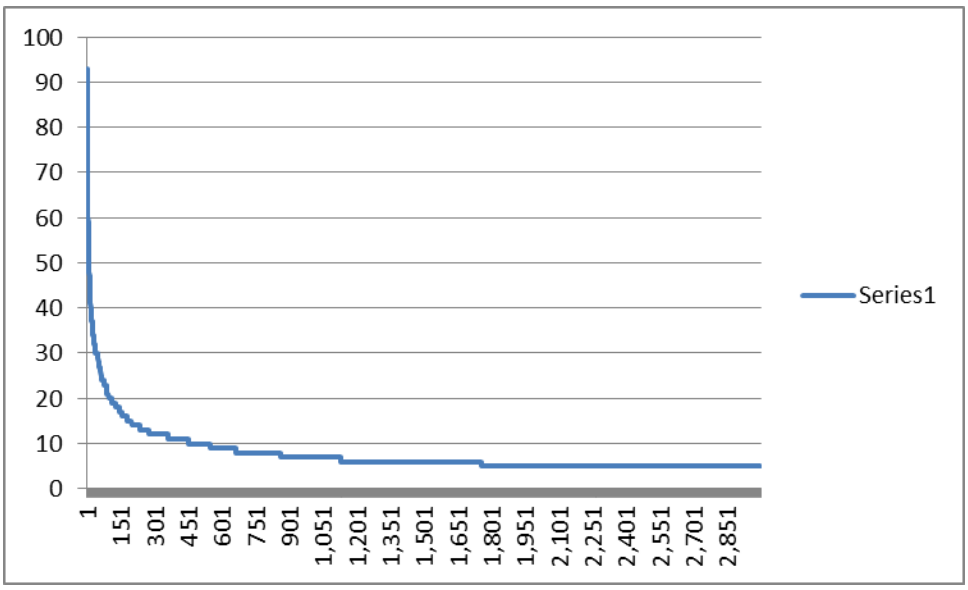


Chart 1e: Comment Corpus 5-Tiếng Chart



APPENDIX H

INTERRATER RELIABILITY CHARTS

Table 1: IRR Morphemes

HMorph * TMorph Crosstabulation

			TMorph					
			1	2	3	4	5	Total
HMorph	1	Count	302	0	0	0	0	302
		% of Total	19.9%	0.0%	0.0%	0.0%	0.0%	19.9%
	2	Count	1	299	4	0	0	304
		% of Total	.1%	19.7%	.3%	0.0%	0.0%	20.1%
	3	Count	0	1	303	0	1	305
		% of Total	0.0%	.1%	20.0%	0.0%	.1%	20.1%
	4	Count	0	0	2	302	0	304
		% of Total	0.0%	0.0%	.1%	19.9%	0.0%	20.1%
	5	Count	0	0	0	2	298	300
		% of Total	0.0%	0.0%	0.0%	.1%	19.7%	19.8%
	6	Count	0	0	0	0	1	1
		% of Total	0.0%	0.0%	0.0%	0.0%	.1%	.1%
Total		Count	303	300	309	304	300	1516
		% of Total	20.0%	19.8%	20.4%	20.1%	19.8%	100.0%

Table 1a: IRR Morphemes Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.990	.003	77.139	0.000
N of Valid Cases		1516			

Table 2: IRR Words

HWord * TWord Crosstabulation

			TWord					
			1	2	3	4	5	Total
HWord	0	Count	16	0	0	0	0	16
		% of Total	1.1%	0.0%	0.0%	0.0%	0.0%	1.1%
	1	Count	443	6	0	0	0	449
		% of Total	29.3%	.4%	0.0%	0.0%	0.0%	29.7%
	2	Count	24	418	18	2	0	462
		% of Total	1.6%	27.6%	1.2%	.1%	0.0%	30.5%
	3	Count	5	88	252	23	3	371
		% of Total	.3%	5.8%	16.7%	1.5%	.2%	24.5%
	4	Count	0	31	52	70	7	160
		% of Total	0.0%	2.0%	3.4%	4.6%	.5%	10.6%
	5	Count	0	4	20	9	22	55
		% of Total	0.0%	.3%	1.3%	.6%	1.5%	3.6%
Total		Count	488	547	342	104	32	1513
		% of Total	32.3%	36.2%	22.6%	6.9%	2.1%	100.0%

Table 2a: IRR Words Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.721	.014	48.048	0.000
N of Valid Cases		1513			

Table 3: IRR Phrases

HPhrase * TPhrase Crosstabulation

			TPhrase		
			0	1	Total
HPhrase	0	Count	524	3	527
		% of Total	35.4%	.2%	35.6%
	1	Count	89	866	955
		% of Total	6.0%	58.4%	64.4%
Total		Count	613	869	1482
		% of Total	41.4%	58.6%	100.0%

Table 3a: IRR Phrases Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.869	.013	33.719	.000
N of Valid Cases		1482			

Table 4: Clauses

HClause * TClause Crosstabulation

			TClause		
			0	1	Total
HClause	0	Count	1447	4	1451
		% of Total	96.7%	.3%	96.9%
	1	Count	14	32	46
		% of Total	.9%	2.1%	3.1%
Total		Count	1461	36	1497
		% of Total	97.6%	2.4%	100.0%

Table 4a: Clauses Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.774	.052	30.201	.000
N of Valid Cases		1497			

Table 5: Sentences

HSentence * TSentence Crosstabulation

			TSentence		
			0	1	Total
HSentence	0	Count	1457	0	1457
		% of Total	97.5%	0.0%	97.5%
	1	Count	29	8	37
		% of Total	1.9%	.5%	2.5%
Total		Count	1486	8	1494
		% of Total	99.5%	.5%	100.0%

Table 5a: Sentences Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.350	.091	17.797	.000
N of Valid Cases		1494			

Table 6: Open Class

HOpen * TOpen Crosstabulation

			TOpen		
			0	1	Total
HOpen	0	Count	323	98	421
		% of Total	21.5%	6.5%	28.0%
	1	Count	19	1065	1084
		% of Total	1.3%	70.8%	72.0%
Total		Count	342	1163	1505
		% of Total	22.7%	77.3%	100.0%

Table 6a: Open Class Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.795	.018	31.153	.000
N of Valid Cases		1505			

Table 7: Closed Class

HClosed * Tclosed Crosstabulation

			Tclosed		
			0	1	Total
HClosed	0	Count	1020	13	1033
		% of Total	67.9%	.9%	68.7%
	1	Count	115	355	470
		% of Total	7.7%	23.6%	31.3%
Total		Count	1135	368	1503
		% of Total	75.5%	24.5%	100.0%

Table 7a: Closed Class Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.789	.018	31.045	.000
N of Valid Cases		1503			

Table 8: Content Forms

HContent * Tcontent Crosstabulation

			Tcontent		
			0	1	Total
HContent	0	Count	313	37	350
		% of Total	20.9%	2.5%	23.4%
	1	Count	20	1126	1146
		% of Total	1.3%	75.3%	76.6%
Total		Count	333	1163	1496
		% of Total	22.3%	77.7%	100.0%

Table 8a: Content Forms Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.892	.014	34.514	.000
N of Valid Cases		1496			

Table 9: Function Forms

HFunction * Tfunction Crosstabulation

			Tfunction		
			0	1	Total
HFunction	0	Count	1069	10	1079
		% of Total	71.6%	.7%	72.2%
	1	Count	54	361	415
		% of Total	3.6%	24.2%	27.8%
Total		Count	1123	371	1494
		% of Total	75.2%	24.8%	100.0%

Table 9a: Function Forms Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.890	.013	34.486	.000
N of Valid Cases		1494			

Table 10: Free Forms

HFree * Tfree Crosstabulation

			Tfree		
			0	1	Total
HFree	0	Count	51	37	88
		% of Total	3.4%	2.5%	5.9%
	1	Count	7	1399	1406
		% of Total	.5%	93.6%	94.1%
Total		Count	58	1436	1494
		% of Total	3.9%	96.1%	100.0%

Table 10a: Free Forms Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.684	.045	27.068	.000
N of Valid Cases		1494			

Table 11: Bound Forms

HBound * Tbound Crosstabulation

			Tbound		
			0	1	Total
HBound	0	Count	1397	4	1401
		% of Total	93.5%	.3%	93.8%
	1	Count	33	60	93
		% of Total	2.2%	4.0%	6.2%
Total		Count	1430	64	1494
		% of Total	95.7%	4.3%	100.0%

Table 11a: Bound Forms Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.752	.039	29.622	.000
N of Valid Cases		1494			

Table 12: Units of Meaning

HUnits * TUnits Crosstabulation

			TUnits					Total
			1	2	3	4	5	
HUnits	1	Count	463	35	4	0	0	502
		% of Total	30.6%	2.3%	.3%	0.0%	0.0%	33.2%
	2	Count	39	423	21	0	0	483
		% of Total	2.6%	28.0%	1.4%	0.0%	0.0%	31.9%
	3	Count	0	80	244	11	2	337
		% of Total	0.0%	5.3%	16.1%	.7%	.1%	22.3%
	4	Count	0	10	48	86	4	148
		% of Total	0.0%	.7%	3.2%	5.7%	.3%	9.8%
	5	Count	2	1	4	12	23	42
		% of Total	.1%	.1%	.3%	.8%	1.5%	2.8%
Total		Count	504	549	321	109	29	1512
		% of Total	33.3%	36.3%	21.2%	7.2%	1.9%	100.0%

Table 12a: Units of Meaning Kappa & Significance

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of	Kap	.749	.014	48.466	0.000
N of Valid Cases		1512			

APPENDIX I

FULL CORPUS CONTENT AND FUNCTION FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Content Forms	1-Tiếng Function Forms
#	AI
ĂN	ANH
ANH	BÀI
BÀI	BỊ
BẠN	CẢ
BỊ	CÁC
BIẾT	CÁI
CÁI	CHỈ
CHỈ	CHỊ
CHỊ	CHO
CHÍNH	CÓ
CHO	CON
CÓ	CÒN
CON	CỦA
CÔNG	CŨNG
DÂN	ĐÃ
ĐẦU	ĐẾN
ĐÂY	ĐƯỢC
ĐỂ	EM
ĐẾN	GÌ
ĐI	HAY
ĐÓ	HƠN
EM	KHI
GIỜ	KHÔNG
HAI	LÀ
HAY	LẠI
HỌ	LÊN
HỌC	MÀ
KHÔNG	MÌNH

LÀ	NÀO
LÀM	NAY
LÊN	NÀY
MẤY	NÊN
MẸ	NHƯ
MÌNH	NHƯNG
MỚI	NHỮNG
MỘT	NÓ
NAM	NỮA
NĂM	Ở
NGÀY	ÔNG
NGƯỜI	PHẢI
NHÀ	QUA
NHÂN	QUÁ
NHIỀU	RA
NHƯ	RỒI
NÓ	SAO
NÓI	SAU
NƯỚC	SẼ
Ở	SỰ
ÔNG	TA
QUA	THẾ
QUỐC	THEO
RA	THÌ
RẤT	THÔI
SAO	TÔI
SỰ	TRÊN
TA	TRONG
THÀNH	TỪ
THẤY	VÀ
THỂ	VẪN
TÔI	VÀO
VÀO	VẬY
VỀ	
VIỆT	

List 1b: 2-Tiếng Forms

2-Tiếng Content Forms	2-Tiếng Function Forms
# #	AI CŨNG
# NĂM	BAO GIỜ
BẮT ĐẦU	BAO NHIÊU
BÂY GIỜ	CÁI GÌ
CÁC BẠN	CHỨ KHÔNG
CÁM ƠN	CÓ LẼ
CẢM ƠN	CÓ THỂ
CHỈ CÓ	CỦA CÁC
CHỈ LÀ	CŨNG NHƯ
CHÍNH TRỊ	ĐÂY LÀ
CHÚNG TA	ĐÓ LÀ
CHÚNG TÔI	ĐÚNG LÀ
CÓ #	LÀM SAO
CÓ MỘT	MỘT CÁCH
CÓ NHIỀU	NÀO CŨNG
CÓ NHỮNG	NHẤT LÀ
CON NGƯỜI	NHƯ THẾ
CỦA ANH	NHƯ VẬY
CỦA MÌNH	SAU KHI
CỦA NGƯỜI	TẠI SAO
CỦA ÔNG	TẤT CẢ
CŨNG CÓ	THẾ NÀO
CŨNG KHÔNG	VẪN CÒN
CŨNG LÀ	
CŨNG NHƯ	
CUỘC SỐNG	
ĐÃ CÓ	
DÂN CHỦ	
DÂN TỘC	
ĐẤT NƯỚC	
ĐẦU TIÊN	
GIA ĐÌNH	
GỌI LÀ	
HÀ NỘI	
HẠNH PHÚC	
HÔM NAY	
KHOA HỌC	
KHÔNG BIẾT	
KHÔNG CÓ	

KHÔNG CÒN	
KHÔNG ĐƯỢC	
KHÔNG PHẢI	
KHÔNG THỂ	
KINH TẾ	
LÀ #	
LÀ MỘT	
LÀ NGƯỜI	
LÀ NHỮNG	
LÀM GÌ	
LÃNH ĐẠO	
MÀ KHÔNG	
MỌI NGƯỜI	
MỘT NGƯỜI	
MỘT SỐ	
NĂM #	
NĂM NAY	
NGHIÊN CỨU	
NGƯỜI DÂN	
NGƯỜI TA	
NGƯỜI VIỆT	
NHÀ NƯỚC	
NHÂN DÂN	
NHIỀU NGƯỜI	
NHƯ THẾ	
NHƯ VẬY	
NHỮNG NGƯỜI	
Ở ĐÂY	
PHẢI LÀ	
QUỐC GIA	
RẤT NHIỀU	
TÁC GIẢ	
THÁNG #	
THẾ GIỚI	
THỜI GIAN	
TÔI KHÔNG	
TRẢ LỜI	
TRUNG QUỐC	
TỰ DO	
VẤN ĐỀ	
VĂN HÓA	
VIỆT NAM	

XÃ HỘI	

List 1c: 3-Tiếng Forms

3-Tiếng Content Forms	3-Tiếng Function Forms
# # #	BAO GIỜ CŨNG
# ĐẾN #	CÀNG NGÀY CÀNG
# MẸ CON	CHO ĐẾN KHI
# NĂM **#**	CŨNG CÓ THỂ
# THÁNG #	ĐẶC BIỆT LÀ
AI CŨNG BIẾT	LÚC NÀO CŨNG
ANH EM {Name}	NÀO CŨNG CÓ
BÀI VIẾT CỦA	TẤT CẢ CÁC
BẠN CÓ THỂ	TẤT CẢ MỌI
BẠN DÂN CHỦ	TẤT CẢ NHỮNG
BÂY GIỜ THÌ	TRONG ĐÓ CÓ
BIẾT BAO NHIÊU	
CÁC BẠN DÂN	
CÁI GÌ CŨNG	
CÁI GỌI LÀ	
CÁM ƠN ANH	
CẢM ƠN ANH	
CÁM ƠN BẠN	
CÁM ƠN CHỊ	
CẢM ƠN EM	
CÂU TRẢ LỜI	
CHỈ CÓ #	
CHỈ CÓ MỘT	
CHỈ LÀ MỘT	
CHỨ KHÔNG PHẢI	
CHƯA BAO GIỜ	
CHÚNG TA CÓ	
CÓ CƠ HỘI	
CÓ KHẢ NĂNG	
CÓ NGHĨA LÀ	
CÓ RẤT NHIỀU	
CÓ THỂ LÀ	
CÓ THỂ LÀM	
CÓ THỂ NÓI	

CÓ THỜI GIAN	
CÓ Ý NGHĨA	
CỦA CHÚNG TA	
CỦA CON NGƯỜI	
CỦA NHỮNG NGƯỜI	
CỦA TRUNG QUỐC	
CỦA VIỆT NAM	
CŨNG LÀ MỘT	
CỨU KHOA HỌC	
ĐÃ GHÉ THĂM	
ĐẢNG CỘNG SẢN	
ĐÂY LÀ MỘT	
ĐÓ LÀ MỘT	
ĐỒNG Ý VỚI	
HA HA HA	
HI HI HI	
HÌ HÌ HÌ	
HỒ CHÍ MINH	
HƠN # NĂM	
KHÔNG BAO GIỜ	
KHÔNG CÓ GÌ	
KHÔNG PHẢI LÀ	
LÀ MỘT NGƯỜI	
LÀ MỘT TRONG	
LÀ NHỮNG NGƯỜI	
LẦN ĐẦU TIÊN	
LIÊN QUAN ĐẾN	
MỘT THỜI GIAN	
MỘT TRONG NHỮNG	
NỀN KINH TẾ	
NẾU KHÔNG CÓ	
NGÀY # THÁNG	
NGHIÊN CỨU KHOA	
NGƯỜI ĐÀN ÔNG	
NGƯỜI PHỤ NỮ	
NGƯỜI TA CÓ	
NGƯỜI TA KHÔNG	
NGƯỜI VIỆT NAM	
NGUYỄN ĐỨC ĐÁT	
NHÀ KHOA HỌC	

NHƯ THẾ NÀO	
NHƯ THẾ NÀY	
NÓI CHUYỆN VỚI	
Ở VIỆT NAM	
QUAN TÂM ĐẾN	
TA CÓ THỂ	
THÂN ÁI GỬI	
THÁNG # NĂM	
TIẾNG ĐỒNG HỒ	
TRÊN THẾ GIỚI	
TRỞ THÀNH MỘT	
TRONG XÃ HỘI	
TRƯỜNG ĐẠI HỌC	
TỪ NĂM #	
VÀ GIA ĐÌNH	
VÀ HẠNH PHÚC	
VÀO NĂM #	

List 1d: 4-Tiếng Forms

4-Tiếng Content Forms	4-Tiếng Function Forms
# # # #	LÚC NÀO CŨNG CÓ
# THÁNG # NĂM	TẤT CẢ NHỮNG GÌ
# TIẾNG ĐỒNG HỒ	
Á HI HI HI	
ẤN PHẨM KHOA HỌC	
ANH {Last} {Middle} {First}	
ANH {Last} {Middle} {First}	
BÀI BÁO KHOA HỌC	
BÀI VIẾT CỦA ANH	
BIỂU TÌNH CHỐNG TRUNG	
BUỔI TỐI VUI VẺ	
CÁC BẠN DÂN CHỦ	
CÁC CUỘC BIỂU TÌNH	
CÁC NHÀ KHOA HỌC	
CÁCH ĐÂY # NĂM	
CẢM ƠN ANH ĐÃ	
CẢM ƠN ANH NHIỀU	
CẢM ƠN EM ĐÃ	

CHẾ ĐỘ CỘNG SẢN	
CHẾ ĐỘ ĐỘC TÀI	
CHIA SẺ THÂN ÁI	
CHỨ KHÔNG PHẢI LÀ	
CHỦ NGHĨA CỘNG SẢN	
CHỦ NGHĨA TƯ BẢN	
CHỦ NGHĨA XÃ HỘI	
CHÚC EM LUÔN VUI	
CHÚC MỪNG NĂM MỚI	
CHÚC MỪNG SINH NHẬT	
CHÚNG TA CÓ THỂ	
CÓ LIÊN QUAN ĐẾN	
CÓ THỂ NÓI LÀ	
CÔNG BỐ QUỐC TẾ	
CỘNG ĐỒNG NGƯỜI VIỆT	
CỘNG SẢN VIỆT NAM	
CÔNG TRÌNH NGHIÊN CỨU	
CŨNG KHÔNG PHẢI LÀ	
ĐÃ GHÉ THĂM VÀ	
DÂN TỘC VIỆT NAM	
ĐẢNG CỘNG SẢN VIỆT	
ĐẢNG VÀ NHÀ NƯỚC	
ĐÂY LÀ LẦN ĐẦU	
ĐỀ CƯƠNG NGHIÊN CỨU	
ĐỂ HÔN EM LẦN	
ĐÓ KHÔNG PHẢI LÀ	
ĐÓ LÀ LÝ DO	
ĐỘC LẬP DÂN TỘC	
ĐỘC LẬP TỰ DO	
EM ĐÃ CHIA SẺ	
GỬI ANH {Last} {Middle}	
GỬI ANH {Last} {Middle}	
GỬI {Last} {Middle} {First}	
GỬI {Last} {Middle} {First}	
HOÀN TOÀN KHÔNG CÓ	
HÔN EM LẦN NỮA	
KHỎE VÀ HẠNH PHÚC	
KHÔNG AI CÓ THỂ	
KHÔNG BAO GIỜ CÓ	

KHÔNG BIẾT BAO NHIÊU	
KHÔNG CÓ NGHĨA LÀ	
KHÔNG CÓ THỜI GIAN	
KHÔNG PHẢI AI CŨNG	
KHÔNG PHẢI LÀ MỘT	
KHÔNG PHẢI LÀ NGƯỜI	
KINH TẾ THỊ TRƯỜNG	
LÀ LẦN ĐẦU TIÊN	
LÀ MỘT TRONG NHỮNG	
LÀM THẾ NÀO ĐỂ	
MẸ CON NHÀ CÚN	
MỘT NGƯỜI ĐÀN ÔNG	
NGÀY # THÁNG #	
NGHIÊN CỨU KHOA HỌC	
NGƯỜI TA CÓ THỂ	
NHÂN DÂN VIỆT NAM	
NƯỚC ÚC PHÁT THÈM	
ƠN EM ĐÃ CHIA	
PHÁT TRIỂN KINH TẾ	
PHÊ BÌNH VĂN HỌC	
QUAN TRỌNG NHẤT LÀ	
SẼ KHÔNG BAO GIỜ	
SỰ PHÁT TRIỂN CỦA	
TAM QUYỀN PHÂN LẬP	
TẬP SAN KHOA HỌC	
TẤT CẢ MỌI NGƯỜI	
THÂN ÁI GỬI ANH	
THÂN ÁI GỬI {Last}	
THÁNG # NĂM #	
THẾ LỰC THÙ ĐỊCH	
TRẢ LỜI CÂU HỎI	
TRÊN CÁC TẬP SAN	
TRONG VÀ NGOÀI NƯỚC	
TRONG VÒNG # NĂM	
TỪ # ĐẾN #	
TỰ DO DÂN CHỦ	
TỰ DO NGÔN LUẬN	
TỪ NĂM # ĐẾN	
ỦNG HỘ TINH THẦN	
VIỆT NAM HIỆN NAY	

XÃ HỘI CHỦ NGHĨA	
XÃ HỘI DÂN SỰ	

List 1e: 5-Tiếng Forms

5-Tiếng Content Forms	5-Tiếng Function Forms
# # # # #	
# BÀI BÁO KHOA HỌC	
# NĂM TRỞ LẠI ĐÂY	
# THÁNG # NĂM #	
ANH BUỔI TỐI VUI VẺ	
ANH {Name} VÀ CHỊ {Name}	
BÀ AUNG SAN SUU KYI	
BÀI BÁO KHOA HỌC TRÊN	
BÀI VIẾT CỦA ÔNG {Name}	
BAN CHẤP HÀNH TRUNG ƯƠNG	
BIỂU TÌNH CHỐNG TRUNG QUỐC	
BLOG FOR YOU GREAT HTTP	
CÁC BẠN DÂN CHỦ KHÔNG	
CÁC BẠN DÂN CHỦ THÌ	
CÁC PHƯƠNG TIỆN TRUYỀN THÔNG	
CÁC TẬP SAN KHOA HỌC	
CÁC TẬP SAN QUỐC TẾ	
CÁC THẾ LỰC THÙ ĐỊCH	
CÁI NƯỚC MÌNH NÓ THẾ	
CÁM ƠN DÌ NHIỀU NHIỀU	
CẢM ƠN EM CHÚC EM	
CẢM ƠN EM ĐÃ CHIA	
CẢM ƠN EM ĐÃ ĐỒNG	
CẢM ƠN EM ĐÃ GHÉ	
CHÂU Á THÁI BÌNH DƯƠNG	
CHỈ BIẾT CÒN ĐẢNG CÒN	
CHỈ CÓ TRUNG QUỐC LÀ	
CHIA SẺ THÂN ÁI GỬI	
CHỦ NGHĨA TƯ BẢN THÂN	
CHÚC ANH BUỔI TỐI VUI	
CHÚC ANH VÀ GIA ĐÌNH	
CHÚC CHỊ VÀ GIA ĐÌNH	
CHÚC EM LUÔN VUI KHỎE	

CÓ TRUNG QUỐC LÀ TỐT	
CỘNG HÒA XÃ HỘI CHỦ	
CỦA CÁC BẠN DÂN CHỦ	
CỦA ĐẢNG VÀ NHÀ NƯỚC	
{Name} VÀ CHỊ {Name} {Name}	
ĐÃ CHIA SẺ THÂN ÁI	
ĐÃ GHÉ THĂM VÀ CHIA	
ĐẢNG CỘNG SẢN VIỆT NAM	
ĐÂY LÀ LẦN ĐẦU TIÊN	
ĐỂ HÔN EM LẦN NỮA	
DƯỚI SỰ LÃNH ĐẠO CỦA	
GHÉ THĂM VÀ CHIA SẺ	
GỬI ANH {Last} {Middle} {First}	
GỬI ANH {Last} {Middle} {First}	
GỬI ANH {Last} {Middle} {First}	
GỬI CHÚ {Last} {Middle} {First}	
GỬI {Name} {Name} {Name} {Name}	
HỘ TINH THẦN CHO {Name}	
HÒA XÃ HỘI CHỦ NGHĨA	
HOÀN TOÀN ĐỒNG Ý VỚI	
HOÀNG SA VÀ TRƯỜNG SA	
HỌC TRÊN CÁC TẬP SAN	
HỘI CHỦ NGHĨA VIỆT NAM	
HỘI NHÀ VĂN VIỆT NAM	
HƠN ĐỘC LẬP TỰ DO	
KHẮP NƠI TRÊN THẾ GIỚI	
KHOA HỌC TRÊN CÁC TẬP	
KHOA HỌC VÀ CÔNG NGHỆ	
KHÔNG CÓ GÌ QUÝ HƠN	
KHÔNG PHẢI LÚC NÀO CŨNG	
LÀ MỘT TRONG NHỮNG NGƯỜI	
LUÔN ỦNG HỘ TINH THẦN	
MẠNH KHỎE VÀ HẠNH PHÚC	
MẠNH KHỎE VÀ VUI VẺ	
MỘT CÔNG TRÌNH NGHIÊN CỨU	
MUỐN LÀM GÌ THÌ LÀM	
NẾU KHÔNG MUỐN NÓI LÀ	
NGÀY # THÁNG # NĂM	
NGÀY MỚI NHIỀU NIỀM VUI	

NGHĨA TƯ BẢN THÂN HỮU	
NHÀ CẦM QUYỀN VIỆT NAM	
NHÂN QUYỀN Ở VIỆT NAM	
NƯỚC CỘNG HÒA XÃ HỘI	
ƠN BẠN ĐÃ GHÉ THĂM	
ƠN EM ĐÃ CHIA SẺ	
ƠN EM ĐÃ ĐỒNG CẢM	
ƠN EM ĐÃ GHÉ THĂM	
SINH RA VÀ LỚN LÊN	
SỐ ẤN PHẨM KHOA HỌC	
SỐ BÀI BÁO KHOA HỌC	
TÀI TRỢ CHO NGHIÊN CỨU	
THÀNH PHỐ HỒ CHÍ MINH	
THANK BLOG FOR YOU GREAT	
THẦY TRÒ CÁC BẠN DÂN	
THỦ TƯỚNG NGUYỄN TẤN DŨNG	
TRÊN CÁC TẬP SAN QUỐC	
TRÒ CÁC BẠN DÂN CHỦ	
TỰ DO CÁI CON C	
ỦNG HỘ TINH THẦN CHO	
VÀ GIA ĐÌNH NĂM MỚI	
VÀ HẠNH PHÚC THÂN ÁI	
VẤN ĐỀ LIÊN QUAN ĐẾN	
VÀO NGÀY # THÁNG #	
VIẾT ĐỀ CƯƠNG NGHIÊN CỨU	
VIỆT NAM VÀ TRUNG QUỐC	
VUI VẺ VÀ HẠNH PHÚC	
XÃ HỘI CHỦ NGHĨA VIỆT	
XÂY DỰNG ĐƯỢC GÌ TRÊN	

APPENDIX J

BLOG CORPUS CONTENT AND FUNCTION FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Content Forms	1-Tiếng Function Forms
#	ANH
ĂN	BỊ
ANH	CẢ
BẠN	CÁC
BỊ	CÁCH
BIẾT	CÁI
CÁCH	CHỈ
CÁI	CHO
CHỈ	CÓ
CHÍNH	CÔ
CHO	CON
CHÚNG	CÒN
CHUYỆN	CỦA
CÓ	CŨNG
CÔ	ĐÃ
CON	ĐẾN
CÔNG	ĐƯỢC
ĐẦU	EM
ĐÂY	GÌ
ĐỂ	HAY
ĐẾN	HƠN
ĐI	KHI
ĐÓ	KHÔNG
EM	LÀ
GIỜ	LẠI
HAI	LÊN
HAY	LÚC
HỌ	MÀ
HỌC	MÌNH

KHÁC	NÀO
KHÔNG	NAY
LÀ	NÀY
LÀM	NÊN
LÊN	NHẤT
MẤY	NHƯ
MẸ	NHƯNG
MÌNH	NHỮNG
MỚI	NÓ
MỘT	Ở
NAM	ÔNG
NĂM	PHẢI
NGÀY	QUA
NGƯỜI	RA
NHÀ	RỒI
NHIỀU	SAU
NHƯ	SẼ
NÓ	SỰ
NÓI	TA
NƯỚC	THẾ
Ở	THEO
ÔNG	THÌ
QUA	TÔI
RA	TRÊN
RẤT	TRONG
SỐ	TRƯỚC
SỰ	TỪ
TA	VÀ
THÀNH	VẪN
THẤY	VÀO
THỂ	VẬY
TÔI	VÌ
VÀO	VIỆC
VỀ	VỚI
VIỆC	
VIỆT	

List 1b: 2-Tiếng Forms

2-Tiếng Content Forms	2-Tiếng Function Forms
# #	BAO GIỜ
# NĂM	BAO NHIÊU
BẮT ĐẦU	CHỨ KHÔNG
BÂY GIỜ	CÓ LẼ
CÁC BẠN	CÓ THỂ
CHỈ CÓ	CỦA CÁC
CHỈ LÀ	CỦA MỘT
CHÍNH TRỊ	CŨNG NHƯ
CHÚNG TA	ĐÂY LÀ
CHÚNG TÔI	ĐÓ LÀ
CÓ #	MỘT CÁCH
CÓ MỘT	NÀO CŨNG
CÓ NHIỀU	NHẤT LÀ
CÓ NHỮNG	NHƯ THẾ
CON NGƯỜI	NHƯ VẬY
CỦA MÌNH	SAU ĐÓ
CỦA NGƯỜI	SAU KHI
CỦA TÔI	TẠI SAO
CŨNG CÓ	TẤT CẢ
CŨNG KHÔNG	THẾ NÀO
CŨNG LÀ	TRONG NHỮNG
CŨNG NHƯ	TRƯỚC KHI
CUỘC SỐNG	VẪN CÒN
CUỐI CÙNG	VỚI NHỮNG
ĐÃ CÓ	
ĐẶC BIỆT	
ĐẠI HỌC	
ĐẦU TIÊN	
GIA ĐÌNH	
GỌI LÀ	
HÀ NỘI	
HÔM NAY	
KHOA HỌC	
KHOẢNG #	
KHÔNG BIẾT	
KHÔNG CÓ	
KHÔNG CÒN	
KHÔNG PHẢI	
KHÔNG THỂ	

KINH TẾ	
LÀ #	
LÀ MỘT	
LÀ NGƯỜI	
LÀ NHỮNG	
LÀM VIỆC	
MÀ KHÔNG	
MỌI NGƯỜI	
MỘT CÁI	
MỘT CHÚT	
MỘT NGƯỜI	
MỘT SỐ	
NĂM #	
NĂM NAY	
NGHIÊN CỨU	
NGƯỜI TA	
NGƯỜI VIỆT	
NHIỀU NGƯỜI	
NHƯ THẾ	
NHƯ VẬY	
NHỮNG NGƯỜI	
Ở ĐÂY	
Ở NHÀ	
PHẢI LÀ	
QUAN TRỌNG	
QUỐC GIA	
RẤT NHIỀU	
TÁC GIẢ	
THÁNG #	
THÀNH PHỐ	
THẾ GIỚI	
THỜI GIAN	
TÔI CÓ	
TÔI ĐÃ	
TÔI KHÔNG	
TRẢ LỜI	
TRUNG QUỐC	
VẤN ĐỀ	
VIỆT NAM	
XÃ HỘI	

List 1c: 3-Tiếng Forms

3-Tiếng Content Forms	3-Tiếng Function Forms
# # #	BAO GIỜ CŨNG
# ĐẾN #	CÀNG NGÀY CÀNG
# NĂM **#**	CHẲNG HẠN NHƯ
# THÁNG #	CHO ĐẾN KHI
AI CŨNG BIẾT	CŨNG CÓ THỂ
ANH EM {Name}	ĐẶC BIỆT LÀ
BÀI BÁO KHOA	ĐÓ LÀ NHỮNG
BẠN CÓ THỂ	LÚC NÀO CŨNG
BÁO KHOA HỌC	NÀO CŨNG CÓ
CÁC QUỐC GIA	TẤT CẢ CÁC
CÁC TẬP SAN	TẤT CẢ ĐỀU
CÂU TRẢ LỜI	TẤT CẢ MỌI
CHỈ CÓ #	TẤT CẢ NHỮNG
CHỈ CÓ MỘT	TRONG ĐÓ CÓ
CHỈ LÀ MỘT	TRONG KHI ĐÓ
CHỨ KHÔNG PHẢI	
CHỦ YẾU LÀ	
CHƯA BAO GIỜ	
CHÚNG TA CÓ	
CÓ CƠ HỘI	
CÓ KHẢ NĂNG	
CÓ LẼ LÀ	
CÓ NGHĨA LÀ	
CÓ RẤT NHIỀU	
CÓ THỂ LÀ	
CÓ THỂ LÀM	
CÓ THỂ NÓI	
CÓ THỜI GIAN	
CÓ Ý NGHĨA	
CÔNG TRÌNH NGHIÊN	
CỦA CHÚNG TA	
CỦA NHỮNG NGƯỜI	
CỦA TRUNG QUỐC	
CỦA VIỆT NAM	
CŨNG LÀ MỘT	
CỨU KHOA HỌC	
ĐẦU TIÊN CỦA	

ĐÂY LÀ MỘT	
ĐÓ LÀ MỘT	
ĐÔNG NAM Á	
ĐƯỢC XEM LÀ	
HAI VỢ CHỒNG	
HƠN # NĂM	
KHÔNG BAO GIỜ	
KHÔNG CÓ GÌ	
KHÔNG PHẢI CHỈ	
KHÔNG PHẢI LÀ	
LÀ MỘT NGƯỜI	
LÀ MỘT TRONG	
LÀ NHỮNG NGƯỜI	
LẦN ĐẦU TIÊN	
LIÊN QUAN ĐẾN	
MỘT NGƯỜI BẠN	
MỘT SỐ NGƯỜI	
MỘT THỜI GIAN	
MỘT TRONG NHỮNG	
NỀN KINH TẾ	
NGÀY # THÁNG	
NGHIÊN CỨU KHOA	
NGHIÊN CỨU SINH	
NGƯỜI ĐÀN ÔNG	
NGƯỜI PHỤ NỮ	
NGƯỜI TA CÓ	
NGƯỜI TA KHÔNG	
NGƯỜI VIỆT NAM	
NHÀ KHOA HỌC	
NHƯ THẾ NÀO	
NHƯ THẾ NÀY	
NHỮNG VẤN ĐỀ	
NÓI CHUYỆN VỚI	
Ở VIỆT NAM	
QUAN TÂM ĐẾN	
QUAN TRỌNG NHẤT	
TA CÓ THỂ	
THÁNG # NĂM	
THẾ KỶ #	
TIẾNG ĐỒNG HỒ	

TÔI CÓ THỂ	
TRÊN THẾ GIỚI	
TRÌNH NGHIÊN CỨU	
TRỞ THÀNH MỘT	
TRONG THỜI GIAN	
TRONG TRƯỜNG HỢP	
TRƯỜNG ĐẠI HỌC	
TỪ NĂM #	
VÀO NĂM #	
VIỆT NAM VÀ	

List 1d: 4-Tiếng Forms

4-Tiếng Content Forms	4-Tiếng Function Forms
# # # #	BẤT CỨ LÚC NÀO
# BÀI BÁO KHOA	LÚC NÀO CŨNG CÓ
# NĂM VỀ TRƯỚC	MỘT CÁI GÌ ĐÓ
# THÁNG # NĂM	MỘT LÚC NÀO ĐÓ
# TIẾNG ĐỒNG HỒ	TẤT CẢ NHỮNG GÌ
ẤN PHẨM KHOA HỌC	
ANH EM NHÀ {Name}	
BÀI BÁO KHOA HỌC	
CÁC ANH CHỊ EM	
CÁC CUỘC BIỂU TÌNH	
CÁC NHÀ KHOA HỌC	
CÁC NHÀ NGHIÊN CỨU	
CÁC TẬP SAN QUỐC	
CÁCH ĐÂY # NĂM	
CẦM QUYỀN VIỆT NAM	
CÂU HỎI NGHIÊN CỨU	
CHẤT LƯỢNG NGHIÊN CỨU	
CHẾ ĐỘ ĐỘC TÀI	
CHỈ SỐ TRÍCH DẪN	
CHIẾN TRANH VIỆT NAM	
CHỨ KHÔNG PHẢI LÀ	
CHỦ NGHĨA TƯ BẢN	
CHỦ NGHĨA XÃ HỘI	
CHÚNG TA CÓ THỂ	
CÓ LIÊN QUAN ĐẾN	

CÓ THỂ NÓI LÀ	
CÓ THỂ NÓI RẰNG	
CÓ THỂ XEM LÀ	
CÔNG BỐ QUỐC TẾ	
CỘNG ĐỒNG NGƯỜI VIỆT	
CÔNG TRÌNH NGHIÊN CỨU	
CÚN VÀ CHUỘT NHẮT	
ĐÃ TRỞ THÀNH MỘT	
ĐẠI HỌC QUỐC GIA	
ĐÂY LÀ LẦN ĐẦU	
ĐỀ CƯƠNG NGHIÊN CỨU	
ĐỂ HÔN EM LẦN	
ĐÓ KHÔNG PHẢI LÀ	
ĐÓ LÀ CHƯA KỂ	
DOANH NGHIỆP NHÀ NƯỚC	
ĐƯỢC XEM LÀ MỘT	
HỆ THỐNG NGÂN HÀNG	
HÔN EM LẦN NỮA	
KẾT QUẢ NGHIÊN CỨU	
KHOA HỌC VIỆT NAM	
KHOA HỌC XÃ HỘI	
KHÔNG AI CÓ THỂ	
KHÔNG BAO GIỜ CÓ	
KHÔNG BIẾT BAO NHIÊU	
KHÔNG CÓ THỜI GIAN	
KHÔNG HIỂU TẠI SAO	
KHÔNG PHẢI AI CŨNG	
KHÔNG PHẢI CHỈ CÓ	
KHÔNG PHẢI LÀ MỘT	
KỶ NIỆM # NĂM	
LÀ LẦN ĐẦU TIÊN	
LÀ MỘT TRONG NHỮNG	
LÀM THẾ NÀO ĐỂ	
LẦN ĐẦU TIÊN TÔI	
LUẬN ÁN TIẾN SĨ	
MỘT NGƯỜI ĐÀN ÔNG	
MỘT NGƯỜI PHỤ NỮ	
MỘT NHÀ KHOA HỌC	
MỘT THỜI GIAN DÀI	
NẾU KHÔNG MUỐN NÓI	

NGÀY # THÁNG #	
NGHIÊN CỨU KHOA HỌC	
NGƯỜI MỸ GỐC VIỆT	
NGƯỜI TA CÓ THỂ	
NHÀ VĂN VIỆT NAM	
NHƯNG TRONG THỰC TẾ	
PHÁT TRIỂN KINH TẾ	
PHÊ BÌNH VĂN HỌC	
QUAN TRỌNG NHẤT LÀ	
SẼ KHÔNG BAO GIỜ	
SỐ ẤN PHẨM KHOA	
SỰ PHÁT TRIỂN CỦA	
TÀI LIỆU THAM KHẢO	
TẦN SỐ TRÍCH DẪN	
TẬP SAN KHOA HỌC	
TẬP SAN QUỐC TẾ	
TẤT CẢ MỌI NGƯỜI	
THÁNG # NĂM #	
THU NHẬP BÌNH QUÂN	
TRẢ LỜI CÂU HỎI	
TRÊN CÁC TẬP SAN	
TRONG THỜI GIAN #	
TRONG VÒNG # NĂM	
TỪ # ĐẾN #	
TỪ NĂM # ĐẾN	
ỦY HỘI SÔNG MEKONG	
VĂN HỌC NGHỆ THUẬT	
VĂN HỌC VIỆT NAM	
VIỆT NAM HIỆN NAY	
XÃ HỘI CHỦ NGHĨA	
XÃ HỘI DÂN SỰ	

List 1e: 5-Tiếng Forms

5-Tiếng Content Forms	5-Tiếng Function Forms
# # # # #	CŨNG LÀ MỘT TRONG NHỮNG
# BÀI BÁO KHOA HỌC	ĐÂY LÀ MỘT TRONG NHỮNG
# NĂM TRỞ LẠI ĐÂY	KHÔNG BIẾT BAO NHIÊU LẦN
# THÁNG # NĂM #	TẤT CẢ MỌI NGƯỜI ĐỀU

ẤN PHẨM KHOA HỌC CỦA	
ANH {Name} VÀ CHỊ {Name}	
ANH {Name} VÀ {Name} {Name}	
BÀI BÁO KHOA HỌC TRÊN	
BÀI DỰ THI SỐ #	
BAN CHẤP HÀNH TRUNG ƯƠNG	
BÁO KHOA HỌC TRÊN CÁC	
BÂY GIỜ LÀ THÁNG #	
BIỂU TÌNH CHỐNG TRUNG QUỐC	
BỐ ĐƯỢC # BÀI BÁO	
BỐ TRÊN CÁC TẬP SAN	
CÁC CHUYÊN GIA BÌNH DUYỆT	
CÁC CÔNG TRÌNH NGHIÊN CỨU	
CÁC PHƯƠNG TIỆN TRUYỀN THÔNG	
CÁC TẬP SAN KHOA HỌC	
CÁC TẬP SAN QUỐC TẾ	
CÁC THẾ LỰC THÙ ĐỊCH	
CÂU HỎI ĐẶT RA LÀ	
CHẤP HÀNH TRUNG ƯƠNG ĐẢNG	
CHÂU Á THÁI BÌNH DƯƠNG	
CHO NGHIÊN CỨU KHOA HỌC	
CHỦ NGHĨA TƯ BẢN THÂN	
{Name} {Name} VÀ {Name} {Name}	
CÓ CHUYỆN GÌ XẢY RA	
CÓ CÔNG BỐ QUỐC TẾ	
CÓ THỂ CHẤP NHẬN ĐƯỢC	
CÔNG BỐ ĐƯỢC # BÀI	
CÔNG BỐ TRÊN CÁC TẬP	
CỦA NGHIÊN CỨU KHOA HỌC	
{Name} VÀ CHỊ {Name} {Name}	
ĐẢNG CỘNG SẢN VIỆT NAM	
ĐÂY LÀ LẦN ĐẦU TIÊN	
ĐỂ HÔN EM LẦN NỮA	
ĐIẾU CÀY NGUYỄN VĂN HẢI	
ĐỐI THOẠI VỀ NHÂN QUYỀN	
ĐƯỢC # BÀI BÁO KHOA	
EM LÀM ƠN IM ĐI	
GIÁO DỤC VÀ ĐÀO TẠO	
GIÁO SƯ VÀ PHÓ GIÁO	
GIỚI LÃNH ĐẠO VIỆT NAM	

GIỮA VIỆT NAM VÀ TRUNG	
HOÀNG SA VÀ TRƯỜNG SA	
HỌC TRÊN CÁC TẬP SAN	
HỘI NHÀ VĂN VIỆT NAM	
HƠN # TIẾNG ĐỒNG HỒ	
KHẮP NƠI TRÊN THẾ GIỚI	
KHOA HỌC CỦA VIỆT NAM	
KHOA HỌC TRÊN CÁC TẬP	
KHOA HỌC VÀ CÔNG NGHỆ	
KHÔNG PHẢI LÚC NÀO CŨNG	
LÀ LẦN ĐẦU TIÊN TÔI	
LÀ MỘT TRONG NHỮNG NGƯỜI	
LÀM NGHIÊN CỨU KHOA HỌC	
LẦN ĐẦU TIÊN TRONG ĐỜI	
MỖI NGÀY MỘT TẤM HÌNH	
MỌI NGƯỜI ĐỀU CÓ QUYỀN	
MỘT BỘ PHẬN KHÔNG NHỎ	
MỘT CÂU HỎI NGHIÊN CỨU	
MỘT CÔNG TRÌNH NGHIÊN CỨU	
MỘT ĐỀ CƯƠNG NGHIÊN CỨU	
NỀN CÔNG NGHIỆP VĂN HÓA	
NẾU KHÔNG MUỐN NÓI LÀ	
NEW ENGLAND JOURNAL OF MEDICINE	
NGÀY # THÁNG # NĂM	
NGÀY # THÁNG # VỪA	
NGHĨA TƯ BẢN THÂN HỮU	
NHÀ CẦM QUYỀN VIỆT NAM	
NHÂN QUYỀN Ở VIỆT NAM	
NHƯ MỘT NGÔN NGỮ THỨ	
NHỮNG CÔNG TRÌNH NGHIÊN CỨU	
NHƯNG TRONG THỰC TẾ THÌ	
NHỮNG VẤN ĐỀ LIÊN QUAN	
Ở VIỆT NAM HIỆN NAY	
SÁNG KIẾN HẠ LƯU MEKONG	
SỐ ẤN PHẨM KHOA HỌC	
SỐ BÀI BÁO KHOA HỌC	
SƯ VÀ PHÓ GIÁO SƯ	
TÀI TRỢ CHO NGHIÊN CỨU	
THÁI LAN VÀ MÃ LAI	
THÀNH PHỐ HỒ CHÍ MINH	

TIẾNG VIỆT NHƯ MỘT NGÔN	
TRÊN CÁC TẬP SAN KHOA	
TRÊN CÁC TẬP SAN QUỐC	
TRONG NGHIÊN CỨU KHOA HỌC	
TRONG THỜI GIAN # NĂM	
TRONG THỜI GIAN GẦN ĐÂY	
TỪ NĂM # ĐẾN #	
VẤN ĐỀ LIÊN QUAN ĐẾN	
VÀO NGÀY # THÁNG #	
VIẾT BÀI BÁO KHOA HỌC	
VIẾT ĐỀ CƯƠNG NGHIÊN CỨU	
VIỆT NAM VÀ TRUNG QUỐC	
VIỆT NHƯ MỘT NGÔN NGỮ	

APPENDIX K

COMMENTS CORPUS CONTENT AND FUNCTION FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Content Forms	1-Tiếng Function Forms
#	AI
ĂN	ANH
ANH	BÀI
BÀI	BỊ
BẠN	CẢ
BỊ	CÁC
BIẾT	CÁI
CÁI	CHỈ
CẢM	CHỊ
CHỈ	CHO
CHỊ	CÓ
CHÍNH	CON
CHO	CÒN
CÓ	CỦA
CON	CŨNG
CÔNG	ĐÃ
DÂN	ĐÂU
ĐỂ	ĐẾN
ĐẾN	ĐƯỢC
ĐI	EM
ĐÓ	GÌ
ĐỌC	HAY
EM	HƠN
GIỜ	KHI
HAY	KHÔNG
HỌ	LÀ
HỌC	LẠI
KHÔNG	LẮM
LÀ	MÀ

LÀM	MÌNH
LUÔN	NÀO
MẤY	NAY
MÌNH	NÀY
MỚI	NÊN
MỘT	NHƯ
NAM	NHƯNG
NĂM	NHỮNG
NGÀY	NÓ
NGƯỜI	NỮA
NHÀ	Ở
NHÂN	ÔNG
NHIỀU	PHẢI
NHƯ	QUA
NÓ	QUÁ
NÓI	RA
NƯỚC	RỒI
Ở	SAO
ƠN	SẼ
ÔNG	SỰ
QUA	TA
QUỐC	THẾ
RA	THÌ
RẤT	THÔI
SAO	TÔI
SỰ	TRÊN
TA	TRONG
THẬT	TỪ
THẤY	VÀ
THỂ	VÀO
TÔI	VẬY
VÀO	VÌ
VỀ	VỚI
VIỆT	
VN	
VUI	

List 1b: 2-Tiếng Forms

2-Tiếng Content Forms	2-Tiếng Function Forms
# NĂM	AI CŨNG
BÀI THƠ	BAO GIỜ
BÀI VIẾT	BAO NHIÊU
BÂY GIỜ	CÁI GÌ
CÁC BẠN	CHỨ KHÔNG
CÁM ƠN	CÓ LẼ
CẢM ƠN	CÓ THỂ
CHẾ ĐỘ	ĐÓ LÀ
CHỈ CÓ	ĐÚNG LÀ
CHỈ LÀ	LÀM SAO
CHIA SẺ	NÀO CŨNG
CHIẾN TRANH	NHẤT LÀ
CHÍNH QUYỀN	NHƯ THẾ
CHÍNH TRỊ	NHƯ VẬY
CHÚC ANH	TẠI SAO
CHÚC MỪNG	TẤT CẢ
CHÚNG TA	THẾ NÀO
CÓ #	AI CŨNG
CÓ GÌ	BAO GIỜ
CÓ MỘT	BAO NHIÊU
CÓ NGƯỜI	CÁI GÌ
CÓ NHIỀU	CHỨ KHÔNG
CÓ NHỮNG	CÓ LẼ
CON NGƯỜI	CÓ THỂ
CỘNG SẢN	ĐÓ LÀ
CỦA ANH	ĐÚNG LÀ
CỦA MÌNH	LÀM SAO
CỦA NGƯỜI	NÀO CŨNG
CỦA ÔNG	NHẤT LÀ
CŨNG CÓ	NHƯ THẾ
CŨNG KHÔNG	NHƯ VẬY
CŨNG LÀ	TẠI SAO
CUỘC SỐNG	TẤT CẢ
DÂN CHỦ	THẾ NÀO
DÂN TỘC	
ĐẤT NƯỚC	
EM CŨNG	
GIA ĐÌNH	
GỌI LÀ	

HA HA	
HẠNH PHÚC	
HI HI	
HÌ HÌ	
HOA KỲ	
HÔM NAY	
KHÔNG BIẾT	
KHÔNG CÓ	
KHÔNG PHẢI	
KHÔNG THỂ	
KINH TẾ	
LÀ MỘT	
LÀ NGƯỜI	
LÀ NHỮNG	
LÀM GÌ	
LÃNH ĐẠO	
MÀ KHÔNG	
MỌI NGƯỜI	
NĂM #	
NGƯỜI DÂN	
NGƯỜI TA	
NGƯỜI VIỆT	
NHÀ NƯỚC	
NHÂN DÂN	
NHƯ THẾ	
NHƯ VẬY	
NHỮNG NGƯỜI	
Ở VN	
ƠN ANH	
ƠN EM	
ÔNG QUỐC	
PHẢI LÀ	
QUỐC GIA	
RẤT NHIỀU	
THAM NHŨNG	
THÂN ÁI	
THẾ GIỚI	
THÌ KHÔNG	
THỜI GIAN	
TRẢ LỜI	
TRUNG QUỐC	
TỰ DO	

VẤN ĐỀ	
VĂN HÓA	
VIỆT NAM	
VUI VẺ	
XÃ HỘI	

List 1c: 3-Tiếng Forms

3-Tiếng Content Forms	3-Tiếng Function Forms
# MẸ CON	CÀNG NGÀY CÀNG
ANH {Last} {Middle}	CŨNG CÓ THỂ
BÀI VIẾT CỦA	LÀM GÌ CÓ
BÀI VIẾT NÀY	LÚC NÀO CŨNG
BẠN DÂN CHỦ	NÀO CŨNG CÓ
BIẾT BAO NHIÊU	TẤT CẢ CÁC
BIỂU TÌNH CHỐNG	TẤT CẢ NHỮNG
CÁC BẠN DÂN	THÌ LÀM SAO
CÁI GÌ CŨNG	TRONG ĐÓ CÓ
CÁI GỌI LÀ	
CÁM ƠN ANH	
CẢM ƠN ANH	
CẢM ƠN BÁC	
CÁM ƠN BẠN	
CẢM ƠN BẠN	
CÁM ƠN CHỊ	
CẢM ƠN CHỊ	
CÁM ƠN DÌ	
CÁM ƠN EM	
CẢM ƠN EM	
CÂU TRẢ LỜI	
CHỈ CÓ #	
CHỈ LÀ MỘT	
CHO MỌI NGƯỜI	
CHỐNG THAM NHŨNG	
CHỨ KHÔNG PHẢI	
CHÙA BÀ ĐANH	
CHƯA BAO GIỜ	
CHÚC EM LUÔN	
CHÚC MỪNG ANH	

CÓ CƠ HỘI	
CÓ KHẢ NĂNG	
CÓ NGHĨA LÀ	
CÓ NHIỀU NGƯỜI	
CÓ PHẢI LÀ	
CÓ THỜI GIAN	
CỦA CÁC BẠN	
CỦA DÂN TỘC	
CỦA NHÀ NƯỚC	
CỦA NHỮNG NGƯỜI	
CỦA ÔNG QUỐC	
CỦA TRUNG QUỐC	
CŨNG LÀ MỘT	
ĐÃ CHIA SẺ	
ĐÃ GHÉ THĂM	
DÂN TỘC VIỆT	
DÂN TỘC VN	
DÂN VIỆT NAM	
ĐẢNG CỘNG SẢN	
DỄ THƯƠNG QUÁ	
ĐỒNG Ý VỚI	
GỬI ANH {Last}	
HA HA HA	
HAPPY NEW YEAR	
HE HE HE	
HI HI HI	
HÍ HÍ HÍ	
HÌ HÌ HÌ	
HỒ CHÍ MINH	
HỘI CHỦ NGHĨA	
HƠN # NĂM	
KHÔNG BAO GIỜ	
KHÔNG CÓ GÌ	
KHÔNG PHẢI LÀ	
LÀ NHỮNG NGƯỜI	
LẦN ĐẦU TIÊN	
MÀ KHÔNG CÓ	
MỪNG SINH NHẬT	
NỀN KINH TẾ	
NGƯỜI TA KHÔNG	

NGƯỜI VIỆT NAM	
NGUYỄN ĐỨC ĐÁT	
NGUYỄN ĐỨC THIỆN	
NGUYỄN HƯNG QUỐC	
NHÂN DÂN VN	
NHIỀU NIỀM VUI	
NHÔ HOA KỲ	
NHƯ THẾ NÀO	
NHƯ THẾ NÀY	
NHƯ VẬY THÌ	
Ở VIỆT NAM	
ƠN ANH ĐÃ	
ƠN EM ĐÃ	
SOUTH CHINA SEA	
THÂN ÁI GỬI	
TRÊN DIỄN ĐÀN	
TRÊN THẾ GIỚI	
VÀ GIA ĐÌNH	
VÀ HẠNH PHÚC	
VUI VẺ VÀ	
XÃ HỘI CHỦ	
Ý KIẾN CỦA	

List 1d: 4-Tiếng Forms

4-Tiếng Content Forms	4-Tiếng Function Forms
Á HI HI HI	
ANH {Last} {Middle} {First}	
ANH {Last} {Middle} {First}	
ANH VÀ GIA ĐÌNH	
BÀI THƠ CỦA ANH	
BÀI VIẾT CỦA ANH	
BÀI VIẾT CỦA ÔNG	
BẠN DÂN CHỦ THÌ	
BUỔI TỐI VUI VẺ	
CÁC BẠN DÂN CHỦ	
CẢI CÁCH RUỘNG ĐẤT	
CẢM ƠN ANH ĐÃ	
CẢM ƠN ANH NHIỀU	

CÁM ƠN BẠN ĐÃ	
CẢM ƠN BẠN ĐÃ	
CÁM ƠN DÌ NHIỀU	
CẢM ƠN EM CHÚC	
CẢM ƠN EM ĐÃ	
CẢM ƠN EM NHIỀU	
CHA TRUYỀN CON NỐI	
CHẾ ĐỘ CỘNG SẢN	
CHẾ ĐỘ ĐỘC TÀI	
CHỊ VÀ GIA ĐÌNH	
CHIA SẺ CÙNG ANH	
CHIA SẺ THÂN ÁI	
CHỨ KHÔNG PHẢI LÀ	
CHỦ NGHĨA CỘNG SẢN	
CHÚ {Last} {Middle} {First}	
CHÚC ANH BUỔI TỐI	
CHÚC ANH LUÔN VUI	
CHÚC ANH NGÀY MỚI	
CHÚC EM LUÔN VUI	
CHÚC MỪNG NĂM MỚI	
CHÚC MỪNG SINH NHẬT	
CỘNG SẢN VIỆT NAM	
CỦA CÁC BẠN DÂN	
{Name} {Name} {Name} {Name}	
CUỐI TUẦN VUI VẺ	
ĐÃ CHIA SẺ THÂN	
ĐÃ GHÉ THĂM VÀ	
DÂN TỘC VIỆT NAM	
ĐẢNG CỘNG SẢN VIỆT	
ĐẢNG VÀ NHÀ NƯỚC	
ĐỘC LẬP DÂN TỘC	
ĐỘC LẬP TỰ DO	
EM ĐÃ CHIA SẺ	
EM ĐÃ GHÉ THĂM	
EM LUÔN VUI KHỎE	
GHÉ THĂM VÀ CHIA	
GIA ĐÌNH NĂM MỚI	
GIÁNG SINH AN LÀNH	
GỬI ANH {Last} {Middle}	

GỬI ANH {Last} {Middle}	
GỬI CHÚ {Last} {Middle}	
GỬI {Name} {Name} {Name}	
GỬI {Last} {Middle} {First}	
GỬI {Last} {Middle} {First}	
GỬI {Last} {Middle} {First}	
GỬI NGƯỜI THÁI BÌNH	
GỬI {Last} {Middle} {First}	
GỬI {Last} {Middle} {First}	
GỬI {Last} {Middle} {First}	
HẠNH PHÚC THÂN ÁI	
HỘ TINH THẦN CHO	
HUGS # MẸ CON	
KHỎE VÀ HẠNH PHÚC	
KHÔNG BAO GIỜ CÓ	
KHÔNG CÓ NGHĨA LÀ	
KHÔNG PHẢI LÀ NGƯỜI	
KINH TẾ THỊ TRƯỜNG	
LÀ MỘT TRONG NHỮNG	
LÀM TAY SAI CHO	
MẠNH KHỎE VÀ VUI	
MẸ CON NHÀ {Name}	
MỚI NHIỀU NIỀM VUI	
NÈ HI HI HI	
NGÀY # THÁNG #	
NHÂN DÂN VIỆT NAM	
NHỮNG BÀI VIẾT CỦA	
NƯỚC ÚC PHÁT THÈM	
ƠN BẠN ĐÃ GHÉ	
ƠN EM CHÚC EM	
ƠN EM ĐÃ CHIA	
ƠN EM ĐÃ GHÉ	
SẼ KHÔNG BAO GIỜ	
TAM QUYỀN PHÂN LẬP	
TẤT CẢ MỌI NGƯỜI	
THĂM VÀ CHIA SẺ	
THÂN ÁI GỬI ANH	
THÂN ÁI GỬI NGUYỄN	

THÁNG # NĂM #	
THẾ LỰC THÙ ĐỊCH	
TRÊN DIỄN ĐÀN NÀY	
TRONG VÀ NGOÀI NƯỚC	
TRƯNG CẦU DÂN Ý	
TỰ DO DÂN CHỦ	
TỰ DO NGÔN LUẬN	
ỦNG HỘ TINH THẦN	
VÀ GIA ĐÌNH MỘT	
VUI VẺ THÂN ÁI	
XÃ HỘI CHỦ NGHĨA	

List 1e: 5-Tiếng Forms

5-Tiếng Content Forms	5-Tiếng Function Forms
# THÁNG # NĂM #	
ĂN VẠ VÀ VU VẠ	
ANH BUỔI TỐI VUI VẺ	
ANH NGÀY MỚI NHIỀU NIỀM	
BÀ AUNG SAN SUU KYI	
BÀI VIẾT CỦA ÔNG QUỐC	
BIẾT CÒN ĐẢNG CÒN TIỀN	
BIẾT KIẾM NGHỀ GÌ KHÁC	
BIỂU TÌNH CHỐNG TRUNG QUỐC	
BLOG FOR YOU GREAT HTTP	
CÁC BẠN DÂN CHỦ KHÔNG	
CÁC BẠN DÂN CHỦ THÌ	
CÁC THẾ LỰC THÙ ĐỊCH	
CẢM ƠN ANH ĐÃ GHÉ	
CẢM ƠN ANH TỚI THĂM	
CÁM ƠN BẠN ĐÃ GHÉ	
CẢM ƠN BẠN ĐÃ GHÉ	
CÁM ƠN DÌ NHIỀU NHIỀU	
CẢM ƠN ĐỒNG HƯƠNG CHÚC	
CẢM ƠN EM CHÚC EM	
CẢM ƠN EM ĐÃ CHIA	
CẢM ƠN EM ĐÃ ĐỒNG	
CẢM ƠN EM ĐÃ GHÉ	
CẢM ƠN EM RẤT NHIỀU	

CẢM ƠN EM THÂN ÁI	
CHỈ BIẾT CÒN ĐẢNG CÒN	
CHỈ CÓ TQ LÀ TỐT	
CHỈ CÓ TRUNG QUỐC LÀ	
CHIA SẺ THÂN ÁI GỬI	
CHÚ {Last} {Middle} {First} CON	
CHÚC ANH BUỔI TỐI VUI	
CHÚC ANH LUÔN VUI VẺ	
CHÚC ANH NGÀY MỚI NHIỀU	
CHÚC ANH VÀ GIA ĐÌNH	
CHÚC CHỊ VÀ GIA ĐÌNH	
CHÚC EM LUÔN VUI KHỎE	
CHÚC EM MỘT NGÀY MỚI	
CHÚC EM VÀ GIA ĐÌNH	
CÓ GÌ QUÝ HƠN ĐỘC	
CÓ TRUNG QUỐC LÀ TỐT	
CỘNG HÒA XÃ HỘI CHỦ	
CỦA CÁC BẠN DÂN CHỦ	
ĐÃ CHIA SẺ THÂN ÁI	
ĐÃ ĐỒNG CẢM THÂN ÁI	
ĐÃ GHÉ THĂM VÀ CHIA	
ĐÃ GHÉ THĂM VÀ ĐỒNG	
ĐẢNG CỘNG SẢN VIỆT NAM	
DỰNG ĐƯỢC GÌ TRÊN SỰ	
ĐƯỢC GÌ TRÊN SỰ DỐI	
DƯỚI SỰ LÃNH ĐẠO CỦA	
EM LUÔN VUI KHỎE THÂN	
GẬY ÔNG ĐẬP LƯNG ÔNG	
GHÉ THĂM VÀ CHIA SẺ	
GHÉ THĂM VÀ ĐỒNG CẢM	
GÌ QUÝ HƠN ĐỘC LẬP	
GÌ TRÊN SỰ DỐI TRÁ	
GỬI ANH {Last} {Middle} {First}	
GỬI ANH {Last} {Middle} {First}	
GỬI ANH {Last} {Middle} {First}	
GỬI ANH {Last} {Middle} {First}	
GỬI CHÚ {Last} {Middle} {First}	
GỬI {Name} {Name} {Name} {Name}	

HÈN VỚI GIẶC ÁC VỚI	
HỘ TINH THẦN CHO CÚN	
HÒA XÃ HỘI CHỦ NGHĨA	
HOÀN TOÀN ĐỒNG Ý VỚI	
HOÀNG SA VÀ TRƯỜNG SA	
HỘI CHỦ NGHĨA VIỆT NAM	
HƠN ĐỘC LẬP TỰ DO	
KHỎE VÀ HẠNH PHÚC THÂN	
KHÔNG CÓ GÌ QUÝ HƠN	
LUÔN ỦNG HỘ TINH THẦN	
LUÔN VUI KHỎE THÂN ÁI	
MẠNH KHỎE VÀ HẠNH PHÚC	
MẠNH KHỎE VÀ VUI VẺ	
NGÀY MỚI NHIỀU NIỀM VUI	
NGÀY NGHỈ CUỐI TUẦN VUI	
NGHỀ GÌ KHÁC ĐỂ SỐNG	
NƯỚC CỘNG HÒA XÃ HỘI	
ƠN ANH ĐÃ GHÉ THĂM	
ƠN BẠN ĐÃ GHÉ THĂM	
ƠN EM CHÚC EM LUÔN	
ƠN EM ĐÃ CHIA SẺ	
ƠN EM ĐÃ ĐỒNG CẢM	
ƠN EM ĐÃ GHÉ THĂM	
THÂN ÁI GỬI ANH {Last}	
THANK BLOG FOR YOU GREAT	
THẦY TRÒ CÁC BẠN DÂN	
THỦ TƯỚNG NGUYỄN TẤN DŨNG	
TRÒ CÁC BẠN DÂN CHỦ	
TỰ DO CÁI CON C	
ỦNG HỘ TINH THẦN CHO	
VÀ GIA ĐÌNH MỘT NĂM	
VÀ GIA ĐÌNH NĂM MỚI	
VÀ HẠNH PHÚC THÂN ÁI	
VỚI GIẶC ÁC VỚI DÂN	
VUI VẺ THÂN ÁI GỬI	
VUI VẺ VÀ HẠNH PHÚC	
XÃ HỘI CHỦ NGHĨA VIỆT	
XÂY DỰNG ĐƯỢC GÌ TRÊN	
XIN CHIA SẺ CÙNG ANH	

APPENDIX L

3-CORPORA FULL FORM WORD LISTS

List 1: 1-Tiếng Full Form Words List

Full Corpus	Blog Corpus	Comments Corpus
#	#	#
AI	ĂN	AI
ĂN	ANH	ĂN
ANH	BẠN	ANH
BÀI	BỊ	BÀI
BẠN	BIẾT	BẠN
BỊ	CẢ	BỊ
BIẾT	CÁC	BIẾT
CẢ	CÁCH	CẢ
CÁC	CÁI	CÁC
CÁI	CHỈ	CÁI
CHỈ	CHÍNH	CHỈ
CHỊ	CHO	CHỊ
CHÍNH	CHÚNG	CHÍNH
CHO	CHUYỆN	CHO
CÓ	CÓ	CÓ
CON	CÔ	CON
CÒN	CON	CÒN
CỦA	CÒN	CỦA
CŨNG	CỦA	CŨNG
ĐÃ	CŨNG	ĐÃ
DÂN	ĐÃ	DÂN
ĐẦU	ĐẦU	ĐÂU
ĐÂY	ĐÂY	ĐỂ
ĐỂ	ĐỂ	ĐẾN

ĐẾN	ĐẾN	ĐI
ĐI	ĐI	ĐÓ
ĐÓ	ĐÓ	ĐỌC
ĐƯỢC	ĐƯỢC	ĐƯỢC
EM	EM	EM
GÌ	GÌ	GÌ
GIỜ	GIỜ	GIỜ
HAI	HAI	HAY
HAY	HAY	HỌ
HỌ	HỌ	HỌC
HỌC	HỌC	HƠN
HƠN	HƠN	KHI
KHI	KHÁC	KHÔNG
KHÔNG	KHI	LÀ
LÀ	KHÔNG	LẠI
LẠI	LÀ	LÀM
LÀM	LẠI	LẮM
LÊN	LÀM	LUÔN
MÀ	LÊN	MÀ
MẤY	LÚC	MẤY
MẸ	MÀ	MÌNH
MÌNH	MẤY	MỚI
MỚI	MẸ	MỘT
MỘT	MÌNH	NAM
NAM	MỚI	NĂM
NĂM	MỘT	NÀO
NÀO	NAM	NAY
NAY	NĂM	NÀY
NÀY	NÀO	NÊN
NÊN	NAY	NGÀY
NGÀY	NÀY	NGƯỜI
NGƯỜI	NÊN	NHÀ
NHÀ	NGÀY	NHIỀU
NHIỀU	NGƯỜI	NHƯ
NHƯ	NHÀ	NHƯNG

NHƯNG	NHẤT	NHỮNG
NHỮNG	NHIỀU	NÓ
NÓ	NHƯ	NÓI
NÓI	NHƯNG	NỮA
NỮA	NHỮNG	NƯỚC
NƯỚC	NÓ	Ở
Ở	NÓI	ÔNG
ÔNG	NƯỚC	PHẢI
PHẢI	Ở	QUA
QUA	ÔNG	QUÁ
QUÁ	PHẢI	RA
RA	QUA	RẤT
RẤT	RA	RỒI
RỒI	RẤT	SAO
SAO	RỒI	SẼ
SAU	SAU	SỰ
SẼ	SẼ	TA
SỰ	SỐ	THẬT
TA	SỰ	THẤY
THÀNH	TA	THẾ
THẤY	THÀNH	THÌ
THẾ	THẤY	THÔI
THEO	THẾ	TÔI
THÌ	THEO	TRÊN
THÔI	THÌ	TRONG
TÔI	TÔI	TỪ
TRÊN	TRÊN	VÀ
TRONG	TRONG	VÀO
TỪ	TRƯỚC	VẬY
VÀ	TỪ	VỀ
VẪN	VÀ	VÌ
VÀO	VẪN	VIỆT
VẬY	VÀO	VN
VỀ	VẬY	VỚI
VÌ	VỀ	VUI

VIỆT	VÌ	
VỚI	VIỆT	
	VỚI	

List 2: 2-Tiếng Full Form Words List

Full Corpus	Blog Corpus	Comments Corpus
# #	# #	BÀI THƠ
BAO GIỜ	BAO GIỜ	BÀI VIẾT
BAO NHIÊU	BAO NHIÊU	BAO GIỜ
BẮT ĐẦU	BẮT ĐẦU	BAO NHIÊU
BÂY GIỜ	BÂY GIỜ	BÂY GIỜ
CÁM ƠN	CHÍNH TRỊ	CÁM ƠN
CẢM ƠN	CHÚNG TA	CẢM ƠN
CHÍNH TRỊ	CHÚNG TÔI	CHẾ ĐỘ
CHÚNG TA	CÓ LẼ	CHIA SẺ
CHÚNG TÔI	CÓ THỂ	CHIẾN TRANH
CÓ LẼ	CON NGƯỜI	CHÍNH QUYỀN
CÓ THỂ	CUỘC SỐNG	CHÍNH TRỊ
CON NGƯỜI	CUỐI CÙNG	CHÚNG TA
CUỘC SỐNG	ĐẶC BIỆT	CÓ LẼ
DÂN CHỦ	ĐẠI HỌC	CÓ THỂ
DÂN TỘC	ĐẦU TIÊN	CON NGƯỜI
ĐẤT NƯỚC	GIA ĐÌNH	CỘNG SẢN
ĐẦU TIÊN	HÀ NỘI	CUỘC SỐNG
GIA ĐÌNH	HÔM NAY	DÂN CHỦ
HÀ NỘI	KHOA HỌC	DÂN TỘC
HẠNH PHÚC	KHÔNG THỂ	ĐẤT NƯỚC
HÔM NAY	KINH TẾ	GIA ĐÌNH
KHOA HỌC	NGHIÊN CỨU	HA HA
KHÔNG THỂ	NGƯỜI TA	HẠNH PHÚC
KINH TẾ	NGƯỜI VIỆT	HI HI
LÀM SAO	NHẤT LÀ	HÌ HÌ
LÃNH ĐẠO	NHƯ THẾ	HOA KỲ
NGHIÊN CỨU	NHƯ VẬY	HÔM NAY
NGƯỜI DÂN	QUAN TRỌNG	KHÔNG THỂ
NGƯỜI TA	QUỐC GIA	KINH TẾ
NGƯỜI VIỆT	SAU KHI	LÀM SAO
NHÀ NƯỚC	TÁC GIẢ	LÃNH ĐẠO

NHÂN DÂN	TẠI SAO	NGƯỜI DÂN
NHẤT LÀ	TẤT CẢ	NGƯỜI TA
NHƯ THẾ	THÁNG #	NGƯỜI VIỆT
NHƯ VẬY	THÀNH PHỐ	NHÀ NƯỚC
QUỐC GIA	THẾ GIỚI	NHÂN DÂN
SAU KHI	THẾ NÀO	NHẤT LÀ
TÁC GIẢ	THỜI GIAN	NHƯ THẾ
TẠI SAO	TRẢ LỜI	NHƯ VẬY
TẤT CẢ	TRUNG QUỐC	QUỐC GIA
THÁNG #	TRƯỚC KHI	TẠI SAO
THẾ GIỚI	VẪN CÒN	TẤT CẢ
THẾ NÀO	VẤN ĐỀ	THAM NHŨNG
THỜI GIAN	VIỆT NAM	THÂN ÁI
TRẢ LỜI	XÃ HỘI	THẾ GIỚI
TRUNG QUỐC		THẾ NÀO
TỰ DO		THỜI GIAN
VẪN CÒN		TRẢ LỜI
VẤN ĐỀ		TRUNG QUỐC
VĂN HÓA		TỰ DO
VIỆT NAM		VẤN ĐỀ
XÃ HỘI		VĂN HÓA
		VIỆT NAM
		VUI VẺ
		XÃ HỘI

List 3: 3-Tiếng Full Form Words List

Full Corpus	Blog Corpus	Comments Corpus
# # #	# # #	CÂU TRẢ LỜI
CÂU TRẢ LỜI	CÂU TRẢ LỜI	HA HA HA
HA HA HA	NỀN KINH TẾ	HE HE HE
HI HI HI	NGHIÊN CỨU SINH	HI HI HI
HÌ HÌ HÌ	NHÀ KHOA HỌC	HÍ HÍ HÍ
NHÀ KHOA HỌC	TIẾNG ĐỒNG HỒ	HÌ HÌ HÌ
TIẾNG ĐỒNG HỒ	TRƯỜNG ĐẠI HỌC	
TRƯỜNG ĐẠI HỌC		

APPENDIX M

Table 1: By-Country Blogs Sampling Data

Australia			US			VN		
Blogger	Blog Words	Sampled Words	Blogger	Blog Words	Sampled Words	Blogger	Blog Words	Sampled Words
Aus-1	19,426	19,426	US-1	29,837	29,837	VN-1	26,631	26,631
Aus2	340,461	31,490	US-2	67,483	31,194	VN-2	47,097	30419
Aus-3	48,132	30,921	US-3	2,944	2,944	VN-3	57,526	30,397
Aus-4	21,956	21,956	US-4	11,245	11,245	VN-4	53,951	30,563
Aus-5	14,648	14,648	US-5	44,988	31,019	VN-5	18,435	18,435
Aus-6	21,682	21,682	US-6	19,069	19,069	VN-6	23,300	23,300
Aus-7	288,500	30,449	US-7	45,355	30,677	VN-7	13,632	13,632
Aus-8	53,451	30,911	US-8	35,244	30,162	VN-8	40,431	30,078
Aus-9	5,012	5,012	US-9	73,465	30,897	VN-9	14,626	14,626
Aus-10	17,118	17,118	US-10	34,661	30,531	VN-10	28,203	28,203
Aus-11	669	669	US-11	102,534	30,872	VN-11	58,956	30,092
Aus-12	12,252	12,252	US-12	19,546	19,546	VN-12	3,817	3,817
Aus-13	1,825	1,825	US-13	85,661	30,606	VN-13	7,877	7,877
Aus-14	713	713	US-14	26,324	26,324	VN-14	26,877	26,877
Aus-15	1,285	1,285	US-15	9,517	9,517	VN-15	70,304	30,496
Aus-16	3,977	3,977	US-16	48,003	30,623	VN-16	77,629	30,698
Aus-17	1,306	1,306	US-17	42,149	30,878	VN-17	74,556	30,394
Aus-18	2,089	2,089	US-18	50,689	30,705	VN-18	75,474	30,131
Aus-19	1,594	1,594	US-19	32,638	30,629	VN-19	21,123	21,123
Aus-20	1,522	1,522	US-20	39,898	30,233	VN-20	69,997	30,858
Aus-21	18,119	18,119	US-21	88,495	31,068	VN-21	11,901	11,901
Aus-22	27,136	27,136	US-22	179,592	30,227	VN-22	22,096	22,096
Aus-23	31,639	31,639	US-23	79,359	30,829	VN-23	66,587	30,537
Aus-24	38,613	30,715	US-24	3,085	3,085	VN-24	111,729	30,445
Aus-25	1,748	1,748	US-25	7,083	7,083	VN-25	88,694	30,459

APPENDIX N

AUS CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
#	7,216	24
LÀ	5,003	25
CÓ	4,941	25
VÀ	4,795	25
MỘT	4,168	25
CỦA	3,787	25
KHÔNG	3,516	25
CHO	2,799	25
TÔI	2,775	22
NGƯỜI	2,670	25
MÌNH	2,562	24
TRONG	2,390	24
ĐƯỢC	2,360	25
VỚI	2,269	25
NHỮNG	2,219	25
NÀY	1,917	25
THÌ	1,857	25
ĐỂ	1,823	25
CŨNG	1,819	25
CÁC	1,806	25
Ở	1,805	25
KHI	1,767	25
MÀ	1,766	25
NHƯ	1,758	25
LÀM	1,734	24
ĐÃ	1,696	24

VÀO	1,617	25
RA	1,592	25
ĐÓ	1,573	24
BẠN	1,545	24
LẠI	1,490	24
ĐẾN	1,477	25
VỀ	1,434	25
NHƯNG	1,410	25
CON	1,409	24
ĐI	1,396	25
CÁI	1,382	25
NÓI	1,280	22
PHẢI	1,230	25
CHỈ	1,187	24
TA	1,183	24
THỂ	1,174	25
NƯỚC	1,153	22
TỪ	1,144	24
CÒN	1,140	24
SẼ	1,128	23
NHIỀU	1,100	24
NHÀ	1,096	24
HỌC	1,088	23
TRÊN	1,039	24
ANH	1,032	22
NĂM	1,025	24
SỰ	1,015	22
ÔNG	1,014	24
HAY	985	25
RẤT	981	23
THẤY	971	25
BÁNH	968	13
ĂN	967	23
NÀO	939	25
VÌ	937	25
NÓ	930	23
THẾ	928	25
SAU	921	24

RỒI	916	25
VIỆT	913	21
CHÚNG	901	24
ĐẦU	898	21
BIẾT	896	25
CẢ	885	22
LÊN	859	24
NGÀY	856	22
CÔNG	846	22
HƠN	843	25
EM	836	21
ĐÂY	807	24
GÌ	795	24
CÁCH	783	25
HAI	779	23
BỊ	760	25
THÀNH	757	24
HỌ	748	21
NÊN	748	25
CHÍNH	742	19
CÔ	733	23
QUA	730	25
MỚI	707	24
KHÁC	700	23
NAM	697	21
NHẤT	669	22
VẬY	669	23
GIỜ	651	23
CHUYỆN	632	24
SỐ	612	21
NHÂN	611	19
CÙNG	595	22
TỚ	577	4
TRƯỚC	577	25
THEO	574	24
ĐANG	573	24

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
CÓ THỂ	769	23
VIỆT NAM	564	18
LÀ MỘT	517	21
CHÚNG TA	425	18
KHÔNG CÓ	384	23
NGƯỜI TA	309	19
CÓ MỘT	307	22
NHỮNG NGƯỜI	263	18
KHOẢNG #	258	15
NĂM #	234	17
CỦA MÌNH	232	20
ĐÓ LÀ	228	23
NGHIÊN CỨU	226	11
CÁC BẠN	213	16
# #	210	14
MỌI NGƯỜI	207	18
KHÔNG PHẢI	203	19
THỜI GIAN	201	19
ĐÂY LÀ	195	19
NHƯ THẾ	194	20
TẤT CẢ	191	18
MỘT CÁCH	189	17
# NĂM	182	19
KHOA HỌC	182	7
Ở ĐÂY	182	18
SAU ĐÓ	177	19
ĐẾN KHI	176	15
# PHÚT	172	15
CÔNG THỨC	171	5
KHÔNG THỂ	170	17
CHÚNG TÔI	169	12
BÂY GIỜ	168	20
CHỈ CÓ	168	22
CŨNG CÓ	168	19
NHƯ VẬY	168	22

THẾ GIỚI	163	17
BẮT ĐẦU	159	17
ĐẦU TIÊN	157	16
MỘT NGƯỜI	157	15
VẤN ĐỀ	157	15
KHÔNG BIẾT	152	20
CHÍNH TRỊ	151	8
GIA ĐÌNH	148	17
CỦA TÔI	145	18
ĐẶC BIỆT	145	18
LÀ NHỮNG	143	16
LÀ NGƯỜI	141	18
QUAN TRỌNG	141	15
BAO GIỜ	140	17
CUỐI CÙNG	139	18
VĂN HÓA	139	8
CŨNG LÀ	138	18
CỦA NGƯỜI	136	20
NHẤT LÀ	136	21
TRƯỚC KHI	136	19
MỘT SỐ	135	19
CÓ #	134	18
CÓ NHIỀU	134	19
SAU KHI	133	18
XÃ HỘI	133	11
THẾ NÀO	132	18
HÔM NAY	129	17
NÀO CŨNG	129	18
RẤT NHIỀU	129	18
CỦA CÁC	128	15
NGƯỜI VIỆT	126	16
MÓN ĂN	125	6
TÔI KHÔNG	125	15
TUY NHIÊN	124	13
KẾT QUẢ	123	14
CỦA HỌ	122	16
LÀ #	122	18
CŨNG KHÔNG	121	15

LÃNH ĐẠO	120	6
TÔI ĐÃ	120	14
HỖN HỢP	118	6
BẠN CÓ	117	13
CÓ NHỮNG	115	18
VỚI NHỮNG	115	16
CHỈ LÀ	114	17
ĐÃ CÓ	114	21
THỊ TRƯỜNG	114	7
CON NGƯỜI	112	17
GỌI LÀ	112	16
Ở VIỆT	112	13
TRẢ LỜI	112	19
THÁNG #	111	18
CHÍNH PHỦ	110	7
MÀ KHÔNG	110	20
TIẾP TỤC	109	15
ĐỐI VỚI	108	11
NHIỀU NGƯỜI	107	17
THÀNH PHỐ	106	16
ĐẦU TƯ	105	6
MỘT CÁI	105	16
TRỞ THÀNH	105	15
CÂU CHUYỆN	104	18
CHO ĐẾN	104	19
VẪN CÒN	103	18
CÓ LẼ	102	16

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
Ở VIỆT NAM	112	13
BẠN CÓ THỂ	76	11
NHƯ THẾ NÀO	65	15
KHÔNG PHẢI LÀ	58	13
NHÀ LÃNH ĐẠO	57	4
CHO ĐẾN KHI	54	15

TA CÓ THỂ	51	14
KHÔNG BAO GIỜ	46	17
LÚC NÀO CŨNG	44	13
# # #	43	6
TẤT CẢ CÁC	43	11
TRÊN THẾ GIỚI	42	13
ĐÂY LÀ MỘT	41	11
ĐÓ LÀ MỘT	40	12
CỦA CHÚNG TA	39	13
KHOẢNG # PHÚT	39	8
MỘT TRONG NHỮNG	39	11
CÓ THỂ LÀM	38	11
LẦN ĐẦU TIÊN	38	13
CHỨ KHÔNG PHẢI	36	11
CŨNG CÓ THỂ	36	12
LÀ MỘT TRONG	36	10
NGHIÊN CỨU SINH	36	1
VỀ NHÂN QUYỀN	36	1
CÓ THỂ NÓI	35	8
NÓI CHUYỆN VỚI	35	11
# ĐẾN #	34	8
CÓ NGHĨA LÀ	34	12
CUNG TRẦM TƯỞNG	34	1
TIẾNG ĐỒNG HỒ	34	11
MỘT THỜI GIAN	32	14
NGƯỜI VIỆT NAM	32	13
CHÚNG TA CÓ	31	10
CUỘC ĐỐI THOẠI	31	2
# ĐỘ C	30	5
CHƯA BAO GIỜ	30	10
ĐẢNG LAO ĐỘNG	30	2
LÀ NHỮNG NGƯỜI	30	11
TRỞ THÀNH MỘT	30	10
CHỈ LÀ MỘT	29	11
CÓ KHẢ NĂNG	29	8
ĐƯỢC XEM LÀ	29	8
NẾU KHÔNG CÓ	29	9
TẤT CẢ NHỮNG	29	7

CÓ CƠ HỘI	28	10
CỘNG ĐỒNG NGƯỜI	28	2
ĐỒNG NGƯỜI VIỆT	28	2
NHÂN QUYỀN Ở	28	1
NHƯ THẾ NÀY	28	13
VỀ VIỆT NAM	28	11
KHÔNG CÓ GÌ	27	11
TRONG VÒNG #	27	8
CÂU TRẢ LỜI	26	11
CHÚNG TA PHẢI	26	6
CỨU KHOA HỌC	26	1
ĐẶC BIỆT LÀ	26	11
LÀ MỘT SỰ	26	9
NGHIÊN CỨU KHOA	26	1
QUAN TRỌNG NHẤT	26	9
CÔNG NGHIỆP VĂN	25	1
CŨNG LÀ MỘT	25	10
NỀN CÔNG NGHIỆP	25	1
NGHIỆP VĂN HÓA	25	1
{Last} {Middle} {First}	25	1
TRONG THỜI GIAN	25	10
VÀO HỖN HỢP	25	3
VỚI NHỮNG NGƯỜI	25	8
# QUẢ TRỨNG	24	3
ẤN PHẨM KHOA	24	1
CÓ THỂ THAY	24	6
LIÊN QUAN ĐẾN	24	7
MỘT NGƯỜI BẠN	24	12
NHÀ CẦM QUYỀN	24	2
NHÀ KHOA HỌC	24	3
PHẨM KHOA HỌC	24	1
BLOG VIỆT LUẬN	23	6
THÌA CÀ PHÊ	23	2
VÀO TỦ LẠNH	23	3
VIỆT NAM VÀ	23	4
CÁC CUỘC ĐỐI	22	2
CHỈ CÓ MỘT	22	12
CHÚNG TA KHÔNG	22	7

CÓ THỂ ĐƯỢC	22	9
CÓ THỂ LÀ	22	9
GẮN LIỀN VỚI	22	5
Ở ĐÂY LÀ	22	10
QUAN TÂM ĐẾN	22	8
TẤT CẢ MỌI	22	8
TRONG ĐÓ CÓ	22	9
VÀO NĂM #	22	6
BÀI PHÁT BIỂU	21	3
CHỈ CÓ #	21	11
CHÍNH TRỊ GIA	21	3
CHÚNG TÔI ĐÃ	21	7
CÓ THỂ DÙNG	21	7
CÔNG THỨC NÀY	21	3
CỦA NGƯỜI VIỆT	21	8
CỦA VIỆT NAM	21	6
HỌC VIỆT NAM	21	2
NHÀ ĐẦU TƯ	21	2

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
CỘNG ĐỒNG NGƯỜI VIỆT	28	2
NGHIÊN CỨU KHOA HỌC	26	1
CÔNG NGHIỆP VĂN HÓA	25	1
NỀN CÔNG NGHIỆP VĂN	25	1
ẤN PHẨM KHOA HỌC	24	1
CÁC CUỘC ĐỐI THOẠI	22	2
LÀ MỘT TRONG NHỮNG	22	9
BÀI BÁO KHOA HỌC	20	1
CHÚNG TA CÓ THỂ	20	7
NHÂN QUYỀN Ở VIỆT	19	1
QUYỀN Ở VIỆT NAM	19	1
# # # #	17	1
# THÌA CÀ PHÊ	17	2
ĐỐI THOẠI VỀ NHÂN	17	1
NGƯỜI TA CÓ THỂ	17	10

THỊ TRƯỜNG CHỨNG KHOÁN	17	1
THOẠI VỀ NHÂN QUYỀN	17	1
# TIẾNG ĐỒNG HỒ	16	6
SỐ ẤN PHẨM KHOA	16	1
CÔNG BỐ QUỐC TẾ	15	1
ĐÀO TẠO TIẾN SĨ	15	1
THÁNG # NĂM #	15	10
VĂN HÓA BÌNH DÂN	15	1
KHOA HỌC VIỆT NAM	14	1
# CỦ HÀNH TÂY	13	3
CẦM QUYỀN VIỆT NAM	13	2
CON SỐ THỐNG KÊ	13	2
CUỘC ĐỐI THOẠI VỀ	13	1
TẬP SAN QUỐC TẾ	13	1
TẤT CẢ MỌI NGƯỜI	13	5
TRÊN CÁC TẬP SAN	13	1
TỪ # ĐẾN #	13	8
VĂN HÓA ĐẠI CHÚNG	13	1
BẢN DỊCH CỦA TÚ	12	1
CÁC CHÍNH TRỊ GIA	12	3
CỰU CHIẾN BINH ÚC	12	1
ĐẾN KHI HỖN HỢP	12	2
DỊCH CỦA TÚ TRINH	12	1
HỮU NỮ NHAN NHƯ	12	1
NỮ NHAN NHƯ NGỌC	12	1
TÌNH TRẠNG NHÂN QUYỀN	12	1
TRẠNG NHÂN QUYỀN Ở	12	1
CÁI NƯỚC MÌNH NÓ	11	1
CÓ THỂ LÀM ĐƯỢC	11	7
CÓ THỂ THAY BẰNG	11	3
CỦA NỀN CÔNG NGHIỆP	11	1
MỘT NHÀ LÃNH ĐẠO	11	3
NGÀY # THÁNG #	11	4
NHÀ CẦM QUYỀN VIỆT	11	2
NƯỚC MÌNH NÓ THẾ	11	1
TRUYỀN THÔNG XÃ HỘI	11	1
# QUẢ TRỨNG GÀ	10	2
BÀI PHÁT BIỂU CỦA	10	2

BẠN CÓ THỂ THAY	10	2
CÁ HỒI HUN KHÓI	10	1
CÁC MẠNG LƯỚI TRUYỀN	10	1
CÁC NHÀ KHOA HỌC	10	3
CÁC TẬP SAN QUỐC	10	1
CÔNG TRÌNH NGHIÊN CỨU	10	1
ĐỐI VỚI NHỮNG NGƯỜI	10	5
KHÔNG AI CÓ THỂ	10	6
LÀM THẾ NÀO ĐỂ	10	5
LƯỚI TRUYỀN THÔNG XÃ	10	1
MẠNG LƯỚI TRUYỀN THÔNG	10	1
Ở NHIỆT ĐỘ PHÒNG	10	3
SẼ KHÔNG BAO GIỜ	10	7
VẤN ĐỀ NHÂN QUYỀN	10	2
VỀ MẶT CHÍNH TRỊ	10	2
# BÀI BÁO KHOA	9	1
# BỘT MÌ #	9	2
# THÁNG # NĂM	9	7
BẠN CÓ THỂ DÙNG	9	3
CHIẾN TRANH VIỆT NAM	9	2
CHO VÀO TỦ LẠNH	9	3
CHỦ NGHĨA XÃ HỘI	9	1
DÂN BIỂU CHRIS HAYES	9	1
HABIBI YA NOUR EL	9	1
HỘI ĐỒNG THÀNH PHỐ	9	2
KẾT QUẢ NGHIÊN CỨU	9	1
MỸ VÀ TRUNG QUỐC	9	1
NGƯỜI DÂN VIỆT NAM	9	2
NGƯỜI VIỆT TỴ NẠN	9	1
SIDE OF THE WORLD	9	1
THE WRONG SIDE OF	9	1
THỜI CHIẾN TRANH LẠNH	9	1
ỦY HỘI SÔNG MEKONG	9	1
VỀ NHÂN QUYỀN VỚI	9	1
WRONG SIDE OF THE	9	1
XÃ HỘI CHỦ NGHĨA	9	4
XIN BẤM VÀO LINK	9	1
# ĐƯỜNG CÁT #	8	3

# Ở ĐÂY NHÉ	8	1
# TSP MUỐI #	8	3
ĂN GIAN VÀ ĂN	8	1
ÁNH SÁNG NHÂN TẠO	8	1
BÁC SĨ VÀ Y	8	1
BỐ ĐƯỢC # BÀI	8	1
CHỦ ĐỀ THÁNG #	8	2
CƠ SỞ VẬT CHẤT	8	2
CÔNG BỐ ĐƯỢC #	8	1

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
NỀN CÔNG NGHIỆP VĂN HÓA	25	1
NHÂN QUYỀN Ở VIỆT NAM	19	1
ĐỐI THOẠI VỀ NHÂN QUYỀN	17	1
SỐ ẤN PHẨM KHOA HỌC	16	1
CUỘC ĐỐI THOẠI VỀ NHÂN	13	1
BẢN DỊCH CỦA TÚ TRINH	12	1
HỮU NỮ NHAN NHƯ NGỌC	12	1
TÌNH TRẠNG NHÂN QUYỀN Ở	12	1
CỦA NỀN CÔNG NGHIỆP VĂN	11	1
NHÀ CẦM QUYỀN VIỆT NAM	11	2
CÁC MẠNG LƯỚI TRUYỀN THÔNG	10	1
CÁC TẬP SAN QUỐC TẾ	10	1
CÁI NƯỚC MÌNH NÓ THẾ	10	1
LƯỚI TRUYỀN THÔNG XÃ HỘI	10	1
MẠNG LƯỚI TRUYỀN THÔNG XÃ	10	1
TRÊN CÁC TẬP SAN QUỐC	10	1
# BÀI BÁO KHOA HỌC	9	1
# THÁNG # NĂM #	9	7
THE WRONG SIDE OF THE	9	1
WRONG SIDE OF THE WORLD	9	1
# QUẢ TRỨNG GÀ #	8	2
ĂN GIAN VÀ ĂN CƯỚP	8	1
BÁC SĨ VÀ Y TÁ	8	1
CÁC CUỘC ĐỐI THOẠI VỀ	8	1

CÔNG BỐ ĐƯỢC # BÀI	8	1
CỘNG ĐỒNG NGƯỜI VIỆT Ở	8	1
CỦA CỘNG ĐỒNG NGƯỜI VIỆT	8	1
HỌC TRÊN CÁC TẬP SAN	8	1
KHOA HỌC TRÊN CÁC TẬP	8	1
PHẦN # Ở ĐÂY NHÉ	8	1
THOẠI VỀ NHÂN QUYỀN VỚI	8	1
TRẠNG NHÂN QUYỀN Ở VIỆT	8	1
BỐ ĐƯỢC # BÀI BÁO	7	1
BO MUỐN ĐI BỘ VỀ	7	1
ĐƯỢC # BÀI BÁO KHOA	7	1
HAPPY NEW YEAR HAPPY NEW	7	1
LỤC BÁT CUNG TRẦM TƯỞNG	7	1
NEW YEAR HAPPY NEW YEAR	7	1
NHỮNG CON SỐ THỐNG KÊ	7	1
ON THE WRONG SIDE OF	7	1
PHẨM CỦA NỀN CÔNG NGHIỆP	7	1
SẢN PHẨM CỦA NỀN CÔNG	7	1
THE COMMITTEE RECOMMENDS THAT THE	7	1
XEM PHẦN # Ở ĐÂY	7	1
BÀI BÁO KHOA HỌC TRÊN	6	1
BẤM VÀO LINK SAU ĐÂY	6	1
BẠN CHẲNG BUỒN NGHĨ ĐẾN	6	1
BAN CHẤP HÀNH TRUNG ƯƠNG	6	1
BÁO KHOA HỌC TRÊN CÁC	6	1
CHẲNG BUỒN NGHĨ ĐẾN VIỆC	6	1
CHÂU Á THÁI BÌNH DƯƠNG	6	1
CHO BÁC SĨ VÀ Y	6	1
ĐA VĂN HÓA SỰ VỤ	6	1
DỊCH TỪ BẢN TIẾNG ANH	6	1
GIỮA MỸ VÀ TRUNG QUỐC	6	1
HAPPY NEW YEAR MAY WE	6	1
HƯ VÔ HOÁ BẤT HẠNH	6	1
KHẮP NƠI TRÊN THẾ GIỚI	6	6
NEW YEAR MAY WE ALL	6	1
THÀNH PHỐ HỒ CHÍ MINH	6	4
THƠ CỦA JUAN RAMÓN JIMÉNEZ	6	1
VỀ CÁC CUỘC ĐỐI THOẠI	6	1

VỀ CHIẾN TRANH VIỆT NAM	6	2
VỀ TÌNH TRẠNG NHÂN QUYỀN	6	1
YEAR HAPPY NEW YEAR MAY	6	1
YEAR MAY WE ALL HAVE	6	1
# NƯỚC TRANH # GHẾ	5	1
BÀI PHÁT BIỂU CỦA DÂN	5	1
BÀI PHÁT BIỂU CỦA ÔNG	5	2
BẤT HẠNH VÀ THANH TẨY	5	1
CHÍNH TRỊ VÀ QUÂN SỰ	5	1
CHƯƠNG TRÌNH HẬU TIẾN SĨ	5	1
CHUYÊN ĐỀ TRUYỆN CỰC NGẮN	5	1
CỘNG ĐỒNG NVTD ÚC CHÂU	5	1
CỦA ỦY HỘI SÔNG MEKONG	5	1
ĐẢNG CẦM QUYỀN LAO ĐỘNG	5	1
ĐƯA RAP VÀO JAZZ MÓN	5	1
GIỚI LÃNH ĐẠO VIỆT NAM	5	1
HOÁ BẤT HẠNH VÀ THANH	5	1
HỘI FOOD PHOTOGRAPHY TRÊN FACEBOOK	5	3
HỐI LỘ CHO BÁC SĨ	5	1
JAZZ MÓN THỜI TRANG QUANH	5	1
LỘ CHO BÁC SĨ VÀ	5	1
LÒ ĐẾN # ĐỘ C	5	1
MÀU VÀNG ÚA VÀ MÀU	5	1
MÓN THỜI TRANG QUANH NĂM	5	1
NGÀY # THÁNG # NĂM	5	3
OF THE SOUND OF LONELINESS	5	1
ONCE UPON A TIME IN	5	1
PHÁT BIỂU CỦA DÂN BIỂU	5	1
RAP VÀO JAZZ MÓN THỜI	5	1
SÔI RỒI VẶN NHỎ LỬA	5	1
SPEED OF THE SOUND OF	5	1
TẨY XƯƠNG BẰNG CÁCH NGÂM	5	1
TỔ CHỨC ROOM TO READ	5	1
TRỨNG Ở NHIỆT ĐỘ PHÒNG	5	2
ÚA VÀ MÀU XANH LÁ	5	1
UPON A TIME IN CABRAMATTA	5	1
VÀNG ÚA VÀ MÀU XANH	5	1
VÀO JAZZ MÓN THỜI TRANG	5	1

APPENDIX O

AUS CORPUS A-CURVE CHARTS

Chart 1a: Aus Corpus 1-Tiếng Chart

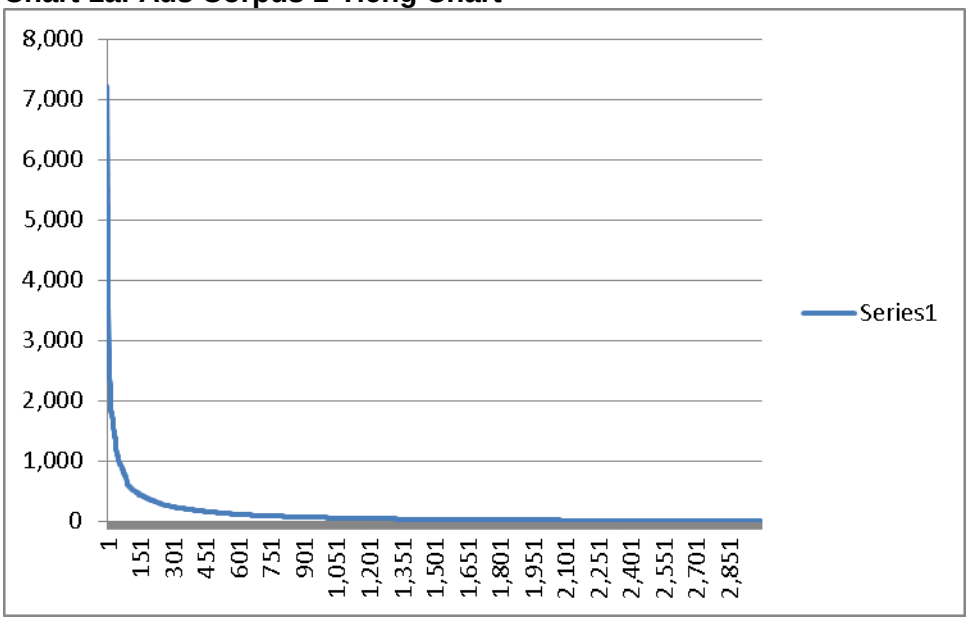


Chart 1b: Aus Corpus 2-Tiếng Chart

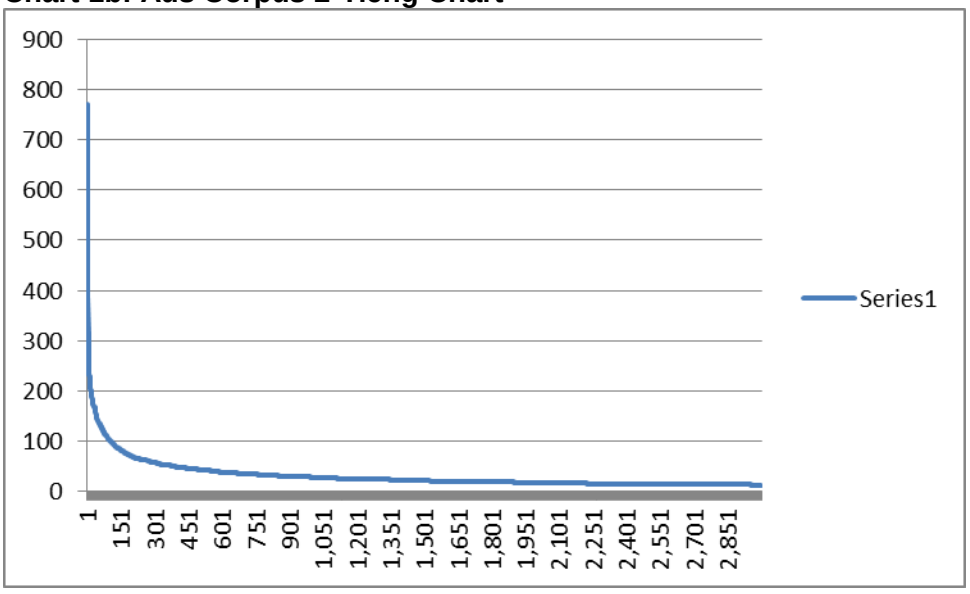


Chart 1c: Aus Corpus 3-Tiếng Chart

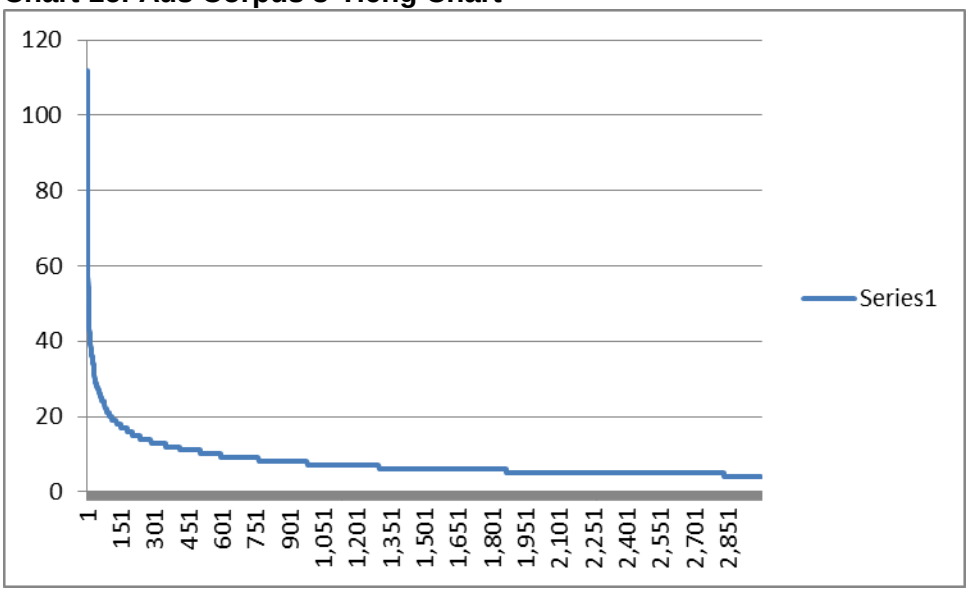


Chart 1d: Aus Corpus 4-Tiếng Chart

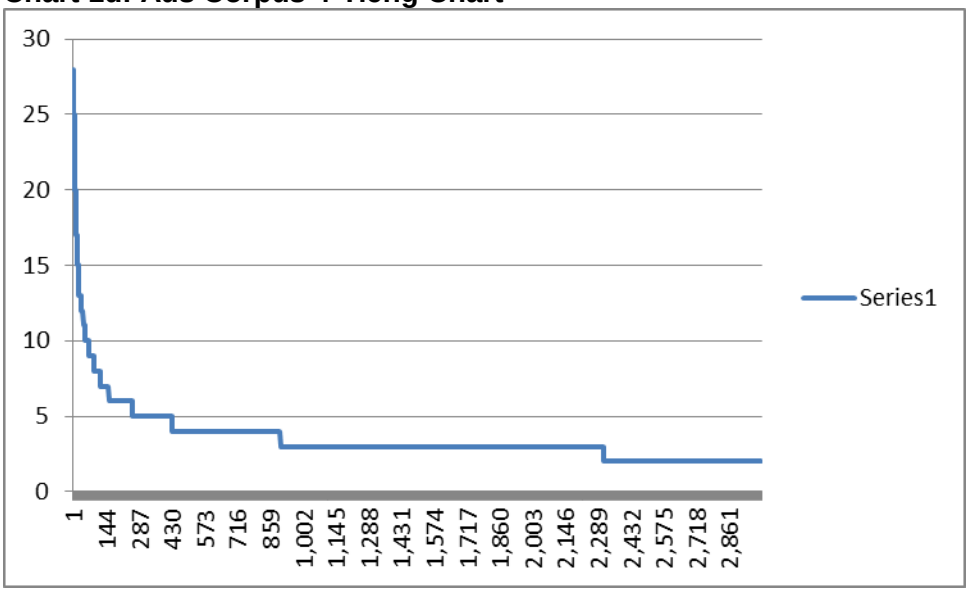
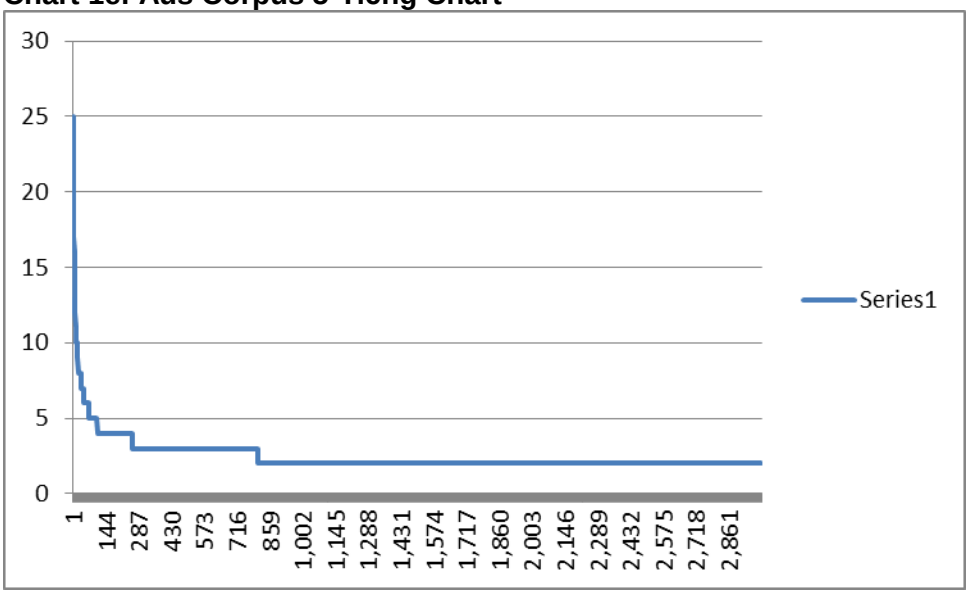


Chart 1e: Aus Corpus 5-Tiếng Chart



APPENDIX P

US CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
#	9,562	25
LÀ	8,136	25
CÓ	7,957	25
KHÔNG	7,012	25
VÀ	6,563	25
CHO	5,902	25
CỦA	5,830	25
MỘT	5,664	25
MÌNH	4,803	25
THÌ	4,389	25
ĐI	4,339	25
NGƯỜI	4,325	25
TÔI	4,191	24
NÀY	3,919	25
CÁI	3,801	25
LÀM	3,732	25
ĐƯỢC	3,725	25
MÀ	3,676	25
CŨNG	3,652	25
CON	3,622	25
TRONG	3,426	25
LẠI	3,304	25
NHỮNG	3,189	25
RA	3,169	25
VỚI	2,945	25
ĐỂ	2,937	25

ĐÃ	2,917	25
Ở	2,916	25
RỒI	2,845	25
VỀ	2,839	25
NHƯ	2,827	25
NHÀ	2,716	25
PHẢI	2,477	25
NĂM	2,361	25
NGÀY	2,347	25
CÒN	2,313	25
KHI	2,310	25
ĐÓ	2,259	25
NHƯNG	2,253	25
EM	2,214	23
MẸ	2,207	24
ĂN	2,106	24
VÀO	2,021	25
CHỈ	1,958	25
ANH	1,947	25
NÓI	1,943	25
NÊN	1,930	25
CÁC	1,911	24
ĐẾN	1,880	25
BIẾT	1,862	25
NÀO	1,829	25
BẠN	1,823	25
QUA	1,806	25
THẤY	1,796	25
TỪ	1,768	25
NHIỀU	1,751	25
NÓ	1,722	25
VÌ	1,704	25
GÌ	1,677	25
SẼ	1,660	25
TỚI	1,653	24
MỚI	1,651	25
ÔNG	1,639	25
NAY	1,626	25

MẤY	1,615	25
LÊN	1,581	25
ĐẦU	1,547	25
HAI	1,540	25
TRÊN	1,538	25
NƯỚC	1,527	25
GIỜ	1,500	24
VẬY	1,487	25
CẢ	1,471	25
BỊ	1,425	25
HAY	1,418	25
SAU	1,396	25
TA	1,344	24
THẾ	1,317	25
HƠN	1,315	25
RẤT	1,314	25
CÔ	1,301	25
ĐÂY	1,275	25
HÌNH	1,244	24
TRƯỚC	1,224	25
CHỊ	1,217	24
QUÁ	1,201	25
THỂ	1,201	25
HÔM	1,185	24
THÔI	1,183	23
NHỎ	1,178	25
BÀ	1,173	24
HỌC	1,169	25
BA	1,168	25
VIỆT	1,167	25
SAO	1,152	24
VẪN	1,146	25
ĐƯỜNG	1,124	24
LÚC	1,118	25
HẾT	1,113	25
THEO	1,112	25

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
CÓ THỂ	754	25
KHÔNG CÓ	705	25
VIỆT NAM	614	24
LÀ MỘT	585	25
HÔM NAY	503	24
GIA ĐÌNH	495	24
KHÔNG BIẾT	443	24
CÓ MỘT	414	24
CỦA MÌNH	414	24
NĂM NAY	403	24
BÂY GIỜ	398	23
NĂM #	395	24
THỜI GIAN	389	25
KHÔNG PHẢI	383	24
# NĂM	378	25
NHỮNG NGƯỜI	371	25
NGƯỜI TA	362	23
CHỈ CÓ	326	24
NÀO CŨNG	304	24
MỌI NGƯỜI	298	23
ĐI LÀM	276	23
ĐẦU TIÊN	273	25
CŨNG CÓ	271	24
CŨNG KHÔNG	270	24
BAO GIỜ	269	24
TẤT CẢ	268	24
Ở NHÀ	267	23
NHƯ VẬY	264	25
CÓ LẼ	261	24
ĐÓ LÀ	256	23
CHÚNG TÔI	254	15
BẮT ĐẦU	253	24
NHƯ THẾ	253	23
THÁNG #	244	22
HÔM QUA	242	21

CUỐI CÙNG	234	25
VẪN CÒN	233	24
ĐÂY LÀ	232	25
MÀ KHÔNG	232	23
BAO NHIÊU	231	22
MỘT NGƯỜI	231	24
CÓ NHIỀU	230	24
CHÚNG TA	225	20
CÓ NHỮNG	225	22
MỘT CÁI	221	20
THÀNH PHỐ	221	22
CHỤP HÌNH	220	16
CÓ #	218	21
TÔI KHÔNG	218	17
VẬY MÀ	218	20
SAU KHI	213	23
CHO CON	208	20
LÀM SAO	208	22
MỘT NGÀY	208	23
KHÔNG THỂ	207	23
TRƯỚC KHI	206	24
MÌNH KHÔNG	205	20
VỀ NHÀ	205	21
CỦA TÔI	202	20
LÀ NGƯỜI	202	24
KHÔNG CÒN	201	24
KINH TẾ	201	16
NHẤT LÀ	201	24
NÀY LÀ	199	23
BẠN BÈ	195	23
KHOẢNG #	195	22
# THÁNG	194	23
CHO MÌNH	193	22
MỘT CHÚT	193	24
KHÔNG ĐƯỢC	192	23
CỦA NGƯỜI	191	25
LÀ #	191	22
RẤT NHIỀU	191	24

THÌ MÌNH	191	17
CHUẨN BỊ	190	19
ANH EM	188	18
GIÁNG SINH	187	18
MÌNH CÓ	186	21
TRẢ LỜI	185	24
VẤN ĐỀ	185	17
LÀM CHO	183	24
NGƯỜI VIỆT	182	19
THẾ GIỚI	182	22
THẾ LÀ	182	21
ĐẶC BIỆT	180	25
PHẢI LÀ	178	24
CÁC BẠN	173	19
MẸ CON	173	17
DOANH NGHIỆP	171	3
Ở ĐÂY	171	24
# GIỜ	168	22
LÀM VIỆC	167	24
MỘT SỐ	167	20
LÁI XE	166	22
CỦA ÔNG	165	24
HOA KỲ	165	14
VẬY LÀ	165	22
CHỨ KHÔNG	164	23
NHỮNG NGÀY	164	21
CÁM ƠN	163	19

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
KHÔNG PHẢI LÀ	123	22
LẦN ĐẦU TIÊN	96	21
# THÁNG #	91	16
Ở VIỆT NAM	83	18
NHƯ THẾ NÀY	82	18
KHÔNG BAO GIỜ	78	16

LÚC NÀO CŨNG	74	20
NGÀY # THÁNG	68	16
MỘT TRONG NHỮNG	66	22
TƯỜNG VI #	66	2
TIẾNG ĐỒNG HỒ	64	14
NGƯỜI MỸ GỐC	61	5
CÁC DOANH NGHIỆP	58	2
TẤT CẢ CÁC	58	15
NHƯ THẾ NÀO	57	18
CHƯA BAO GIỜ	56	20
LÀ MỘT TRONG	55	22
CHỈ CÓ #	53	17
CỦA VIỆT NAM	52	14
ĐI LÀM VỀ	52	14
NÓI CHUYỆN VỚI	52	15
MỸ GỐC VIỆT	51	4
MÙA GIÁNG SINH	49	13
CŨNG CÓ THỂ	48	21
CHỨ KHÔNG PHẢI	47	20
CÓ THỜI GIAN	47	14
ĐÂY LÀ MỘT	46	13
ĐẶC BIỆT LÀ	45	14
CÓ RẤT NHIỀU	44	18
CỦA NHỮNG NGƯỜI	44	18
KHÔNG BIẾT CÓ	44	15
MÌNH CÓ THỂ	44	10
TẤT CẢ MỌI	44	12
NÀO CŨNG CÓ	42	19
NÓI CHUNG LÀ	42	12
# MẸ CON	41	9
MỘT THỜI GIAN	41	16
NỀN KINH TẾ	41	4
ANH CHỊ EM	40	11
ĐÔNG NAM Á	40	3
MÌNH KHÔNG CÓ	40	12
# NĂM #	38	9
CÓ CẢM GIÁC	38	12
HƠN # NĂM	38	18

TRƯỜNG ĐẠI HỌC	38	17
# TIẾNG ĐỒNG	37	13
BIẾT BAO NHIÊU	37	16
CÁC ANH CHỊ	37	11
CHO GIA ĐÌNH	37	14
THÁNG # NĂM	37	11
CHỈ CÓ MỘT	36	15
CỦA CHÚNG TA	36	11
LÀ MỘT NGƯỜI	36	17
MỌI NGƯỜI ĐỀU	36	7
TẤT CẢ NHỮNG	36	15
CÓ CƠ HỘI	35	16
KHÔNG CÓ GÌ	35	16
# GIỜ SÁNG	34	13
HAI MẸ CON	34	10
TRÊN THẾ GIỚI	34	11
CHỈ LÀ MỘT	33	14
CHO MỌI NGƯỜI	33	14
CÓ THỂ LÀM	33	16
NGÀY HÔM NAY	33	15
NGƯỜI VIỆT NAM	33	13
# NĂM NAY	32	9
CÂU TRẢ LỜI	32	17
CHỪNG # PHÚT	32	4
CŨNG LÀ MỘT	32	17
GẦN # NĂM	32	13
LÀM SAO MÀ	32	14
LÂU LẮM RỒI	32	16
CÁC NGÂN HÀNG	31	1
KHÔNG THỂ NÀO	31	15
TÔI CÓ THỂ	31	11
TRONG VÒNG #	31	13
TỪ NĂM #	31	11
VỚI GIA ĐÌNH	31	13
ANH EM {Name}	30	1
CẢ MỌI NGƯỜI	30	9
CÓ KHẢ NĂNG	30	14
CÓ THỂ LÀ	30	12

ĐẦU TIÊN CỦA	30	17
GIA ĐÌNH MÌNH	30	12
II CÁCH LÀM	30	1
NGÀY NÀO CŨNG	30	11
TRÊN ĐƯỜNG VỀ	30	12
TRONG ĐÓ CÓ	30	14
XỨ CAO BỒI	30	4
# NĂM TRƯỚC	29	12
HAI VỢ CHỒNG	29	8
MỘT NGƯỜI BẠN	29	14
THÌ CÓ THỂ	29	12
THÌ LÀM SAO	29	13
THỦ ĐÔ DC	29	1
TÔI KHÔNG BIẾT	29	11
TÔI KHÔNG CÓ	29	8
VÀI TẤM HÌNH	29	9
CỦA GIA ĐÌNH	28	13
HÔM NAY LÀ	28	14

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
NGÀY # THÁNG #	60	15
LÀ MỘT TRONG NHỮNG	41	22
NGƯỜI MỸ GỐC VIỆT	39	4
# TIẾNG ĐỒNG HỒ	36	13
THÁNG # NĂM #	31	9
TẤT CẢ MỌI NGƯỜI	30	9
DOANH NGHIỆP NHÀ NƯỚC	24	2
MỌI NGƯỜI ĐỀU CÓ	23	1
BIỂN ĐÔNG NAM Á	22	1
ANH EM CHÚNG TÔI	21	3
I NGUYÊN LIỆU AMP	21	1
NGƯỜI ĐỀU CÓ QUYỀN	21	1
# MUỖNG CÀ PHÊ	20	2
BẢO HIỂM SỨC KHỎE	20	2
LÀ LẦN ĐẦU TIÊN	19	12

BẢO HIỂM Y TẾ	18	3
{Name} VÀ {Name} {Name}	18	1
KHÔNG CÓ THỜI GIAN	18	9
KHÔNG PHẢI LÀ MỘT	18	6
MẸ CON NHÀ {Name}	18	1
TỐI CAO PHÁP VIỆN	17	1
# THÁNG # NĂM	16	8
HỆ THỐNG NGÂN HÀNG	16	1
TRẠI TỊ NẠN BATAAN	16	1
ANH EM NHÀ {Name}	15	1
CHÍNH PHỦ VIỆT NAM	15	1
ĐÂY LÀ LẦN ĐẦU	15	10
ĐÔNG NAM Á CHÂU	15	1
TÔI VÀ BẠN TÔI	15	1
TỪ # ĐẾN #	15	6
ANH {Name} VÀ {Name}	14	1
CÁC THẦY CÔ GIÁO	14	1
KHÔNG BIẾT BAO NHIÊU	14	9
Á CHÂU THÁI BÌNH	13	2
CÁC ANH CHỊ EM	13	7
CÁC NHÀ ĐẦU TƯ	13	3
DOANH NGHIỆP TƯ NHÂN	13	2
KHÔNG AI CÓ THỂ	13	7
LẦN ĐẦU TIÊN TÔI	13	6
LỰA CHỌN THAY THẾ	13	1
MỘT MÙA GIÁNG SINH	13	8
NATIONAL GALLERY OF ART	13	1
XỨ CAO BỒI NÀY	13	1
# NĂM VỀ TRƯỚC	12	7
CÁC DOANH NGHIỆP NHÀ	12	2
CÁCH ĐÂY # NĂM	12	5
CHÂU THÁI BÌNH DƯƠNG	12	2
CHÚNG TA CÓ THỂ	12	3
CÔNG ĂN VIỆC LÀM	12	4
ĐI TỚI ĐI LUI	12	9
NÓI NGƯỜI MỸ GỐC	12	1
TIẾNG NÓI NGƯỜI MỸ	12	1
VOICE OF VIETNAMESE AMERICANS	12	1

HÔM NAY LÀ NGÀY	11	8
KINH TẾ VIỆT NAM	11	2
MẮM CHANH TỎI ỚT	11	1
NĂM TRỞ LẠI ĐÂY	11	1
NGÔN QUỐC TẾ NHÂN	11	1
NGƯỜI CÔ GIÁO TRẺ	11	1
NƯỚC MẮM CHANH TỎI	11	2
QUỐC TẾ NHÂN QUYỀN	11	1
SẼ KHÔNG BAO GIỜ	11	9
TRONG VÒNG # NĂM	11	6
TUYÊN NGÔN QUỐC TẾ	11	1
# GIỜ ĐỒNG HỒ	10	3
CÁC DOANH NGHIỆP TƯ	10	2
CHIẾC LÁ CUỐI CÙNG	10	1
CHUA CHUA NGỌT NGỌT	10	2
CÓ CẢM GIÁC NHƯ	10	7
DÙNG NÓNG VỚI CƠM	10	1
LÀM GÌ THÌ LÀM	10	7
MẤY MẸ CON NHÀ	10	2
MỘT NGÀY NÀO ĐÓ	10	5
MỘT SỐ TRƯỜNG HỢP	10	3
MỸ GỐC Á CHÂU	10	1
QUỐC HỘI HOA KỲ	10	1
TẤT CẢ NHỮNG GÌ	10	7
TRÊN BIỂN ĐÔNG NAM	10	1
TRONG ĐÔI MẮT EM	10	1
TỰ DO NGÔN LUẬN	10	5
Á THÁI BÌNH DƯƠNG	9	3
AN SINH XÃ HỘI	9	2
CẢ NHÀ KÉO NHAU	9	6
CÁC HÃNG BẢO HIỂM	9	1
CÁC TRƯỜNG ĐẠI HỌC	9	6
CHÂU Á THÁI BÌNH	9	3
CHẾ ĐỘ ĐỘC TÀI	9	3
CÔ GIÁO TRẺ NĂM	9	1
CÓ NHIỀU LỰA CHỌN	9	2
EM MÃI LÀ #	9	1
GIÁO TRẺ NĂM XƯA	9	1

GỐC Á CHÂU THÁI	9	1
I VẬT LIỆU AMP	9	1
KHÔNG BIẾT CÓ PHẢI	9	5
KHÔNG CÓ GÌ ĐỂ	9	7
LÀ NGƯỜI ĐẦU TIÊN	9	9
LUẬT BẢO HIỂM SỨC	9	1
MẤY CON CẢM CÚM	9	1
MỘT NỀN KINH TẾ	9	4
MỘT NGƯỜI ĐÀN BÀ	9	5

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
MỌI NGƯỜI ĐỀU CÓ QUYỀN	19	1
# THÁNG # NĂM #	16	8
NGÀY # THÁNG # NĂM	16	8
ANH {Name} VÀ {Name} {Name}	14	1
Á CHÂU THÁI BÌNH DƯƠNG	12	2
BIỂN ĐÔNG NAM Á CHÂU	12	1
CÁC DOANH NGHIỆP NHÀ NƯỚC	12	2
NÓI NGƯỜI MỸ GỐC VIỆT	12	1
TIẾNG NÓI NGƯỜI MỸ GỐC	12	1
ĐÂY LÀ LẦN ĐẦU TIÊN	11	8
NGÔN QUỐC TẾ NHÂN QUYỀN	11	1
TUYÊN NGÔN QUỐC TẾ NHÂN	11	1
NƯỚC MẮM CHANH TỎI ỚT	10	1
TRÊN BIỂN ĐÔNG NAM Á	10	1
CHÂU Á THÁI BÌNH DƯƠNG	9	3
CÔ GIÁO TRẺ NĂM XƯA	9	1
GỐC Á CHÂU THÁI BÌNH	9	1
LUẬT BẢO HIỂM SỨC KHỎE	9	1
MẤY MẸ CON NHÀ CÚN	9	1
MÙA XUÂN TRONG ĐÔI MẮT	9	1
NGƯỜI MỸ GỐC Á CHÂU	9	1
XUÂN TRONG ĐÔI MẮT EM	9	1
CÁC DOANH NGHIỆP TƯ NHÂN	8	2
CÁC TÁC PHẨM NGHỆ THUẬT	8	2

CỘNG ĐỒNG NGƯỜI MỸ GỐC	8	2
CÔNG PHÁP QUỐC TẾ LUẬT	8	1
EM MÃI LÀ # TUỔI	8	1
MỖI NGÀY MỘT TẤM HÌNH	8	1
MỸ GỐC Á CHÂU THÁI	8	1
NGÀY MỘT TẤM HÌNH #	8	1
PHILIPPINES TRẠI TỊ NẠN BATAAN	8	1
# NĂM TRỞ LẠI ĐÂY	7	1
BÌNH TRÊN BIỂN ĐÔNG NAM	7	1
CÁC CÔNG TY BẢO HIỂM	7	2
CÁC LỰA CHỌN THAY THẾ	7	1
CẦU NGUYỆN CHO HÒA BÌNH	7	2
CHO CÁC THẦY CÔ GIÁO	7	1
CHO HÒA BÌNH TRÊN BIỂN	7	1
CỘNG ĐỒNG PHI VÀ VIỆT	7	1
CUỘC CÁCH MẠNG KHKT LẦN	7	1
ĐỪNG XOA EM ĐÊM NAY	7	1
GIỮ LẠI MỘT NÉT THU	7	1
HÒA BÌNH TRÊN BIỂN ĐÔNG	7	1
KHOẢNG # PHÚT LÁI XE	7	4
NGƯỜI CÔ GIÁO TRẺ NĂM	7	1
PHÁP QUỐC TẾ LUẬT BIỂN	7	1
PHÍ XỬ LÝ HỒ SƠ	7	1
TRONG BƯỚC THU VỀ #	7	1
# MUỖNG CÀ PHÊ DẦU	6	1
ANH CÚN VÀ CHỊ CHUỘT	6	1
BỘ GIAO THÔNG VẬN TẢI	6	2
CÁC DOANH NGHIỆP VIỆT NAM	6	1
CÁI ĐẸP CỦA HÌNH XẤU	6	1
CÁI XẤU CỦA HÌNH ĐẸP	6	1
CHIỀU THÀNH PHỐ MƯA BAY	6	2
CHỪNG # MUỖNG CÀ PHÊ	6	2
CỦA TẤT CẢ MỌI NGƯỜI	6	4
{Name} VÀ CHỊ {Name} {Name}	6	1
ĐÂY LÀ MỘT TRONG NHỮNG	6	4
ĐÂY NGHE EM VỀ ĐÂY	6	1
DỰ ÁN ĐẦU TƯ CÔNG	6	1
DU HỌC SINH VIỆT NAM	6	1

GDP BÌNH QUÂN ĐẦU NGƯỜI	6	1
HÀNH TRÌNH TÌM TỰ DO	6	1
HÔM QUA ĐI LÀM VỀ	6	4
HƠN # TIẾNG ĐỒNG HỒ	6	5
KÍNH MỜI QUÝ ĐỒNG HƯƠNG	6	1
LÀ MỘT TRONG NHỮNG NGƯỜI	6	6
LẦN ĐẦU TIÊN TRONG ĐỜI	6	5
NÊM NẾM CHO VỪA ĂN	6	1
NÊM NẾM LẠI CHO VỪA	6	1
NGHỊCH LÝ CỦA LỰA CHỌN	6	1
NGƯỜI ĐỀU CÓ QUYỀN ĐƯỢC	6	1
NGUYỆN CHO HÒA BÌNH TRÊN	6	1
NHÂN QUYỀN TẠI VIỆT NAM	6	1
TẠI BIỂN ĐÔNG NAM Á	6	1
TẤT CẢ MỌI NGƯỜI ĐỀU	6	5
THƯ CHO ÔNG GIÀ NOEL	6	2
TRONG BẢN TUYÊN NGÔN NÀY	6	1
ỦNG HỘ TINH THẦN CHO	6	1
VÀ MÙA XUÂN TRONG ĐÔI	6	1
VỀ ĐÂY NGHE EM VỀ	6	1
VỀ LẠI THỦ ĐÔ DC	6	1
# THÁNG ĐẦU NĂM #	5	1
ĂN CHUNG VỚI NƯỚC MẮM	5	1
BIỂU TÌNH CẦU NGUYỆN CHO	5	1
CÁC DỰ ÁN ĐẦU TƯ	5	1
CÁC HÃNG BẢO HIỂM TƯ	5	1
CAN'T LOOK AT THE STARS	5	1
CHO ANH EM NHÀ {Name}	5	1
CHỨC SWIPES FOR THE HOMELESS	5	1
{Name} {Name} VÀ {Name} {Name}	5	1
CỜ BẠC BẤT HỢP PHÁP	5	1
CÓ NHỮNG CHIỀU THÀNH PHỐ	5	2
CON NGƯỜI LÀ ĐỘNG VẬT	5	1
CỦA CHÍNH PHỦ VIỆT NAM	5	1
CỦA NGƯỜI MỸ GỐC VIỆT	5	1
CỦA NGƯỜI VIỆT HẢI NGOẠI	5	3
CŨNG LÀ MỘT TRONG NHỮNG	5	2
ĐÀN BÀ LÀ ĐỘNG VẬT	5	1

APPENDIX Q

US CORPUS A-CURVE CHARTS

Chart 1a: US Corpus 1-Tiếng Chart

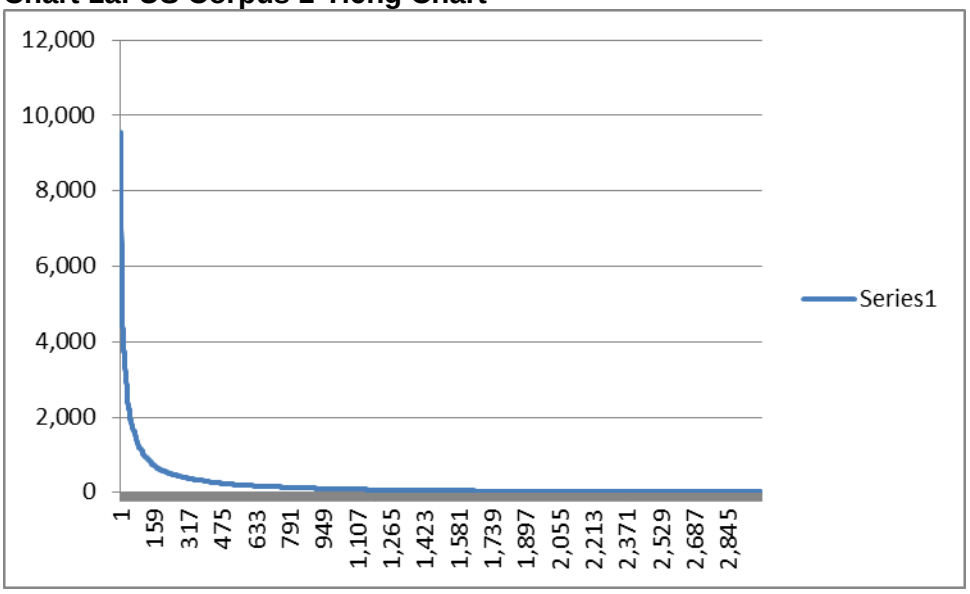


Chart 1b: US Corpus 2-Tiếng Chart

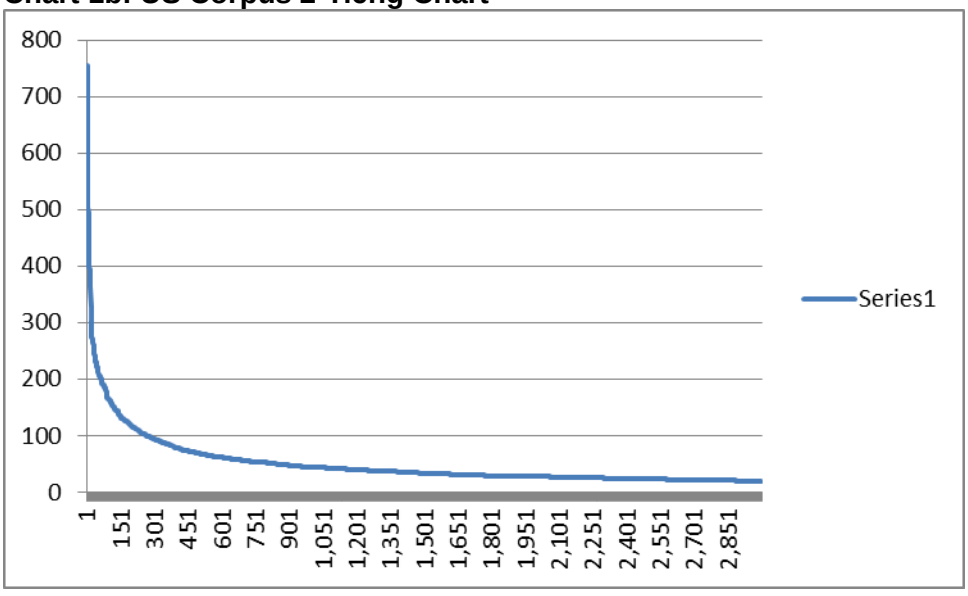


Chart 1c: US Corpus 3-Tiếng Chart

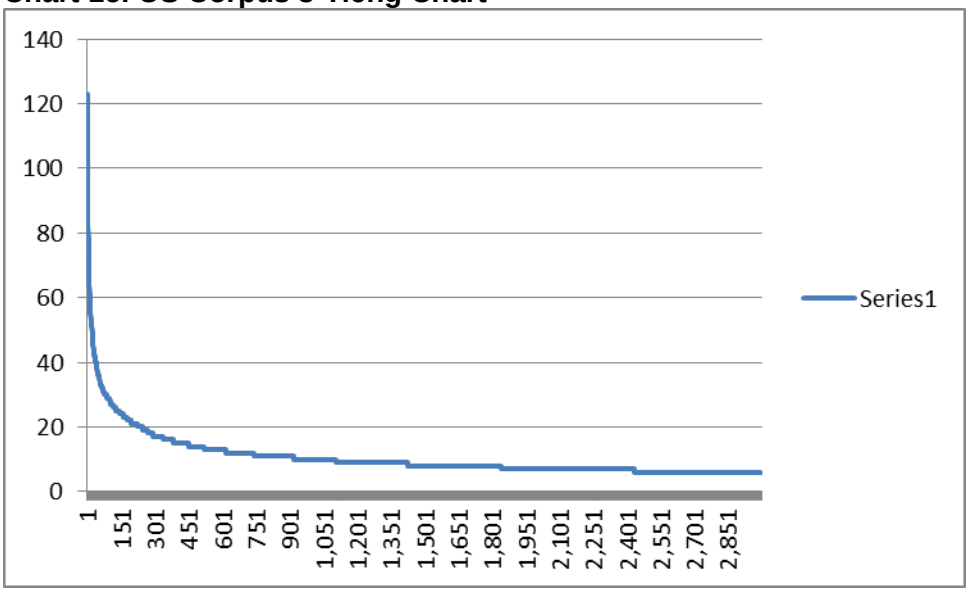


Chart 1d: US Corpus 4-Tiếng Chart

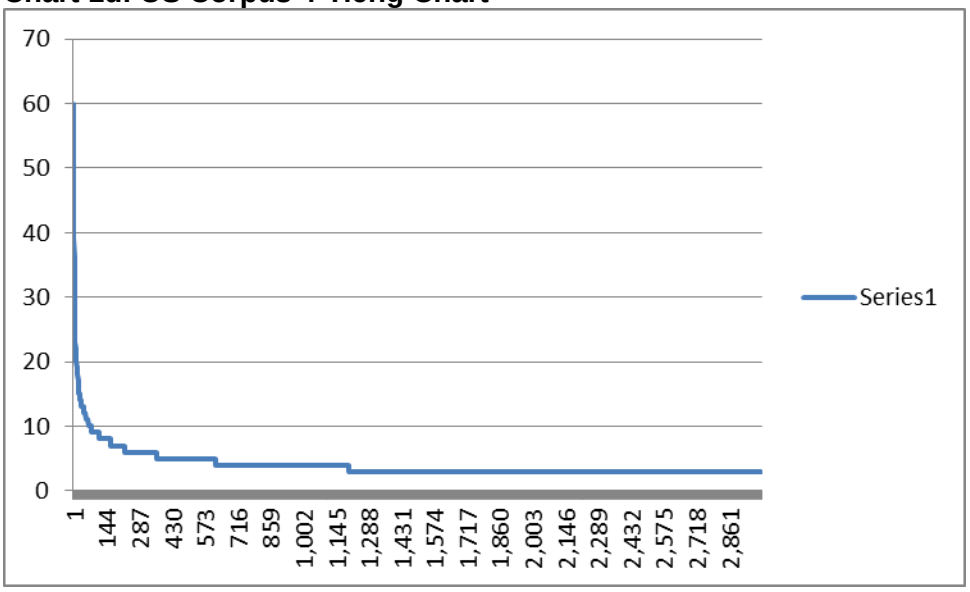
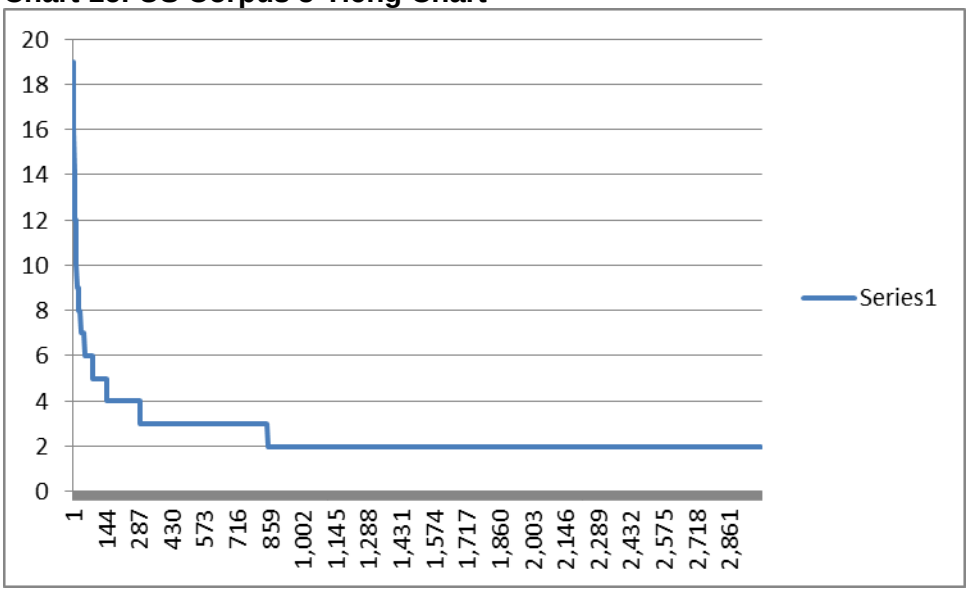


Chart 1e: US Corpus 5-Tiếng Chart



APPENDIX R

VN CORPUS FORM LISTS

List 1a: 1-Tiếng Forms

1-Tiếng Forms	Frequency	Texts
LÀ	8,248	25
CÓ	7,814	25
MỘT	7,229	25
KHÔNG	7,064	25
CỦA	5,766	25
#	5,639	25
VÀ	5,602	25
TÔI	5,348	24
NGƯỜI	5,123	25
NHỮNG	4,975	25
MÌNH	4,056	25
CHO	3,942	25
ĐƯỢC	3,802	25
ĐI	3,753	25
TRONG	3,753	25
NHƯ	3,492	25
CŨNG	3,436	25
ĐÃ	3,286	25
VỚI	3,132	25
LẠI	3,108	25
MÀ	3,080	25
CON	3,062	25
VỀ	2,936	25
RA	2,922	25
ĐỂ	2,906	25
LÀM	2,887	25
Ở	2,864	25
ANH	2,854	25
THÌ	2,789	25
ĐÓ	2,724	25

NHÀ	2,627	25
CÁI	2,574	25
NÀY	2,566	25
ĐẾN	2,505	25
EM	2,479	25
KHI	2,475	25
PHẢI	2,375	25
NHƯNG	2,338	25
RỒI	2,186	25
CÁC	2,170	25
CHỈ	2,144	25
BẠN	2,118	25
NGÀY	2,086	25
CÒN	2,077	25
VÀO	2,033	25
THẾ	1,986	25
TỪ	1,953	25
THẤY	1,934	25
NHIỀU	1,905	25
NĂM	1,896	25
NÓI	1,885	25
THỂ	1,846	25
NÀO	1,841	25
GÌ	1,814	25
NÓ	1,813	25
TA	1,722	25
BIẾT	1,714	25
CÔ	1,701	25
TRÊN	1,676	25
CẢ	1,648	25
VÌ	1,572	25
SỰ	1,554	25
MỚI	1,540	25
ĐẦU	1,517	25
LÊN	1,481	25
SẼ	1,476	25
HAI	1,475	24
HƠN	1,457	25

HỌC	1,413	25
QUA	1,399	25
ÔNG	1,395	25
MẸ	1,387	25
ĂN	1,364	25
RẤT	1,361	25
HAY	1,331	25
CÔNG	1,320	25
GIỜ	1,316	25
THÀNH	1,308	25
NÊN	1,290	25
SAU	1,285	25
VẪN	1,282	25
VIỆC	1,268	25
ĐÂY	1,222	25
NƯỚC	1,205	25
VĂN	1,205	24
ẤY	1,198	25
TÌNH	1,198	25
CẢM	1,180	25
ĐƯỜNG	1,180	25
LÚC	1,129	25
SAO	1,129	25
YÊU	1,121	25
AI	1,106	25
VẬY	1,104	25
CHUYỆN	1,099	25
NHẤT	1,093	25
TRƯỚC	1,075	25
THEO	1,056	25
CÁCH	1,047	25
SỐNG	1,027	25

List 1b: 2-Tiếng Forms

2-Tiếng Forms	Frequency	Texts
CÓ THỂ	978	25
LÀ MỘT	710	25
NGƯỜI TA	561	25
KHÔNG CÓ	542	24
KHÔNG PHẢI	465	24
CÓ MỘT	463	24
NHỮNG NGƯỜI	458	24
CỦA MÌNH	443	25
ĐÓ LÀ	442	25
HÀ NỘI	431	24
THỜI GIAN	401	25
CHÚNG TÔI	400	20
BAO GIỜ	352	25
NHƯ THẾ	351	24
MỘT NGƯỜI	347	25
BÂY GIỜ	336	25
CUỘC SỐNG	336	24
KHÔNG THỂ	336	23
CON NGƯỜI	317	25
GIA ĐÌNH	314	23
VIỆT NAM	306	20
KHÔNG BIẾT	304	25
TÌNH YÊU	302	22
HÔM NAY	300	25
TẤT CẢ	298	24
CÓ NHỮNG	291	25
BẮT ĐẦU	287	25
NHƯ VẬY	287	22
THẾ NÀO	284	25
NĂM #	281	23
# NĂM	272	24
TÔI KHÔNG	272	21
CHỈ CÓ	265	24
CŨNG KHÔNG	263	23
CŨNG CÓ	262	25

CÓ LẼ	261	24
CHỈ LÀ	259	25
ĐẦU TIÊN	259	24
TÔI ĐÃ	254	22
CẢM GIÁC	247	23
CÔNG VIỆC	247	24
NÀO CŨNG	238	25
CỦA TÔI	237	21
LÀ NHỮNG	237	24
THÁNG #	235	21
Ở ĐÂY	234	23
THÀNH PHỐ	234	23
MỌI NGƯỜI	231	23
NHÀ VĂN	227	19
ĐÂY LÀ	226	24
MỘT CÁI	223	25
LÀ NGƯỜI	222	25
CÂU CHUYỆN	218	23
CHÚNG TA	218	22
HẠNH PHÚC	218	23
VỚI NHỮNG	216	24
KHÔNG CÒN	212	25
CON ĐƯỜNG	211	23
LÀM VIỆC	211	24
CỦA NGƯỜI	210	23
NHƯ MỘT	207	23
CŨNG LÀ	204	25
TÁC GIẢ	204	19
CỦA MỘT	203	24
NHỮNG NGÀY	203	22
XÃ HỘI	203	21
CUỘC ĐỜI	202	23
MÀ KHÔNG	201	24
MỘT SỐ	201	22
SÀI GÒN	201	19
TIỂU THUYẾT	196	14
ĐÃ CÓ	195	24
THẾ GIỚI	195	23

CÁC BẠN	193	23
CUỐN SÁCH	188	17
NHẤT LÀ	188	25
RẤT NHIỀU	188	22
PHẢI LÀ	187	24
CÁI GÌ	186	23
CÓ NHIỀU	185	24
TÔI CŨNG	184	21
CON GÁI	183	21
BAO NHIÊU	180	24
MỘT CÁCH	180	22
MỘT NGÀY	179	24
Ở NHÀ	179	23
CÀ PHÊ	177	20
NHÂN VẬT	177	20
TÁC PHẨM	177	18
ĐẶC BIỆT	175	21
TÔI CÓ	173	23
SAU KHI	171	22
TRONG NHỮNG	170	24
CUỐI CÙNG	169	24
NGÀY #	168	24
MỘT CHÚT	165	25
PHỤ NỮ	165	22
NGƯỜI BẠN	164	24
ĐIỆN THOẠI	163	24
GÌ ĐÓ	162	24

List 1c: 3-Tiếng Forms

3-Tiếng Forms	Frequency	Texts
KHÔNG PHẢI LÀ	126	23
KHÔNG BAO GIỜ	115	23
LẦN ĐẦU TIÊN	79	20
NHƯ THẾ NÀO	77	18
LÚC NÀO CŨNG	73	20
# THÁNG #	69	11

TẤT CẢ NHỮNG	69	22
BẠN CÓ THỂ	63	15
NGƯỜI ĐÀN ÔNG	63	19
ĐÓ LÀ MỘT	62	16
CŨNG CÓ THỂ	60	18
MỘT TRONG NHỮNG	60	19
NGÀY # THÁNG	59	9
NGƯỜI PHỤ NỮ	59	17
CHƯA BAO GIỜ	58	18
ĐỖ HỒNG NGỌC	55	2
NHỮNG CÂU CHUYỆN	55	19
NHƯ THẾ NÀY	53	13
MỘT NGƯỜI BẠN	52	19
CHỨ KHÔNG PHẢI	50	16
CHỈ LÀ MỘT	48	19
CỦA NHỮNG NGƯỜI	48	16
LÀ MỘT TRONG	45	15
CÁI GÌ ĐÓ	44	15
TA CÓ THỂ	44	18
CUỘC SỐNG CỦA	43	13
KHÔNG CÓ GÌ	43	17
ĐẶC BIỆT LÀ	42	12
Ở VIỆT NAM	42	15
BÂY GIỜ LÀ	41	12
CŨNG LÀ MỘT	41	18
NHỮNG NGƯỜI BẠN	41	14
TÔI CÓ THỂ	41	12
CÓ THỂ NÓI	39	15
ĐIỀU GÌ ĐÓ	39	16
Ở SÀI GÒN	39	10
CÓ CẢM GIÁC	38	15
MỘT CUỐN SÁCH	38	10
NGƯỜI ĐÀN BÀ	38	14
CÓ THỂ LÀ	37	17
TRỞ THÀNH MỘT	37	14
CHO ĐẾN KHI	36	15
CÓ NGHĨA LÀ	36	18
EM LẦN NỮA	36	2

LÀ MỘT NGƯỜI	36	16
MỐI QUAN HỆ	36	17
NÀO CŨNG CÓ	36	15
NHỮNG CON ĐƯỜNG	36	13
TẤT CẢ CÁC	36	19
CÂU TRẢ LỜI	35	14
CÓ RẤT NHIỀU	35	15
ĐỂ HÔN EM	35	1
HÔN EM LẦN	35	1
LÀM THẾ NÀO	35	14
MỘT CÂU CHUYỆN	35	14
CHO NGƯỜI TA	34	14
Ở HÀ NỘI	34	13
PHÁT HIỆN RA	34	15
QUÁN CÀ PHÊ	34	13
CÓ NHỮNG NGƯỜI	33	16
CỦA CON NGƯỜI	33	16
CỦA MỘT NGƯỜI	33	14
HỘI NHÀ VĂN	33	3
# NĂM TRƯỚC	32	11
TRONG CUỘC SỐNG	32	12
CÓ THỂ LÀM	31	16
CỦA NHÀ VĂN	31	9
NÓI CHUYỆN VỚI	31	16
VỚI NHỮNG NGƯỜI	31	13
CUỐN TIỂU THUYẾT	30	12
ĐÂY LÀ MỘT	30	10
ĐÓ LÀ NHỮNG	30	13
KHÔNG HIỂU SAO	30	14
NGƯỜI TA KHÔNG	30	17
QUAN TÂM ĐẾN	30	12
CẤP THÂM NIÊN	29	1
CHỈ CÓ MỘT	29	16
CỦA TÁC GIẢ	29	6
ĐÃ TRỞ THÀNH	29	17
ĐẦU TIÊN CỦA	29	13
HAI VỢ CHỒNG	29	6
LÀ NHỮNG NGƯỜI	29	16

LIÊN QUAN ĐẾN	29	11
NHỮNG CUỐN SÁCH	29	10
PHỤ CẤP THÂM	29	1
THẾ NÀO ĐỂ	29	13
TRONG ĐÓ CÓ	29	14
BẮT ĐẦU TỪ	28	14
CÁC TÁC PHẨM	28	9
CHẲNG BAO GIỜ	28	14
CÙ LAO RÙA	28	1
ĐỂ CÓ THỂ	28	13
HAI MẸ CON	28	11
NGÀY HÔM NAY	28	13
TRẦN THU TRANG	28	1
ĐƠN GIẢN LÀ	27	13
HỒ CHÍ MINH	27	9
KHÔNG THỂ NÀO	27	11
MỘT NGƯỜI ĐÀN	27	13
NHỮNG CHI TIẾT	27	11

List 1d: 4-Tiếng Forms

4-Tiếng Forms	Frequency	Texts
NGÀY # THÁNG #	58	8
ĐỂ HÔN EM LẦN	35	1
HÔN EM LẦN NỮA	35	1
LÀ MỘT TRONG NHỮNG	31	13
PHỤ CẤP THÂM NIÊN	29	1
BÂY GIỜ LÀ THÁNG	23	1
LÀM THẾ NÀO ĐỂ	22	9
TẤT CẢ NHỮNG GÌ	22	10
MỘT NGƯỜI ĐÀN ÔNG	21	12
VĂN MIẾU TRẤN BIÊN	21	1
PHỤ NỮ VIỆT NAM	20	4
TIỂU THUYẾT CHIẾN TRANH	20	1
BỘ PHẬN KHÔNG NHỎ	19	2
GIỜ LÀ THÁNG #	18	1
TÍN NGƯỠNG THỜ MẪU	18	1

LÀ LẦN ĐẦU TIÊN	17	10
BẢO TÀNG PHỤ NỮ	16	1
KỶ NIỆM # NĂM	16	6
NHÀ VĂN VIỆT NAM	16	3
TÁC PHẨM VĂN HỌC	16	8
TẤT CẢ MỌI NGƯỜI	16	10
VĂN HỌC NGHỆ THUẬT	16	5
KHÔNG PHẢI LÀ MỘT	15	11
TÀNG PHỤ NỮ VIỆT	15	1
VIẾT VỀ CHIẾN TRANH	15	1
CHẾ ĐỘ PHỤ CẤP	14	1
HỘI NHÀ VĂN VIỆT	14	3
CUỘC SỐNG CỦA MÌNH	13	6
LÀM ƠN IM ĐI	13	6
SẼ KHÔNG BAO GIỜ	13	8
THÁNG # NĂM #	13	6
CÁCH ĐÂY # NĂM	12	7
CÓ CÁI GÌ ĐÓ	12	8
CÔNG TÁC XÃ HỘI	12	3
CỦA {Last} {Middle} {First}	12	1
ĐÂY KHÔNG PHẢI LÀ	12	11
EM LÀM ƠN IM	12	6
KHÔNG NHẤT THIẾT PHẢI	12	7
MỘT ĐIỀU GÌ ĐÓ	12	6
NHỮNG NGƯỜI PHỤ NỮ	12	7
NHỮNG NGƯỜI XUNG QUANH	12	6
THÀNH PHỐ BIÊN HÒA	12	1
THỊ TRẤN TORTILLA FLAT	12	1
VẤN ĐỀ XÃ HỘI	12	3
VĂN NGHỆ QUÂN ĐỘI	12	1
# THÁNG # NĂM	11	5
BẤT CỨ LÚC NÀO	11	8
COCKTAIL CHO TÌNH YÊU	11	1
CUỘC SỐNG GIA ĐÌNH	11	6
ĐẦU TIÊN TRONG ĐỜI	11	7
KHÔNG PHẢI LÚC NÀO	11	8
MỘT BỘ PHẬN KHÔNG	11	2
MỘT CÁI GÌ ĐÓ	11	6

MỘT NGƯỜI PHỤ NỮ	11	8
NGƯỜI TA CÓ THỂ	11	9
NHỮNG CÁNH ĐỒNG HOA	11	1
NHỮNG NGƯỜI ĐÀN ÔNG	11	7
TẠP CHÍ VĂN NGHỆ	11	2
# THỜI GIAN HOÀN	10	1
# TIẾNG ĐỒNG HỒ	10	6
ANH VÀ EM #	10	1
CÁC EM HỌC SINH	10	5
CẢNH SÁT GIAO THÔNG	10	5
ĐÂY LÀ LẦN ĐẦU	10	7
ĐỘ PHỤ CẤP THÂM	10	1
GIẢI THƯỞNG HỘI NHÀ	10	1
KHAI THẬT ĐÂU ANH	10	1
KHÔNG BAO GIỜ CÓ	10	8
LẦN ĐẦU TIÊN TRONG	10	6
LÚC NÀO CŨNG CÓ	10	7
MẬT MÃ TÂY TẠNG	10	1
NXB HỘI NHÀ VĂN	10	2
ÔNG GIÀ BA BỊ	10	1
TÁC GIẢ SỬ DỤNG	10	1
TẤT CẢ NHỮNG ĐIỀU	10	8
THỜI GIAN HOÀN THÀNH	10	1
THỜI GIAN KHỞI CÔNG	10	1
THƯỞNG HỘI NHÀ VĂN	10	1
BS {Last} {Middle} {First}	9	1
CẦN PHẢI KHAI THẬT	9	1
CHÚNG TA CÓ THỂ	9	6
CHÚNG TÔI QUYẾT ĐỊNH	9	3
CÓ THỂ XEM LÀ	9	1
CỦA CÁC TÁC GIẢ	9	6
HƠN BAO GIỜ HẾT	9	5
KHOA HỌC KỸ THUẬT	9	4
KHÔNG BAO GIỜ ĐƯỢC	9	7
KHÔNG BAO GIỜ QUÊN	9	5
KHÔNG CẦN PHẢI KHAI	9	1
KHÔNG PHẢI LÀ NGƯỜI	9	7
MỘT CUỐN TIỂU THUYẾT	9	6

PHẢI KHAI THẬT ĐÂU	9	1
RẠCH GẦM XOÀI MÚT	9	1
TÔI CHƯA BAO GIỜ	9	6
TÔI KHÔNG BAO GIỜ	9	4
TP HỒ CHÍ MINH	9	5
TRÁI TIM BẠC NHƯỢC	9	1
TRUNG TÂM THÀNH PHỐ	9	7
ẢNH SÁCH THẬT BÌA	8	1
BẤM VÀO ĐỂ XEM	8	1

List 1e: 5-Tiếng Forms

5-Tiếng Forms	Frequency	Texts
ĐỂ HÔN EM LẦN NỮA	35	1
BÂY GIỜ LÀ THÁNG #	18	1
BẢO TÀNG PHỤ NỮ VIỆT	15	1
TÀNG PHỤ NỮ VIỆT NAM	15	1
HỘI NHÀ VĂN VIỆT NAM	14	3
EM LÀM ƠN IM ĐI	12	6
MỘT BỘ PHẬN KHÔNG NHỎ	11	2
# THỜI GIAN HOÀN THÀNH	10	1
CHẾ ĐỘ PHỤ CẤP THÂM	10	1
ĐỘ PHỤ CẤP THÂM NIÊN	10	1
GIẢI THƯỞNG HỘI NHÀ VĂN	10	1
CẦN PHẢI KHAI THẬT ĐÂU	9	1
KHÔNG CẦN PHẢI KHAI THẬT	9	1
NGÀY # THÁNG # NĂM	9	5
PHẢI KHAI THẬT ĐÂU ANH	9	1
# THÁNG # NĂM #	8	4
ẢNH SÁCH THẬT BÌA #	8	1
BẤM VÀO ĐỂ XEM HÌNH	8	1
BÌA BẤM VÀO ĐỂ XEM	8	1
BỘ THIẾT KẾ BÌA BẤM	8	1
CHÍ VĂN NGHỆ QUÂN ĐỘI	8	1
ĐÂY LÀ LẦN ĐẦU TIÊN	8	6
ĐỂ XEM HÌNH CỠ LỚN	8	1
KẾ BÌA BẤM VÀO ĐỂ	8	1

QUY TẮC LÀM CHA MẸ	8	1
TẠP CHÍ VĂN NGHỆ QUÂN	8	1
THIẾT KẾ BÌA BẤM VÀO	8	1
THƯỞNG HỘI NHÀ VĂN VIỆT	8	1
TOÀN BỘ THIẾT KẾ BÌA	8	1
VÀO ĐỂ XEM HÌNH CỠ	8	1
ẤM KỂ TEA NGHE #	7	1
BAC SI {Last} {Middle} {First}	7	1
BAO GIỜ CHO ĐẾN THÁNG	7	2
BIÊN HÒA THỜI GIAN KHỞI	7	1
CÁNH ĐỒNG HOA HƯỚNG DƯƠNG	7	1
GIÁM KHẢO KHÔNG CHÍNH THỨC	7	1
HIỆP SĨ KHÔNG HIỆN HỮU	7	1
HÒA THỜI GIAN KHỞI CÔNG	7	1
KHI CẬU GẶP CẬU TA	7	1
LẦN ĐẦU TIÊN TRONG ĐỜI	7	4
PHỐ BIÊN HÒA THỜI GIAN	7	1
THÀNH PHỐ BIÊN HÒA THỜI	7	1
AN TOÀN VỆ SINH THỰC	6	2
ANH EM NHÀ TÂY SƠN	6	1
BÀN TAY NHỎ DƯỚI MƯA	6	1
BÀY TÍN NGƯỠNG THỜ MẪU	6	1
CÔNG TÁC XÃ HỘI TRONG	6	1
ĐI VỀ PHÍA KHÔNG NHAU	6	1
GIA ĐÌNH NỮ HỘ SINH	6	1
GIAO TIẾP PHI NGÔN NGỮ	6	1
GIỜ CHO ĐẾN THÁNG MƯỜI	6	1
{Name} {Name} CÙNG BÀN LUẬN	6	1
HỘI NGHỊ VIẾT VĂN TRẺ	6	1
HỘI VĂN HỌC NGHỆ THUẬT	6	2
KHÔNG PHẢI LÚC NÀO CŨNG	6	5
LÀ MỘT NGƯỜI ĐÀN ÔNG	6	4
LÀM TAN NÁT LÒNG NHAU	6	1
LỜI LÀM TAN NÁT LÒNG	6	1
MỘT LỜI LÀM TAN NÁT	6	1
NGHỊ VIẾT VĂN TRẺ TOÀN	6	1
NHÀ THƠ HỒ NGỌC SƠN	6	1
NHÂN VIÊN CÔNG TÁC XÃ	6	1

NHỮNG CÁNH ĐỒNG HOA HƯỚNG	6	1
NHỮNG QUY TẮC LÀM CHA	6	1
PHẢI LẤY NGƯỜI NHƯ ANH	6	1
TÁC XÃ HỘI TRONG BỆNH	6	1
THƯ CHO BÉ SƠ SINH	6	1
TOÀN VỆ SINH THỰC PHẨM	6	2
TRONG DÒNG SÔNG CỦA HERACLITUS	6	1
TRƯNG BÀY TÍN NGƯỠNG THỜ	6	1
TRUYỆN NGẮN CỦA R CARVER	6	2
VÂN VÂN VÀ VÂN VÂN	6	4
VIÊN CÔNG TÁC XÃ HỘI	6	1
VIẾT VĂN TRẺ TOÀN QUỐC	6	1
VỚI TẤT CẢ MỌI NGƯỜI	6	6
AI LÊN XỨ HOA ĐÀO	5	1
BẬC THẦY TRUYỆN NGẮN TỐI	5	2
BÂY GIỜ LÀ THÁNG MƯỜI	5	1
BÌA COCKTAIL CHO TÌNH YÊU	5	1
BIÊN PHÒNG TỐNG LÊ CHÂN	5	1
CÁC HỌA SĨ VIỆT NAM	5	1
CÁC THẾ LỰC THÙ ĐỊCH	5	1
CẤP THÂM NIÊN NHÀ GIÁO	5	1
CHIA SẺ CÙNG CÁC BẠN	5	2
CHO NHỮNG NGƯỜI XUNG QUANH	5	4
CHỦ TỊCH HỒ CHÍ MINH	5	3
CÓ THỂ BẠN MUỐN ĐỌC	5	1
CÔNG TY TNHH MỘT THÀNH	5	1
ĐẠI TÁ NGUYỄN CÔNG TUẤN	5	1
DOANH NGHIỆP KINH DOANH XĂNG	5	1
ĐỒN BIÊN PHÒNG TỐNG LÊ	5	1
ĐÔNG BẮC KÝ SỰ #	5	1
DỰ HỘI NGHỊ VIẾT VĂN	5	1
HAI NGƯỜI KHÔNG NHÌN MẮT	5	1
HỘI CHỢ HÀNG THỦ CÔNG	5	2
JUST THE WAY YOU ARE	5	1
KHÔNG BIẾT BAO NHIÊU LẦN	5	5
KỶ NIỆM # NĂM NGÀY	5	4
LÀ LẦN ĐẦU TIÊN TÔI	5	4
MỘT THÀNH VIÊN IN QUÂN	5	1

APPENDIX S

VN CORPUS A-CURVE CHARTS

Chart 1a: VN Corpus 1-Tiếng Chart

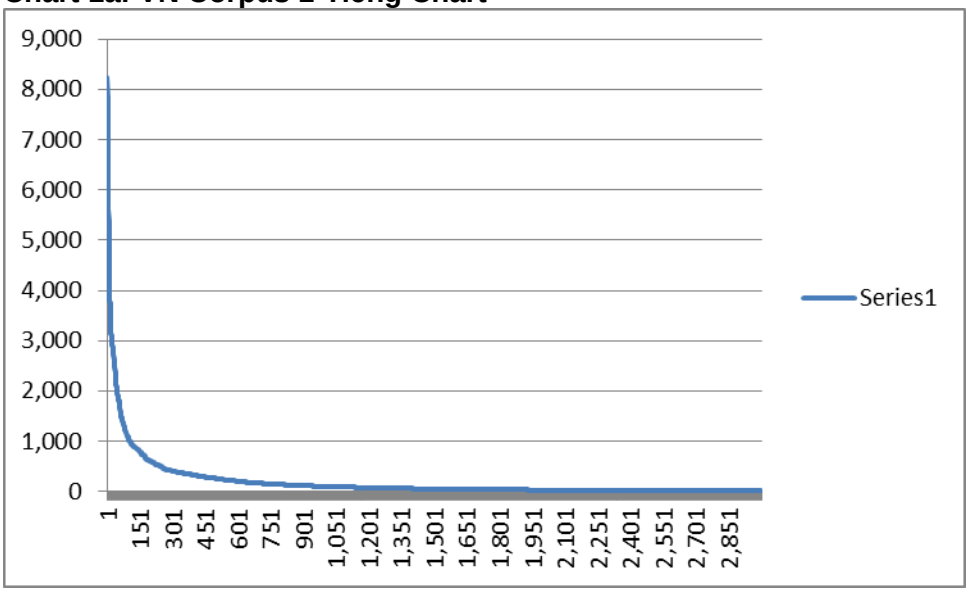


Chart 1b: VN Corpus 2-Tiếng Chart

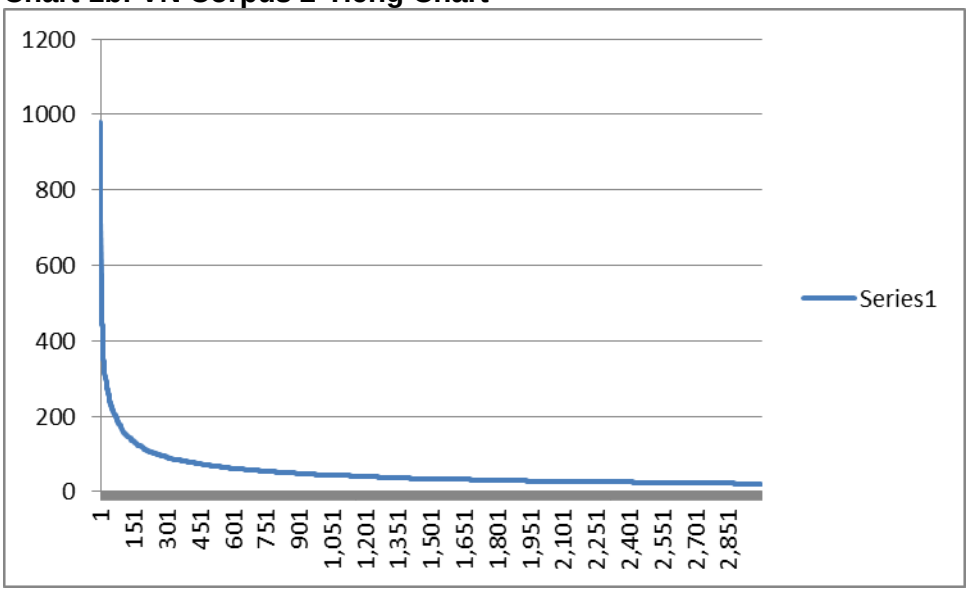


Chart 1c: VN Corpus 3-Tiếng Chart

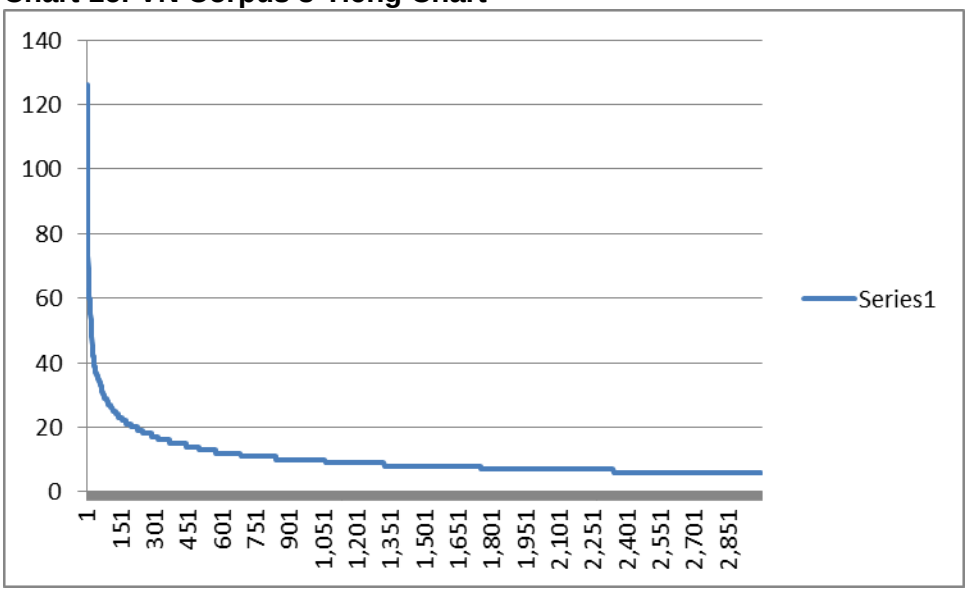


Chart 1d: VN Corpus 4-Tiếng Chart

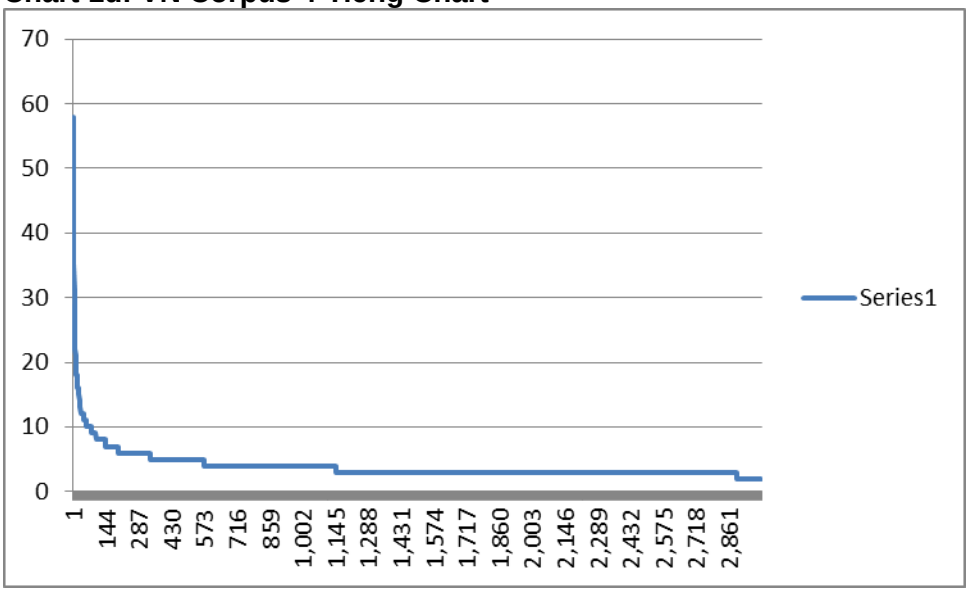
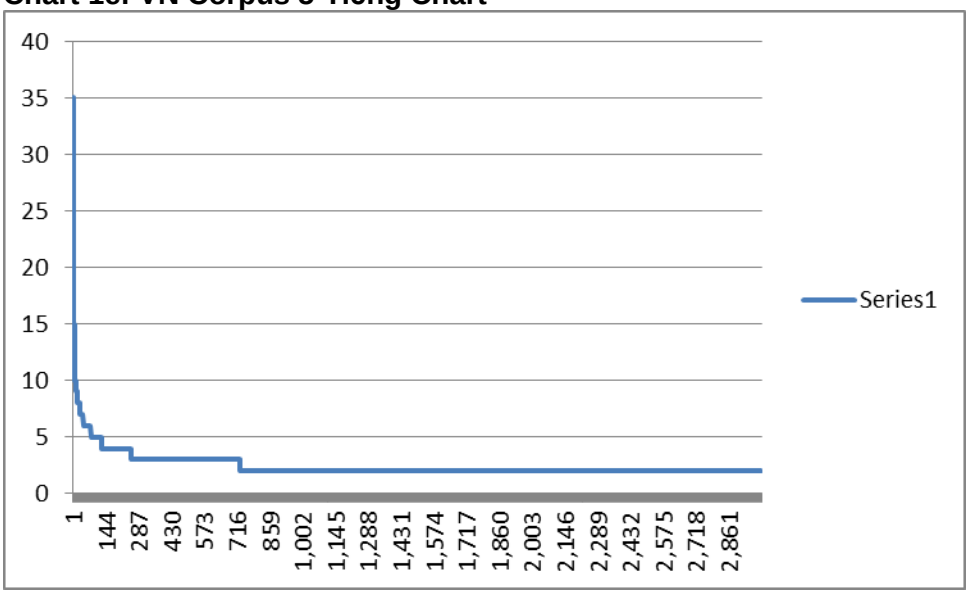


Chart 1e: VN Corpus 5-Tiếng Chart



APPENDIX T

3-CORPORA FULL FORM WORD LISTS

List 1: 1-Tiếng Full Form Words List

AUS Corpus	US Corpus	VN Corpus
#	#	#
ĂN	ĂN	AI
ANH	ANH	ĂN
BẠN	BA	ANH
BÁNH	BÀ	ẤY
BỊ	BẠN	BẠN
BIẾT	BỊ	BIẾT
CẢ	BIẾT	CẢ
CÁC	CẢ	CÁC
CÁCH	CÁC	CÁCH
CÁI	CÁI	CÁI
CHỈ	CHỈ	CẢM
CHÍNH	CHỊ	CHỈ
CHO	CHO	CHO
CHÚNG	CÓ	CHUYỆN
CHUYỆN	CÔ	CÓ
CÓ	CON	CÔ
CÔ	CÒN	CON
CON	CỦA	CÒN
CÒN	CŨNG	CỦA

CỦA	ĐÃ	CŨNG
CÙNG	ĐẦU	ĐÃ
CŨNG	ĐÂY	ĐẦU
ĐÃ	ĐỂ	ĐÂY
ĐANG	ĐẾN	ĐỂ
ĐẦU	ĐI	ĐẾN
ĐÂY	ĐÓ	ĐI
ĐỂ	ĐƯỢC	ĐÓ
ĐẾN	ĐƯỜNG	ĐƯỢC
ĐI	EM	ĐƯỜNG
ĐÓ	GÌ	EM
ĐƯỢC	GIỜ	GÌ
EM	HAI	GIỜ
GÌ	HAY	HAI
GIỜ	HẾT	HAY
HAI	HÌNH	HỌC
HAY	HỌC	HƠN
HỌ	HÔM	KHI
HỌC	HƠN	KHÔNG
HƠN	KHI	LÀ
KHÁC	KHÔNG	LẠI
KHI	LÀ	LÀM
KHÔNG	LẠI	LÊN
LÀ	LÀM	LÚC
LẠI	LÊN	MÀ
LÀM	LÚC	MẸ
LÊN	MÀ	MÌNH
MÀ	MẤY	MỚI

MÌNH	MẸ	MỘT
MỚI	MÌNH	NĂM
MỘT	MỚI	NÀO
NAM	MỘT	NÀY
NĂM	NĂM	NÊN
NÀO	NÀO	NGÀY
NÀY	NAY	NGƯỜI
NÊN	NÀY	NHÀ
NGÀY	NÊN	NHẤT
NGƯỜI	NGÀY	NHIỀU
NHÀ	NGƯỜI	NHƯ
NHẤT	NHÀ	NHƯNG
NHIỀU	NHIỀU	NHỮNG
NHƯ	NHỎ	NÓ
NHƯNG	NHƯ	NÓI
NHỮNG	NHƯNG	NƯỚC
NÓ	NHỮNG	Ở
NÓI	NÓ	ÔNG
NƯỚC	NÓI	PHẢI
Ở	NƯỚC	QUA
ÔNG	Ở	RA
PHẢI	ÔNG	RẤT
QUA	PHẢI	RỒI
RA	QUA	SAO
RẤT	QUÁ	SAU
RỒI	RA	SẼ
SAU	RẤT	SỐNG
SẼ	RỒI	TA

SỐ	SAO	THÀNH
TA	SAU	THẤY
THÀNH	SẼ	THẾ
THẤY	TA	THEO
THẾ	THẤY	THÌ
THEO	THẾ	TÌNH
THÌ	THEO	TÔI
TỚ	THÌ	TRÊN
TÔI	THÔI	TRONG
TRÊN	TÔI	TRƯỚC
TRONG	TỚI	TỪ
TRƯỚC	TRÊN	VÀ
TỪ	TRONG	VĂN
VÀ	TRƯỚC	VẪN
VÀO	TỪ	VÀO
VẬY	VÀ	VẬY
VỀ	VẪN	VỀ
VÌ	VÀO	VÌ
VIỆT	VẬY	VỚI
VỚI	VỀ	YÊU
	VÌ	
	VIỆT	
	VỚI	

List 2: 2-Tiếng Full Form Words List

AUS Corpus	US Corpus	VN Corpus
# #	BẠN BÈ	BAO GIỜ
BAO GIỜ	BAO GIỜ	BAO NHIÊU

BẮT ĐẦU	BAO NHIÊU	BẮT ĐẦU
BÂY GIỜ	BẮT ĐẦU	BÂY GIỜ
CÂU CHUYỆN	BÂY GIỜ	CÀ PHÊ
CHÍNH PHỦ	CÁM ƠN	CÁI GÌ
CHÍNH TRỊ	CHUẨN BỊ	CẢM GIÁC
CHÚNG TA	CHÚNG TA	CÂU CHUYỆN
CHÚNG TÔI	CHÚNG TÔI	CHÚNG TA
CÓ LẼ	CÓ LẼ	CHÚNG TÔI
CÓ THỂ	CÓ THỂ	CÓ LẼ
CON NGƯỜI	CUỐI CÙNG	CÓ THỂ
CÔNG THỨC	ĐẶC BIỆT	CON NGƯỜI
CUỐI CÙNG	ĐẦU TIÊN	CÔNG VIỆC
ĐẶC BIỆT	DOANH NGHIỆP	CUỘC ĐỜI
ĐẦU TIÊN	GIA ĐÌNH	CUỘC SỐNG
ĐẦU TƯ	GIÁNG SINH	CUỐI CÙNG
ĐỐI VỚI	HOA KỲ	CUỐN SÁCH
GIA ĐÌNH	HÔM NAY	ĐẶC BIỆT
HÔM NAY	HÔM QUA	ĐẦU TIÊN
HỖN HỢP	KHÔNG THỂ	ĐIỆN THOẠI
KẾT QUẢ	KINH TẾ	GIA ĐÌNH
KHOA HỌC	LÀM SAO	HÀ NỘI
KHÔNG THỂ	NGƯỜI TA	HẠNH PHÚC
LÃNH ĐẠO	NGƯỜI VIỆT	HÔM NAY
MÓN ĂN	NHƯ THẾ	KHÔNG THỂ
NGHIÊN CỨU	NHƯ VẬY	NGÀY #
NGƯỜI TA	SAU KHI	NGƯỜI TA
NGƯỜI VIỆT	TẤT CẢ	NHÀ VĂN
NHƯ THẾ	THÁNG #	NHÂN VẬT

NHƯ VẬY	THÀNH PHỐ	NHƯ THẾ
QUAN TRỌNG	THẾ GIỚI	NHƯ VẬY
SAU KHI	THỜI GIAN	PHỤ NỮ
TẤT CẢ	TRẢ LỜI	SÀI GÒN
THÁNG #	TRƯỚC KHI	SAU KHI
THÀNH PHỐ	VẪN CÒN	TÁC GIẢ
THẾ GIỚI	VẤN ĐỀ	TÁC PHẨM
THẾ NÀO	VIỆT NAM	TẤT CẢ
THỊ TRƯỜNG		THÁNG #
THỜI GIAN		THÀNH PHỐ
TIẾP TỤC		THẾ GIỚI
TRẢ LỜI		THẾ NÀO
TRỞ THÀNH		THỜI GIAN
TRƯỚC KHI		TIỂU THUYẾT
TUY NHIÊN		TÌNH YÊU
VẪN CÒN		VIỆT NAM
VẤN ĐỀ		XÃ HỘI
VĂN HÓA		
VIỆT NAM		
XÃ HỘI		

List 3: 3-Tiếng Full Form Words List

AUS Corpus	US Corpus	VN Corpus
# # #	CÂU TRẢ LỜI	CÂU TRẢ LỜI
CÂU TRẢ LỜI	NỀN KINH TẾ	MỐI QUAN HỆ
CHÍNH TRỊ GIA	TIẾNG ĐỒNG HỒ	
NỀN CÔNG NGHIỆP	TRƯỜNG ĐẠI HỌC	

NGHIÊN CỨU SINH		
NHÀ CẦM QUYỀN		
NHÀ ĐẦU TƯ		
NHÀ KHOA HỌC		
NHÀ LÃNH ĐẠO		
TIẾNG ĐỒNG HỒ		

List 4: 4-Tiếng Full Form Words List

AUS Corpus	US Corpus	VN Corpus
# # # #	CHUA CHUA NGỌT NGỌT	

APPENDIX U

3 CORPUS OPEN CLASS CONTENT FORMS LISTS

List 1: 1-Tiếng Open Class Content Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
#	7,216	#	9,562	LÀ	8,248
LÀ	5,003	LÀ	8,136	CÓ	7,814
CÓ	4,941	CÓ	7,957	MỘT	7,229
MỘT	4,168	CHO	5,902	#	5,639
CHO	2,799	MỘT	5,664	TÔI	5,348
TÔI	2,775	MÌNH	4,803	NGƯỜI	5,123
NGƯỜI	2,670	ĐI	4,339	MÌNH	4,056
MÌNH	2,562	NGƯỜI	4,325	CHO	3,942
ĐỂ	1,823	TÔI	4,191	ĐI	3,753
Ở	1,805	LÀM	3,732	CON	3,062
LÀM	1,734	CON	3,622	VỀ	2,936
VÀO	1,617	RA	3,169	RA	2,922
RA	1,592	ĐỂ	2,937	ĐỂ	2,906
BẠN	1,545	Ở	2,916	LÀM	2,887
ĐẾN	1,477	VỀ	2,839	Ở	2,864
VỀ	1,434	NHÀ	2,716	ANH	2,854
CON	1,409	NĂM	2,361	NHÀ	2,627
ĐI	1,396	NGÀY	2,347	ĐẾN	2,505
NÓI	1,280	EM	2,214	EM	2,479
TA	1,183	MẸ	2,207	BẠN	2,118

THỂ	1,174	ĂN	2,106	NGÀY	2,086
NƯỚC	1,153	VÀO	2,021	VÀO	2,033
TỪ	1,144	ANH	1,947	TỪ	1,953
NHIỀU	1,100	NÓI	1,943	THẤY	1,934
NHÀ	1,096	ĐẾN	1,880	NHIỀU	1,905
HỌC	1,088	BIẾT	1,862	NĂM	1,896
ANH	1,032	BẠN	1,823	NÓI	1,885
NĂM	1,025	QUA	1,806	THỂ	1,846
SỰ	1,015	THẤY	1,796	TA	1,722
ÔNG	1,014	TỪ	1,768	BIẾT	1,714
HAY	985	NHIỀU	1,751	CÔ	1,701
THẤY	971	MỚI	1,651	SỰ	1,554
BÁNH	968	ÔNG	1,639	MỚI	1,540
ĂN	967	LÊN	1,581	ĐẦU	1,517
VIỆT	913	ĐẦU	1,547	LÊN	1,481
CHÚNG	901	HAI	1,540	HAI	1,475
ĐẦU	898	NƯỚC	1,527	HỌC	1,413
BIẾT	896	GIỜ	1,500	QUA	1,399
LÊN	859	BỊ	1,425	ÔNG	1,395
NGÀY	856	HAY	1,418	MẸ	1,387
CÔNG	846	TA	1,344	ĂN	1,364
EM	836	CÔ	1,301	HAY	1,331
HAI	779	HÌNH	1,244	CÔNG	1,320
BỊ	760	CHỊ	1,217	GIỜ	1,316
THÀNH	757	THỂ	1,201	THÀNH	1,308
CHÍNH	742	HÔM	1,185	VIỆC	1,268
CÔ	733	NHỎ	1,178	NƯỚC	1,205
MỚI	707	BÀ	1,173	TÌNH	1,198

NAM	697	HỌC	1,169	ĐƯỜNG	1,180
NHẤT	669	BA	1,168	CẢM	1,180
GIỜ	651	VIỆT	1,167	SAO	1,129
CHUYỆN	632	SAO	1,152	YÊU	1,121
SỐ	612	ĐƯỜNG	1,124	CHUYỆN	1,099
NHÂN	611			NHẤT	1,093
TỚ	577			SỐNG	1,027

List 2: 2-Tiếng Open Class Content Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
VIỆT NAM	564	VIỆT NAM	614	LÀ MỘT	710
LÀ MỘT	517	LÀ MỘT	585	NGƯỜI TA	561
CHÚNG TA	425	HÔM NAY	503	CÓ MỘT	463
NGƯỜI TA	309	GIA ĐÌNH	495	NHỮNG NGƯỜI	458
CÓ MỘT	307	KHÔNG BIẾT	443	CỦA MÌNH	443
NHỮNG NGƯỜI	263	CÓ MỘT	414	HÀ NỘI	431
KHOẢNG #	258	CỦA MÌNH	414	THỜI GIAN	401
NĂM #	234	NĂM NAY	403	CHÚNG TÔI	400
CỦA MÌNH	232	NĂM #	395	MỘT NGƯỜI	347
NGHIÊN CỨU	226	THỜI GIAN	389	CUỘC SỐNG	336
CÁC BẠN	213	# NĂM	378	CON NGƯỜI	317
# #	210	NHỮNG NGƯỜI	371	GIA ĐÌNH	314
MỌI NGƯỜI	207	NGƯỜI TA	362	VIỆT NAM	306
THỜI GIAN	201	CHỈ CÓ	326	KHÔNG BIẾT	304
# NĂM	182	MỌI NGƯỜI	298	TÌNH YÊU	302
KHOA HỌC	182	ĐI LÀM	276	HÔM NAY	300

Ở ĐÂY	182	ĐẦU TIÊN	273	BẮT ĐẦU	287
# PHÚT	172	Ở NHÀ	267	NĂM #	281
CÔNG THỨC	171	CHÚNG TÔI	254	TÔI KHÔNG	272
CHÚNG TÔI	169	BẮT ĐẦU	253	# NĂM	272
CHỈ CÓ	168	THÁNG #	244	CHỈ CÓ	265
THẾ GIỚI	163	HÔM QUA	242	CHỈ LÀ	259
BẮT ĐẦU	159	CUỐI CÙNG	234	ĐẦU TIÊN	259
ĐẦU TIÊN	157	MỘT NGƯỜI	231	TÔI ĐÃ	254
MỘT NGƯỜI	157	CHÚNG TA	225	CÔNG VIỆC	247
VẤN ĐỀ	157	THÀNH PHỐ	221	CẢM GIÁC	247
KHÔNG BIẾT	152	MỘT CÁI	221	CỦA TÔI	237
CHÍNH TRỊ	151	CHỤP HÌNH	220	THÁNG #	235
GIA ĐÌNH	148	CÓ #	218	Ở ĐÂY	234
CỦA TÔI	145	TÔI KHÔNG	218	THÀNH PHỐ	234
ĐẶC BIỆT	145	MỘT NGÀY	208	MỌI NGƯỜI	231
LÀ NGƯỜI	141	CHO CON	208	NHÀ VĂN	227
QUAN TRỌNG	141	VỀ NHÀ	205	MỘT CÁI	223
CUỐI CÙNG	139	MÌNH KHÔNG	205	LÀ NGƯỜI	222
VĂN HÓA	139	LÀ NGƯỜI	202	HẠNH PHÚC	218
CỦA NGƯỜI	136	CỦA TÔI	202	CHÚNG TA	218
MỘT SỐ	135	KINH TẾ	201	CÂU CHUYỆN	218
CÓ #	134	BẠN BÈ	195	CON ĐƯỜNG	211
XÃ HỘI	133	KHOẢNG #	195	LÀM VIỆC	211
HÔM NAY	129	# THÁNG	194	CỦA NGƯỜI	210
RẤT NHIỀU	129	MỘT CHÚT	193	TÁC GIẢ	204
NGƯỜI VIỆT	126	CHO MÌNH	193	NHỮNG NGÀY	203
MÓN ĂN	125	CỦA NGƯỜI	191	XÃ HỘI	203
TÔI KHÔNG	125	RẤT NHIỀU	191	CUỘC ĐỜI	202

KẾT QUẢ	123	LÀ #	191	SÀI GÒN	201
CỦA HỌ	122	CHUẨN BỊ	190	MỘT SỐ	201
LÀ #	122	ANH EM	188	TIỂU THUYẾT	196
LÃNH ĐẠO	120	GIÁNG SINH	187	THẾ GIỚI	195
TÔI ĐÃ	120	MÌNH CÓ	186	CÁC BẠN	193
HỖN HỢP	118	TRẢ LỜI	185	CUỐN SÁCH	188
BẠN CÓ	117	VẤN ĐỀ	185	RẤT NHIỀU	188
CHỈ LÀ	114	NGƯỜI VIỆT	182	TÔI CŨNG	184
THỊ TRƯỜNG	114	THẾ GIỚI	182	CON GÁI	183
CON NGƯỜI	112	ĐẶC BIỆT	180	Ở NHÀ	179
GỌI LÀ	112	MẸ CON	173	MỘT NGÀY	179
Ở VIỆT	112	CÁC BẠN	173	TÁC PHẨM	177
TRẢ LỜI	112	DOANH NGHIỆP	171	NHÂN VẬT	177
THÁNG #	111	Ở ĐÂY	171	CÀ PHÊ	177
CHÍNH PHỦ	110	# GIỜ	168	ĐẶC BIỆT	175
TIẾP TỤC	109	MỘT SỐ	167	TÔI CÓ	173
NHIỀU NGƯỜI	107	LÀM VIỆC	167	CUỐI CÙNG	169
THÀNH PHỐ	106	LÁI XE	166	NGÀY #	168
ĐẦU TƯ	105	CỦA ÔNG	165	PHỤ NỮ	165
MỘT CÁI	105	HOA KỲ	165	MỘT CHÚT	165
TRỞ THÀNH	105	NHỮNG NGÀY	164	NGƯỜI BẠN	164
CÂU CHUYỆN	104	CÁM ƠN	163	ĐIỆN THOẠI	163

List 3: 3-Tiếng Open Class Content Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
Ở VIỆT NAM	112	LẦN ĐẦU TIÊN	96	LẦN ĐẦU TIÊN	79
BẠN CÓ THỂ	76	# THÁNG #	91	# THÁNG #	69

NHÀ LÃNH ĐẠO	57	Ở VIỆT NAM	83	NGƯỜI ĐÀN ÔNG	63
TA CÓ THỂ	51	NGÀY # THÁNG	68	BẠN CÓ THỂ	63
# # #	43	TƯỜNG VI #	66	ĐÓ LÀ MỘT	62
TRÊN THẾ GIỚI	42	TIẾNG ĐỒNG HỒ	64	NGÀY # THÁNG	59
ĐÂY LÀ MỘT	41	NGƯỜI MỸ GỐC	61	NGƯỜI PHỤ NỮ	59
ĐÓ LÀ MỘT	40	CÁC DOANH NGHIỆP	58	ĐỖ HỒNG NGỌC	55
CỦA CHÚNG TA	39	CHỈ CÓ #	53	NHỮNG CÂU CHUYỆN	55
KHOẢNG # PHÚT	39	NÓI CHUYỆN VỚI	52	MỘT NGƯỜI BẠN	52
CÓ THỂ LÀM	38	CỦA VIỆT NAM	52	CỦA NHỮNG NGƯỜI	48
LẦN ĐẦU TIÊN	38	ĐI LÀM VỀ	52	CHỈ LÀ MỘT	48
NGHIÊN CỨU SINH	36	MỸ GỐC VIỆT	51	TA CÓ THỂ	44
VỀ NHÂN QUYỀN	36	MÙA GIÁNG SINH	49	CUỘC SỐNG CỦA	43
CÓ THỂ NÓI	35	CÓ THỜI GIAN	47	KHÔNG CÓ GÌ	43
NÓI CHUYỆN VỚI	35	ĐÂY LÀ MỘT	46	Ở VIỆT NAM	42
# ĐẾN #	34	MÌNH CÓ THỂ	44	NHỮNG NGƯỜI BẠN	41
CÓ NGHĨA LÀ	34	CÓ RẤT NHIỀU	44	BÂY GIỜ LÀ	41
CUNG TRẦM TƯỞNG	34	CỦA NHỮNG NGƯỜI	44	CŨNG LÀ MỘT	41
TIẾNG ĐỒNG HỒ	34	KHÔNG BIẾT CÓ	44	TÔI CÓ THỂ	41
MỘT THỜI GIAN	32	NÓI CHUNG LÀ	42	CÓ THỂ NÓI	39
NGƯỜI VIỆT NAM	32	# MẸ CON	41	Ở SÀI GÒN	39
CHÚNG TA CÓ	31	MỘT THỜI GIAN	41	ĐIỀU GÌ ĐÓ	39
CUỘC ĐỐI THOẠI	31	ĐÔNG NAM Á	40	NGƯỜI ĐÀN BÀ	38
# ĐỘ C	30	MÌNH KHÔNG CÓ	40	CÓ CẢM GIÁC	38
ĐẢNG LAO ĐỘNG	30	ANH CHỊ EM	40	MỘT CUỐN SÁCH	38
LÀ NHỮNG NGƯỜI	30	# NĂM #	38	CÓ THỂ LÀ	37
TRỞ THÀNH MỘT	30	TRƯỜNG ĐẠI HỌC	38	TRỞ THÀNH MỘT	37
CHỈ LÀ MỘT	29	CÓ CẢM GIÁC	38	NHỮNG CON	36

				ĐƯỜNG	
CÓ KHẢ NĂNG	29	HƠN # NĂM	38	MỐI QUAN HỆ	36
ĐƯỢC XEM LÀ	29	THÁNG # NĂM	37	LÀ MỘT NGƯỜI	36
NẾU KHÔNG CÓ	29	CÁC ANH CHỊ	37	CÓ NGHĨA LÀ	36
CÓ CƠ HỘI	28	CHO GIA ĐÌNH	37	EM LẦN NỮA	36
CỘNG ĐỒNG NGƯỜI	28	# TIẾNG ĐỒNG	37	CÂU TRẢ LỜI	35
ĐỒNG NGƯỜI VIỆT	28	BIẾT BAO NHIÊU	37	CÓ RẤT NHIỀU	35
NHÂN QUYỀN Ở	28	CHỈ CÓ MỘT	36	MỘT CÂU CHUYỆN	35
VỀ VIỆT NAM	28	CỦA CHÚNG TA	36	ĐỂ HÔN EM	35
KHÔNG CÓ GÌ	27	LÀ MỘT NGƯỜI	36	HÔN EM LẦN	35
TRONG VÒNG #	27	MỌI NGƯỜI ĐỀU	36	Ở HÀ NỘI	34
CÂU TRẢ LỜI	26	KHÔNG CÓ GÌ	35	CHO NGƯỜI TA	34
CHÚNG TA PHẢI	26	CÓ CƠ HỘI	35	QUÁN CÀ PHÊ	34
CỨU KHOA HỌC	26	# GIỜ SÁNG	34	PHÁT HIỆN RA	34
LÀ MỘT SỰ	26	HAI MẸ CON	34	CÓ NHỮNG NGƯỜI	33
NGHIÊN CỨU KHOA	26	TRÊN THẾ GIỚI	34	CỦA CON NGƯỜI	33
CÔNG NGHIỆP VĂN	25	CÓ THỂ LÀM	33	HỘI NHÀ VĂN	33
CŨNG LÀ MỘT	25	NGÀY HÔM NAY	33	CỦA MỘT NGƯỜI	33
NỀN CÔNG NGHIỆP	25	CHỈ LÀ MỘT	33	# NĂM TRƯỚC	32
NGHIỆP VĂN HÓA	25	CHO MỌI NGƯỜI	33	TRONG CUỘC SỐNG	32
{Last} {Middle} {First}	25	NGƯỜI VIỆT NAM	33	CÓ THỂ LÀM	31
TRONG THỜI GIAN	25	# NĂM NAY	32	VỚI NHỮNG NGƯỜI	31
VÀO HỖN HỢP	25	LÂU LẮM RỒI	32	CỦA NHÀ VĂN	31
VỚI NHỮNG NGƯỜI	25	CÂU TRẢ LỜI	32	NÓI CHUYỆN VỚI	31
# QUẢ TRỨNG	24	CHỪNG # PHÚT	32	NGƯỜI TA KHÔNG	30
ẤN PHẨM KHOA	24	GẦN # NĂM	32	ĐÂY LÀ MỘT	30

CÓ THỂ THAY	24	CŨNG LÀ MỘT	32	CUỐN TIỂU THUYẾT	30
LIÊN QUAN ĐẾN	24	TRONG VÒNG #	31	KHÔNG HIỂU SAO	30
MỘT NGƯỜI BẠN	24	TỪ NĂM #	31	QUAN TÂM ĐẾN	30
NHÀ CẦM QUYỀN	24	TÔI CÓ THỂ	31	LÀ NHỮNG NGƯỜI	29
NHÀ KHOA HỌC	24	CÁC NGÂN HÀNG	31	NHỮNG CUỐN SÁCH	29
PHẨM KHOA HỌC	24	VỚI GIA ĐÌNH	31	LIÊN QUAN ĐẾN	29
BLOG {Name} {Name}	23	ANH EM {Name}	30	ĐẦU TIÊN CỦA	29
THÌA CÀ PHÊ	23	CÓ THỂ LÀ	30	HAI VỢ CHỒNG	29
VÀO TỦ LẠNH	23	GIA ĐÌNH MÌNH	30	CẤP THÂM NIÊN	29
VIỆT NAM VÀ	23	II CÁCH LÀM	30	CỦA TÁC GIẢ	29
CÁC CUỘC ĐỐI	22	NGÀY NÀO CŨNG	30	PHỤ CẤP THÂM	29
CHỈ CÓ MỘT	22	TRÊN ĐƯỜNG VỀ	30	CHỈ CÓ MỘT	29
CHÚNG TA KHÔNG	22	XỨ CAO BỒI	30	ĐÃ TRỞ THÀNH	29
CÓ THỂ LÀ	22	CẢ MỌI NGƯỜI	30	BẮT ĐẦU TỪ	28
Ở ĐÂY LÀ	22	CÓ KHẢ NĂNG	30	{Last} {Middle} {First}	28
QUAN TÂM ĐẾN	22	ĐẦU TIÊN CỦA	30	NGÀY HÔM NAY	28
VÀO NĂM #	22	TÔI KHÔNG CÓ	29	CÁC TÁC PHẨM	28
BÀI PHÁT BIỂU	21	HAI VỢ CHỒNG	29	CÙ LAO RÙA	28
CHỈ CÓ #	21	MỘT NGƯỜI BẠN	29	HAI MẸ CON	28
CHÍNH TRỊ GIA	21	# NĂM TRƯỚC	29	MỘT NGƯỜI ĐÀN	27
CHÚNG TÔI ĐÃ	21	VÀI TẤM HÌNH	29	NHỮNG CHI TIẾT	27
CÓ THỂ DÙNG	21	THỦ ĐÔ DC	29	ĐƠN GIẢN LÀ	27
CÔNG THỨC NÀY	21	TÔI KHÔNG BIẾT	29	HỒ CHÍ MINH	27
CỦA NGƯỜI VIỆT	21	CỦA GIA ĐÌNH	28		
CỦA VIỆT NAM	21				
HỌC VIỆT NAM	21				
NHÀ ĐẦU TƯ	21				

List 4: 4-Tiếng Open Class Content Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
CỘNG ĐỒNG NGƯỜI VIỆT	28	NGÀY # THÁNG #	60	NGÀY # THÁNG #	58
NGHIÊN CỨU KHOA HỌC	26	LÀ MỘT TRONG NHỮNG	41	HÔN EM LẦN NỮA	35
CÔNG NGHIỆP VĂN HÓA	25	NGƯỜI MỸ GỐC VIỆT	39	ĐỂ HÔN EM LẦN	35
NỀN CÔNG NGHIỆP VĂN	25	# TIẾNG ĐỒNG HỒ	36	LÀ MỘT TRONG NHỮNG	31
ẤN PHẨM KHOA HỌC	24	THÁNG # NĂM #	31	PHỤ CẤP THÂM NIÊN	29
CÁC CUỘC ĐỐI THOẠI	22	TẤT CẢ MỌI NGƯỜI	30	BÂY GIỜ LÀ THÁNG	23
LÀ MỘT TRONG NHỮNG	22	DOANH NGHIỆP NHÀ NƯỚC	24	VĂN MIẾU TRẤN BIÊN	21
BÀI BÁO KHOA HỌC	20	MỌI NGƯỜI ĐỀU CÓ	23	MỘT NGƯỜI ĐÀN ÔNG	21
CHÚNG TA CÓ THỂ	20	BIỂN ĐÔNG NAM Á	22	TIỂU THUYẾT CHIẾN TRANH	20
NHÂN QUYỀN Ở VIỆT	19	NGƯỜI ĐỀU CÓ QUYỀN	21	PHỤ NỮ VIỆT NAM	20
QUYỀN Ở VIỆT NAM	19	I NGUYÊN LIỆU AMP	21	BỘ PHẬN KHÔNG NHỎ	19
# # # #	17	ANH EM CHÚNG TÔI	21	BÂY GIỜ LÀ THÁNG #	18
# THÌA CÀ PHÊ	17	# MUỖNG CÀ PHÊ	20	GIỜ LÀ THÁNG #	18
ĐỐI THOẠI VỀ NHÂN	17	BẢO HIỂM SỨC KHỎE	20	TÍN NGƯỠNG THỜ MẪU	18
NGƯỜI TA CÓ THỂ	17	LÀ LẦN ĐẦU TIÊN	19	LÀ LẦN ĐẦU TIÊN	17
THỊ TRƯỜNG CHỨNG KHOÁN	17	BẢO HIỂM Y TẾ	18	NHÀ VĂN VIỆT NAM	16
THOẠI VỀ NHÂN QUYỀN	17	KHÔNG CÓ THỜI GIAN	18	TÁC PHẨM VĂN HỌC	16
# TIẾNG ĐỒNG HỒ	16	KHÔNG PHẢI LÀ MỘT	18	KỶ NIỆM # NĂM	16
SỐ ẤN PHẨM KHOA	16	MẸ CON NHÀ {Name}	18	VĂN HỌC NGHỆ THUẬT	16

CÔNG BỐ QUỐC TẾ	15	{Name} VÀ {Name} {Name}	18	BẢO TÀNG PHỤ NỮ	16
ĐÀO TẠO TIẾN SĨ	15	TỐI CAO PHÁP VIỆN	17	KHÔNG PHẢI LÀ MỘT	15
THÁNG # NĂM #	15	# THÁNG # NĂM	16	TÀNG PHỤ NỮ VIỆT	15
VĂN HÓA BÌNH DÂN	15	HỆ THỐNG NGÂN HÀNG	16	VIẾT VỀ CHIẾN TRANH	15
KHOA HỌC VIỆT NAM	14	TRẠI TỊ NẠN BATAAN	16	BẢO TÀNG PHỤ NỮ VIỆT	15
# CỦ HÀNH TÂY	13	# THÁNG # NĂM #	16	HỘI NHÀ VĂN VIỆT	14
CẦM QUYỀN VIỆT NAM	13	CHÍNH PHỦ VIỆT NAM	15	CHẾ ĐỘ PHỤ CẤP	14
CON SỐ THỐNG KÊ	13	TÔI VÀ BẠN TÔI	15	THÁNG # NĂM #	13
CUỘC ĐỐI THOẠI VỀ	13	ĐÔNG NAM Á CHÂU	15	CUỘC SỐNG CỦA MÌNH	13
TẬP SAN QUỐC TẾ	13	ANH EM NHÀ {Name}	15	LÀM ƠN IM ĐI	13
TRÊN CÁC TẬP SAN	13	TỪ # ĐẾN #	15	NHỮNG NGƯỜI PHỤ NỮ	12
TỪ # ĐẾN #	13	ĐÂY LÀ LẦN ĐẦU	15	CÁCH ĐÂY # NĂM	12
VĂN HÓA ĐẠI CHÚNG	13	KHÔNG BIẾT BAO NHIÊU	14	EM LÀM ƠN IM	12
BẢN DỊCH CỦA TÚ	12	CÁC THẦY CÔ GIÁO	14	CỦA TRẦN THU TRANG	12
CÁC CHÍNH TRỊ GIA	12	ANH {Name} VÀ {Name}	14	VĂN NGHỆ QUÂN ĐỘI	12
CỰU CHIẾN BINH ÚC	12	MỘT MÙA GIÁNG SINH	13	KHÔNG NHẤT THIẾT PHẢI	12
ĐẾN KHI HỖN HỢP	12	CÁC NHÀ ĐẦU TƯ	13	THỊ TRẤN TORTILLA FLAT	12
DỊCH CỦA {Name} {Name}	12	NATIONAL GALLERY OF ART	13	VẤN ĐỀ XÃ HỘI	12
HỮU NỮ NHAN NHƯ	12	LẦN ĐẦU TIÊN TÔI	13	MỘT ĐIỀU GÌ ĐÓ	12
NỮ NHAN NHƯ NGỌC	12	XỨ CAO BỒI NÀY	13	NHỮNG NGƯỜI XUNG QUANH	12
TÌNH TRẠNG NHÂN QUYỀN	12	Á CHÂU THÁI BÌNH	13	THÀNH PHỐ BIÊN HÒA	12
TRẠNG NHÂN QUYỀN Ở	12	LỰA CHỌN THAY THẾ	13	CÔNG TÁC XÃ HỘI	12

CÁI NƯỚC MÌNH NÓ	11	CÁC ANH CHỊ EM	13	# THÁNG # NĂM	11
CÓ THỂ LÀM ĐƯỢC	11	DOANH NGHIỆP TƯ NHÂN	13	NHỮNG NGƯỜI ĐÀN ÔNG	11
CÓ THỂ THAY BẰNG	11	# NĂM VỀ TRƯỚC	12	NGƯỜI TA CÓ THỂ	11
CỦA NỀN CÔNG NGHIỆP	11	CÁC DOANH NGHIỆP NHÀ	12	TẠP CHÍ VĂN NGHỆ	11
MỘT NHÀ LÃNH ĐẠO	11	TIẾNG NÓI NGƯỜI MỸ	12	CUỘC SỐNG GIA ĐÌNH	11
NGÀY # THÁNG #	11	NÓI NGƯỜI MỸ GỐC	12	MỘT BỘ PHẬN KHÔNG	11
NHÀ CẦM QUYỀN VIỆT	11	VOICE OF VIETNAMESE AMERICANS	12	NHỮNG CÁNH ĐỒNG HOA	11
NƯỚC MÌNH NÓ THẾ	11	CÁCH ĐÂY # NĂM	12	COCKTAIL CHO TÌNH YÊU	11
TRUYỀN THÔNG XÃ HỘI	11	CHÚNG TA CÓ THỂ	12	ĐẦU TIÊN TRONG ĐỜI	11
# QUẢ TRỨNG GÀ	10	CÔNG ĂN VIỆC LÀM	12	MỘT NGƯỜI PHỤ NỮ	11
BÀI PHÁT BIỂU CỦA	10	ĐI TỚI ĐI LUI	12	# THỜI GIAN HOÀN	10
BẠN CÓ THỂ THAY	10	CHÂU THÁI BÌNH DƯƠNG	12	# TIẾNG ĐỒNG HỒ	10
CÁ HỒI HUN KHÓI	10	HÔM NAY LÀ NGÀY	11	ANH VÀ EM #	10
CÁC MẠNG LƯỚI TRUYỀN	10	MẮM CHANH TỎI ỚT	11	CẢNH SÁT GIAO THÔNG	10
CÁC NHÀ KHOA HỌC	10	TUYÊN NGÔN QUỐC TẾ	11	MẬT MÃ TÂY TẠNG	10
CÁC TẬP SAN QUỐC	10	NƯỚC MẮM CHANH TỎI	11	# THỜI GIAN HOÀN THÀNH	10
CÔNG TRÌNH NGHIÊN CỨU	10	TRONG VÒNG # NĂM	11	CÁC EM HỌC SINH	10
ĐỐI VỚI NHỮNG NGƯỜI	10	KINH TẾ VIỆT NAM	11	THƯỞNG HỘI NHÀ VĂN	10
LƯỚI TRUYỀN THÔNG XÃ	10	NĂM TRỞ LẠI ĐÂY	11	ÔNG GIÀ BA BỊ	10
MẠNG LƯỚI TRUYỀN THÔNG	10	NGÔN QUỐC TẾ NHÂN	11	ĐÂY LÀ LẦN ĐẦU	10
Ở NHIỆT ĐỘ PHÒNG	10	NGƯỜI CÔ GIÁO TRẺ	11	THỜI GIAN KHỞI CÔNG	10

VẤN ĐỀ NHÂN QUYỀN	10	QUỐC TẾ NHÂN QUYỀN	11	ĐỘ PHỤ CẤP THÂM	10
VỀ MẶT CHÍNH TRỊ	10	# GIỜ ĐỒNG HỒ	10	GIẢI THƯỞNG HỘI NHÀ	10
# BÀI BÁO KHOA	9	MỘT NGÀY NÀO ĐÓ	10	TÁC GIẢ SỬ DỤNG	10
# BỘT MÌ #	9	MỘT SỐ TRƯỜNG HỢP	10	NXB HỘI NHÀ VĂN	10
# THÁNG # NĂM	9	QUỐC HỘI HOA KỲ	10	THỜI GIAN HOÀN THÀNH	10
BẠN CÓ THỂ DÙNG	9	TRÊN BIỂN ĐÔNG NAM	10	LẦN ĐẦU TIÊN TRONG	10
CHIẾN TRANH VIỆT NAM	9	CHIẾC LÁ CUỐI CÙNG	10	KHAI THẬT ĐẦU ANH	10
CHO VÀO TỦ LẠNH	9	DÙNG NÓNG VỚI CƠM	10	TẤT CẢ NHỮNG ĐIỀU	10
CHỦ NGHĨA XÃ HỘI	9	TRONG ĐÔI MẮT EM	10	TP HỒ CHÍ MINH	9
DÂN BIỂU CHRIS HAYES	9	CHUA CHUA NGỌT NGỌT	10	TÔI KHÔNG BAO GIỜ	9
HABIBI YA NOUR EL	9	CÓ CẢM GIÁC NHƯ	10	MỘT CUỐN TIỂU THUYẾT	9
HỘI ĐỒNG THÀNH PHỐ	9	MẤY MẸ CON NHÀ	10	BS {Last} {Middle} {First}	9
KẾT QUẢ NGHIÊN CỨU	9	TỰ DO NGÔN LUẬN	10	CẦN PHẢI KHAI THẬT	9
MỸ VÀ TRUNG QUỐC	9	LÀM GÌ THÌ LÀM	10	KHÔNG BAO GIỜ QUÊN	9
NGƯỜI DÂN VIỆT NAM	9	MỸ GỐC Á CHÂU	10	KHÔNG CẦN PHẢI KHAI	9
NGƯỜI VIỆT TỴ NẠN	9	CÁC DOANH NGHIỆP TƯ	10	CỦA CÁC TÁC GIẢ	9
SIDE OF THE WORLD	9	GỐC Á CHÂU THÁI	9	PHẢI KHAI THẬT ĐÂU	9
THE WRONG SIDE OF	9	LUẬT BẢO HIỂM SỨC	9	CHÚNG TA CÓ THỂ	9
THỜI CHIẾN TRANH LẠNH	9	AN SINH XÃ HỘI	9	KHÔNG PHẢI LÀ NGƯỜI	9
ỦY HỘI SÔNG MEKONG	9	CÓ NHIỀU LỰA CHỌN	9	CHÚNG TÔI QUYẾT ĐỊNH	9
VỀ NHÂN QUYỀN VỚI	9	EM MÃI LÀ #	9	TRÁI TIM BẠC NHƯỢC	9
WRONG SIDE OF	9	CHÂU Á THÁI BÌNH	9	KHOA HỌC KỸ	9

THE				THUẬT	
XÃ HỘI CHỦ NGHĨA	9	KHÔNG BIẾT CÓ PHẢI	9	TRUNG TÂM THÀNH PHỐ	9
XIN BẤM VÀO LINK	9	CÔ GIÁO TRẺ NĂM	9	TÔI CHƯA BAO GIỜ	9
# ĐƯỜNG CÁT #	8	I VẬT LIỆU AMP	9	RẠCH GẦM XOÀI MÚT	9
# Ở ĐÂY NHÉ	8	MỘT NỀN KINH TẾ	9	ẢNH SÁCH THẬT BÌA	8
# TSP MUỐI #	8	CHẾ ĐỘ ĐỘC TÀI	9		
ĂN GIAN VÀ ĂN	8	MẤY CON CẢM CÚM	9		
ÁNH SÁNG NHÂN TẠO	8	Á THÁI BÌNH DƯƠNG	9		
BÁC SĨ VÀ Y	8	CÁC TRƯỜNG ĐẠI HỌC	9		
BỐ ĐƯỢC # BÀI	8	KHÔNG CÓ GÌ ĐỂ	9		
CHỦ ĐỀ THÁNG #	8	LÀ NGƯỜI ĐẦU TIÊN	9		
CƠ SỞ VẬT CHẤT	8	MỘT NGƯỜI ĐÀN BÀ	9		
CÔNG BỐ ĐƯỢC #	8	CẢ NHÀ KÉO NHAU	9		
		CÁC HÃNG BẢO HIỂM	9		
		GIÁO TRẺ NĂM XƯA	9		

List 5: 5-Tiếng Open Class Content Forms

AUS		**US**		**VN**	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
NỀN CÔNG NGHIỆP VĂN HÓA	25	MỌI NGƯỜI ĐỀU CÓ QUYỀN	19	ĐỂ HÔN EM LẦN NỮA	35
NHÂN QUYỀN Ở VIỆT NAM	19	NGÀY # THÁNG # NĂM	16	TÀNG PHỤ NỮ VIỆT NAM	15
ĐỐI THOẠI VỀ NHÂN QUYỀN	17	ANH {Name} VÀ {Name} {Name}	14	HỘI NHÀ VĂN VIỆT NAM	14
SỐ ẤN PHẨM KHOA HỌC	16	BIỂN ĐÔNG NAM Á CHÂU	12	EM LÀM ƠN IM ĐI	12

CUỘC ĐỐI THOẠI VỀ NHÂN	13	TIẾNG NÓI NGƯỜI MỸ GỐC	12	MỘT BỘ PHẬN KHÔNG NHỎ	11
BẢN DỊCH CỦA TÚ TRINH	12	CÁC DOANH NGHIỆP NHÀ NƯỚC	12	ĐỘ PHỤ CẤP THÂM NIÊN	10
HỮU NỮ NHAN NHƯ NGỌC	12	Á CHÂU THÁI BÌNH DƯƠNG	12	CHẾ ĐỘ PHỤ CẤP THÂM	10
TÌNH TRẠNG NHÂN QUYỀN Ở	12	NÓI NGƯỜI MỸ GỐC VIỆT	12	GIẢI THƯỞNG HỘI NHÀ VĂN	10
CỦA NỀN CÔNG NGHIỆP VĂN	11	ĐÂY LÀ LẦN ĐẦU TIÊN	11	PHẢI KHAI THẬT ĐÂU ANH	9
NHÀ CẦM QUYỀN VIỆT NAM	11	NGÔN QUỐC TẾ NHÂN QUYỀN	11	CẦN PHẢI KHAI THẬT ĐÂU	9
CÁC MẠNG LƯỚI TRUYỀN THÔNG	10	TUYÊN NGÔN QUỐC TẾ NHÂN	11	KHÔNG CẦN PHẢI KHAI THẬT	9
CÁC TẬP SAN QUỐC TẾ	10	TRÊN BIỂN ĐÔNG NAM Á	10	NGÀY # THÁNG # NĂM	9
CÁI NƯỚC MÌNH NÓ THẾ	10	NƯỚC MẮM CHANH TỎI ỚT	10	# THÁNG # NĂM #	8
LƯỚI TRUYỀN THÔNG XÃ HỘI	10	MÙA XUÂN TRONG ĐÔI MẮT	9	BẤM VÀO ĐỂ XEM HÌNH	8
MẠNG LƯỚI TRUYỀN THÔNG XÃ	10	NGƯỜI MỸ GỐC Á CHÂU	9	BẤM VÀO ĐỂ XEM	8
TRÊN CÁC TẬP SAN QUỐC	10	CÔ GIÁO TRẺ NĂM XƯA	9	TẠP CHÍ VĂN NGHỆ QUÂN	8
# BÀI BÁO KHOA HỌC	9	GỐC Á CHÂU THÁI BÌNH	9	VÀO ĐỂ XEM HÌNH CỠ	8
# THÁNG # NĂM #	9	LUẬT BẢO HIỂM SỨC KHỎE	9	KẾ BÌA BẤM VÀO ĐỂ	8
THE WRONG SIDE OF THE	9	CHÂU Á THÁI BÌNH DƯƠNG	9	THƯỞNG HỘI NHÀ VĂN VIỆT	8
WRONG SIDE OF THE WORLD	9	MẤY MẸ CON NHÀ {Name}	9	QUY TẮC LÀM CHA MẸ	8
# QUẢ TRỨNG GÀ #	8	XUÂN TRONG ĐÔI MẮT EM	9	BÌA BẤM VÀO ĐỂ XEM	8
ĂN GIAN VÀ ĂN CƯỚP	8	CỘNG ĐỒNG NGƯỜI MỸ GỐC	8	BỘ THIẾT KẾ BÌA BẤM	8
BÁC SĨ VÀ Y TÁ	8	EM MÃI LÀ # TUỔI	8	CHÍ VĂN NGHỆ QUÂN ĐỘI	8
CÁC CUỘC ĐỐI THOẠI VỀ	8	CÁC TÁC PHẨM NGHỆ THUẬT	8	ĐỂ XEM HÌNH CỠ LỚN	8
CÔNG BỐ ĐƯỢC # BÀI	8	MỸ GỐC Á CHÂU	8	THIẾT KẾ BÌA BẤM	8

		THÁI		VÀO	
CỘNG ĐỒNG NGƯỜI VIỆT Ở	8	NGÀY MỘT TẤM HÌNH #	8	TOÀN BỘ THIẾT KẾ BÌA	8
CỦA CỘNG ĐỒNG NGƯỜI VIỆT	8	CÁC DOANH NGHIỆP TƯ NHÂN	8	ĐÂY LÀ LẦN ĐẦU TIÊN	8
HỌC TRÊN CÁC TẬP SAN	8	PHILIPPINES TRẠI TỊ NẠN BATAAN	8	ẢNH SÁCH THẬT BÌA #	8
KHOA HỌC TRÊN CÁC TẬP	8	CÔNG PHÁP QUỐC TẾ LUẬT	8	BAC SI {Last} {Middle} {First}	7
PHẦN # Ở ĐÂY NHÉ	8	MỖI NGÀY MỘT TẤM HÌNH	8	ẤM KỂ TEA NGHE #	7
THOẠI VỀ NHÂN QUYỀN VỚI	8	# NĂM TRỞ LẠI ĐÂY	7	BAO GIỜ CHO ĐẾN THÁNG	7
TRẠNG NHÂN QUYỀN Ở VIỆT	8	CẦU NGUYỆN CHO HÒA BÌNH	7	BIÊN HÒA THỜI GIAN KHỞI	7
BỐ ĐƯỢC # BÀI BÁO	7	KHOẢNG # PHÚT LÁI XE	7	CÁNH ĐỒNG HOA HƯỚNG DƯƠNG	7
BO MUỐN ĐI BỘ VỀ	7	PHÍ XỬ LÝ HỒ SƠ	7	GIÁM KHẢO KHÔNG CHÍNH THỨC	7
ĐƯỢC # BÀI BÁO KHOA	7	CÁC LỰA CHỌN THAY THẾ	7	PHỐ BIÊN HÒA THỜI GIAN	7
HAPPY NEW YEAR HAPPY NEW	7	CUỘC CÁCH MẠNG KHKT LẦN	7	THÀNH PHỐ BIÊN HÒA THỜI	7
LỤC BÁT CUNG TRẦM TƯỞNG	7	PHÁP QUỐC TẾ LUẬT BIỂN	7	HÒA THỜI GIAN KHỞI CÔNG	7
NEW YEAR HAPPY NEW YEAR	7	CÁC CÔNG TY BẢO HIỂM	7	KHI CẬU GẶP CẬU TA	7
NHỮNG CON SỐ THỐNG KÊ	7	CHO CÁC THẦY CÔ GIÁO	7	LẦN ĐẦU TIÊN TRONG ĐỜI	7
ON THE WRONG SIDE OF	7	CỘNG ĐỒNG PHI VÀ VIỆT	7	HIỆP SĨ KHÔNG HIỆN HỮU	7
PHẨM CỦA NỀN CÔNG NGHIỆP	7	ĐỪNG XOA EM ĐÊM NAY	7	BÀY TÍN NGƯỠNG THỜ MẪU	6
SẢN PHẨM CỦA NỀN CÔNG	7	GIỮ LẠI MỘT NÉT THU	7	TRƯNG BÀY TÍN NGƯỠNG THỜ	6
THE COMMITTEE RECOMMENDS THAT THE	7	BÌNH TRÊN BIỂN ĐÔNG NAM	7	TRUYỆN NGẮN CỦA R CARVER	6
XEM PHẦN # Ở ĐÂY	7	CHO HÒA BÌNH TRÊN BIỂN	7	PHẢI LẤY NGƯỜI NHƯ ANH	6
BÀI BÁO KHOA HỌC	6	HÒA BÌNH TRÊN	7	LÀM TAN NÁT LÒNG	6

TRÊN		BIỂN ĐÔNG		NHAU	
BẤM VÀO LINK SAU ĐÂY	6	NGƯỜI CÔ GIÁO TRẺ NĂM	7	NHÂN VIÊN CÔNG TÁC XÃ	6
BẠN CHẲNG BUỒN NGHĨ ĐẾN	6	TRONG BƯỚC THU VỀ #	7	NHỮNG CÁNH ĐỒNG HOA HƯỚNG	6
BAN CHẤP HÀNH TRUNG ƯƠNG	6	# MUỖNG CÀ PHÊ DẦU	6	VIẾT VĂN TRẺ TOÀN QUỐC	6
BÁO KHOA HỌC TRÊN CÁC	6	LẦN ĐẦU TIÊN TRONG ĐỜI	6	GIỜ CHO ĐẾN THÁNG MƯỜI	6
CHẲNG BUỒN NGHĨ ĐẾN VIỆC	6	CÁC DOANH NGHIỆP VIỆT NAM	6	NHỮNG QUY TẮC LÀM CHA	6
CHÂU Á THÁI BÌNH DƯƠNG	6	TẤT CẢ MỌI NGƯỜI ĐỀU	6	VIÊN CÔNG TÁC XÃ HỘI	6
CHO BÁC SĨ VÀ Y	6	CÁI XẤU CỦA HÌNH ĐẸP	6	ĐI VỀ PHÍA KHÔNG NHAU	6
ĐA VĂN HÓA SỰ VỤ	6	HƠN # TIẾNG ĐỒNG HỒ	6	GIAO TIẾP PHI NGÔN NGỮ	6
DỊCH TỪ BẢN TIẾNG ANH	6	ANH {Name} VÀ CHỊ {Name}	6	LÀ MỘT NGƯỜI ĐÀN ÔNG	6
GIỮA MỸ VÀ TRUNG QUỐC	6	CÁI ĐẸP CỦA HÌNH XẤU	6	LỜI LÀM TAN NÁT LÒNG	6
HAPPY NEW YEAR MAY WE	6	CHỪNG # MUỖNG CÀ PHÊ	6	THƯ CHO BÉ SƠ SINH	6
HƯ VÔ HOÁ BẤT HẠNH	6	ĐÂY NGHE EM VỀ ĐÂY	6	NHÀ THƠ HỒ NGỌC SƠN	6
KHẮP NƠI TRÊN THẾ GIỚI	6	NGHỊCH LÝ CỦA LỰA CHỌN	6	CÔNG TÁC XÃ HỘI TRONG	6
NEW YEAR MAY WE ALL	6	TẠI BIỂN ĐÔNG NAM Á	6	MỘT LỜI LÀM TAN NÁT	6
THÀNH PHỐ HỒ CHÍ MINH	6	CỦA TẤT CẢ MỌI NGƯỜI	6	VỚI TẤT CẢ MỌI NGƯỜI	6
THƠ CỦA JUAN RAMÓN JIMÉNEZ	6	HÔM QUA ĐI LÀM VỀ	6	TÁC XÃ HỘI TRONG BỆNH	6
VỀ CÁC CUỘC ĐỐI THOẠI	6	NÊM NẾM LẠI CHO VỪA	6	NGHỊ VIẾT VĂN TRẺ TOÀN	6
VỀ CHIẾN TRANH VIỆT NAM	6	NGƯỜI ĐỀU CÓ QUYỀN ĐƯỢC	6	TOÀN VỆ SINH THỰC PHẨM	6
VỀ TÌNH TRẠNG NHÂN QUYỀN	6	CHIỀU THÀNH PHỐ MƯA BAY	6	TRONG DÒNG SÔNG CỦA HERACLITUS	6
YEAR HAPPY NEW YEAR MAY	6	VỀ ĐÂY NGHE EM VỀ	6	HÀM ĐAN CÙNG BÀN LUẬN	6

YEAR MAY WE ALL HAVE	6	VỀ LẠI THỦ ĐÔ DC	6	GIA ĐÌNH NỮ HỘ SINH	6
# NƯỚC TRANH # GHẾ	5	BỘ GIAO THÔNG VẬN TẢI	6	HỘI NGHỊ VIẾT VĂN TRẺ	6
BÀI PHÁT BIỂU CỦA DÂN	5	{Name} VÀ CHỊ {Name} {Name}	6	HỘI VĂN HỌC NGHỆ THUẬT	6
BÀI PHÁT BIỂU CỦA ÔNG	5	DỰ ÁN ĐẦU TƯ CÔNG	6	BÀN TAY NHỎ DƯỚI MƯA	6
BẤT HẠNH VÀ THANH TẨY	5	DU HỌC SINH VIỆT NAM	6	ANH EM NHÀ TÂY SƠN	6
CHÍNH TRỊ VÀ QUÂN SỰ	5	GDP BÌNH QUÂN ĐẦU NGƯỜI	6	AN TOÀN VỆ SINH THỰC	6
CHƯƠNG TRÌNH HẬU TIẾN SĨ	5	HÀNH TRÌNH TÌM TỰ DO	6	BẬC THẦY TRUYỆN NGẮN TỐI	5
CHUYÊN ĐỀ TRUYỆN CỰC NGẮN	5	KÍNH MỜI QUÝ ĐỒNG HƯƠNG	6	AI LÊN XỨ HOA ĐÀO	5
CỘNG ĐỒNG NVTD ÚC CHÂU	5	LÀ MỘT TRONG NHỮNG NGƯỜI	6	BIÊN PHÒNG TỐNG LÊ CHÂN	5
CỦA ỦY HỘI SÔNG MEKONG	5	NÊM NẾM CHO VỪA ĂN	6	JUST THE WAY YOU ARE	5
ĐẢNG CẦM QUYỀN LAO ĐỘNG	5	NGUYỆN CHO HÒA BÌNH TRÊN	6	KỶ NIỆM # NĂM NGÀY	5
ĐƯA RAP VÀO JAZZ MÓN	5	NHÂN QUYỀN TẠI VIỆT NAM	6	CHỦ TỊCH HỒ CHÍ MINH	5
GIỚI LÃNH ĐẠO VIỆT NAM	5	THƯ CHO ÔNG GIÀ NOEL	6	CẤP THÂM NIÊN NHÀ GIÁO	5
HOÁ BẤT HẠNH VÀ THANH	5	TRONG BẢN TUYÊN NGÔN NÀY	6	ĐẠI TÁ NGUYỄN CÔNG TUẤN	5
HỘI FOOD PHOTOGRAPHY TRÊN FACEBOOK	5	ỦNG HỘ TINH THẦN CHO	6	MỘT THÀNH VIÊN IN QUÂN	5
HỐI LỘ CHO BÁC SĨ	5	VÀ MÙA XUÂN TRONG ĐÔI	6	CHO NHỮNG NGƯỜI XUNG QUANH	5
JAZZ MÓN THỜI TRANG QUANH	5	CỦA NGƯỜI MỸ GỐC VIỆT	5	CÁC HỌA SĨ VIỆT NAM	5
LỘ CHO BÁC SĨ VÀ	5	CÓ NHỮNG CHIỀU THÀNH PHỐ	5	BÌA COCKTAIL CHO TÌNH YÊU	5
LÒ ĐẾN # ĐỘ C	5	CAN'T LOOK AT THE STARS	5	CÓ THỂ BẠN MUỐN ĐỌC	5
MÀU VÀNG ÚA VÀ MÀU	5	# THÁNG ĐẦU NĂM #	5	ĐỒN BIÊN PHÒNG TỐNG LÊ	5
MÓN THỜI TRANG	5	ĂN CHUNG VỚI	5	CÁC THẾ LỰC THÙ	5

QUANH NĂM		NƯỚC MẮM		ĐỊCH	
NGÀY # THÁNG # NĂM	5	{Name} {Name} VÀ {Name} {Name}	5	HỘI CHỢ HÀNG THỦ CÔNG	5
OF THE SOUND OF LONELINESS	5	BIỂU TÌNH CẦU NGUYỆN CHO	5	KHÔNG BIẾT BAO NHIÊU LẦN	5
ONCE UPON A TIME IN	5	CÁC DỰ ÁN ĐẦU TƯ	5	LÀ LẦN ĐẦU TIÊN TÔI	5
PHÁT BIỂU CỦA DÂN BIỂU	5	CỦA CHÍNH PHỦ VIỆT NAM	5	CÔNG TY TNHH MỘT THÀNH	5
RAP VÀO JAZZ MÓN THỜI	5	CŨNG LÀ MỘT TRONG NHỮNG	5	BÂY GIỜ LÀ THÁNG MƯỜI	5
SÔI RỒI VẶN NHỎ LỬA	5	ĐÀN BÀ LÀ ĐỘNG VẬT	5	ĐÔNG BẮC KÝ SỰ #	5
SPEED OF THE SOUND OF	5	CÁC HÃNG BẢO HIỂM TƯ	5	HAI NGƯỜI KHÔNG NHÌN MẮT	5
TẨY XƯƠNG BẰNG CÁCH NGÂM	5	CHO ANH EM NHÀ {Name}	5	CHIA SẺ CÙNG CÁC BẠN	5
TỔ CHỨC ROOM TO READ	5	CHỨC SWIPES FOR THE HOMELESS	5	DOANH NGHIỆP KINH DOANH XĂNG	5
TRỨNG Ở NHIỆT ĐỘ PHÒNG	5	CỜ BẠC BẤT HỢP PHÁP	5	DỰ HỘI NGHỊ VIẾT VĂN	5
ÚA VÀ MÀU XANH LÁ	5	CON NGƯỜI LÀ ĐỘNG VẬT	5		
UPON A TIME IN CABRAMATTA	5	CỦA NGƯỜI VIỆT HẢI NGOẠI	5		
VÀNG ÚA VÀ MÀU XANH	5				
VÀO JAZZ MÓN THỜI TRANG	5				

APPENDIX V

3-CORPUS FUNCTION CLASS FORMS LISTS

List 1: 1-Tiếng Function Class Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
ANH	1,032	ANH	1,947	AI	1,106
BỊ	760	BỊ	1,425	ANH	2,854
CẢ	885	CẢ	1,471	ẤY	1,198
CÁC	1,806	CÁC	1,911	CẢ	1,648
CÁCH	783	CÁI	3,801	CÁC	2,170
CÁI	1,382	CHỈ	1,958	CÁCH	1,047
CHỈ	1,187	CHỊ	1,217	CÁI	2,574
CHO	2,799	CHO	5,902	CHỈ	2,144
CHÚNG	901	CÓ	7,957	CHO	3,942
CÓ	4,941	CÔ	1,301	CÓ	7,814
CÔ	733	CON	3,622	CÔ	1,701
CON	1,409	CÒN	2,313	CON	3,062
CÒN	1,140	CỦA	5,830	CÒN	2,077
CỦA	3,787	CŨNG	3,652	CỦA	5,766
CÙNG	595	ĐÃ	2,917	CŨNG	3,436
CŨNG	1,819	ĐẾN	1,880	ĐÃ	3,286
ĐÃ	1,696	ĐƯỢC	3,725	ĐẾN	2,505
ĐANG	573	EM	2,214	ĐƯỢC	3,802
ĐẾN	1,477	GÌ	1,677	EM	2,479
ĐƯỢC	2,360	HAY	1,418	GÌ	1,814

EM	836	HẾT	1,113	HAY	1,331
GÌ	795	HƠN	1,315	HƠN	1,457
HAY	985	KHI	2,310	KHI	2,475
HỌ	748	LÀ	8,136	LÀ	8,248
HƠN	843	LẠI	3,304	LẠI	3,108
KHÁC	700	LÊN	1,581	LÊN	1,481
KHI	1,767	LÚC	1,118	LÚC	1,129
LÀ	5,003	MÀ	3,676	MÀ	3,080
LẠI	1,490	MÌNH	4,803	MÌNH	4,056
LÊN	859	NÀO	1,829	NÀO	1,841
MÀ	1,766	NAY	1,626	NÀY	2,566
MÌNH	2,562	NÀY	3,919	NÊN	1,290
NÀO	939	NÊN	1,930	NHÀ	2,627
NÀY	1,917	NHÀ	2,716	NHƯ	3,492
NÊN	748	NHƯ	2,827	NHƯNG	2,338
NHÀ	1,096	NHƯNG	2,253	NHỮNG	4,975
NHƯ	1,758	NHỮNG	3,189	NÓ	1,813
NHƯNG	1,410	NÓ	1,722	Ở	2,864
NHỮNG	2,219	Ở	2,916	ÔNG	1,395
NÓ	930	ÔNG	1,639	PHẢI	2,375
Ở	1,805	PHẢI	2,477	QUA	1,399
ÔNG	1,014	QUA	1,806	RA	2,922
PHẢI	1,230	QUÁ	1,201	RỒI	2,186
QUA	730	RA	3,169	SAO	1,129
RA	1,592	RỒI	2,845	SAU	1,285
RỒI	916	SAO	1,152	SẼ	1,476
SAU	921	SAU	1,396	SỰ	1,554
SẼ	1,128	SẼ	1,660	TA	1,722

SỰ	1,015	TA	1,344	THẾ	1,986
TA	1,183	THẾ	1,317	THEO	1,056
THẾ	928	THÌ	4,389	THÌ	2,789
THÌ	1,857	THÔI	1,183	TÔI	5,348
TỚ	577	TÔI	4,191	TRÊN	1,676
TÔI	2,775	TỚI	1,653	TRONG	3,753
TRÊN	1,039	TRÊN	1,538	TRƯỚC	1,075
TRONG	2,390	TRONG	3,426	TỪ	1,953
TRƯỚC	577	TRƯỚC	1,224	VÀ	5,602
TỪ	1,144	TỪ	1,768	VĂN	1,205
VÀ	4,795	VÀ	6,563	VẪN	1,282
VÀO	1,617	VẪN	1,146	VÀO	2,033
VẬY	669	VÀO	2,021	VẬY	1,104
VỀ	1,434	VẬY	1,487	VỀ	2,936
VÌ	937	VỀ	2,839	VÌ	1,572
VỚI	2,269	VÌ	1,704	VIỆC	1,268
		VỚI	2,945	VỚI	3,132

List 2: 2-Tiếng Function Class Forms

AUS		**US**		**VN**	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
BAO GIỜ	140	BAO GIỜ	269	BAO GIỜ	352
CHO ĐẾN	104	BAO NHIÊU	231	BAO NHIÊU	180
CHÚNG TA	425	CHÚNG TA	225	CÁI GÌ	186
CHÚNG TÔI	169	CHÚNG TÔI	254	CHÚNG TA	218
CÓ LẼ	102	CÓ LẼ	261	CHÚNG TÔI	400
CÓ THỂ	769	CÓ THỂ	754	CÓ LẼ	261

CỦA CÁC	128	ĐÓ LÀ	256	CÓ THỂ	978
ĐẾN KHI	176	LÀM CHO	183	CỦA MỘT	203
ĐÓ LÀ	228	LÀM SAO	208	ĐÓ LÀ	442
ĐỐI VỚI	108	NÀO CŨNG	304	GÌ ĐÓ	162
MỘT CÁCH	189	NHẤT LÀ	201	MỘT CÁCH	180
NÀO CŨNG	129	NHƯ THẾ	253	NÀO CŨNG	238
NHẤT LÀ	136	NHƯ VẬY	264	NHẤT LÀ	188
NHƯ THẾ	194	SAU KHI	213	NHƯ THẾ	351
NHƯ VẬY	168	TẤT CẢ	268	NHƯ VẬY	287
SAU ĐÓ	177	TRƯỚC KHI	206	SAU KHI	171
SAU KHI	133	VẪN CÒN	233	TẤT CẢ	298
TẤT CẢ	191	VẬY MÀ	218	THẾ NÀO	284
THẾ NÀO	132			TRONG NHỮNG	170
TRƯỚC KHI	136			VỚI NHỮNG	216
TUY NHIÊN	124				
VẪN CÒN	103				
VỚI NHỮNG	115				

List 3: 3-Tiếng Function Class Forms

AUS		**US**		**VN**	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
CHO ĐẾN KHI	54	CŨNG CÓ THỂ	48	CHẲNG BAO GIỜ	28
CÓ THỂ ĐƯỢC	22	ĐẶC BIỆT LÀ	45	CHO ĐẾN KHI	36
CŨNG CÓ THỂ	36	HÔM NAY LÀ	28	CŨNG CÓ THỂ	60
ĐẶC BIỆT LÀ	26	LÀ MỘT TRONG	55	ĐẶC BIỆT LÀ	42
GẮN LIỀN VỚI	22	LÀM SAO MÀ	32	ĐỂ CÓ THỂ	28
LÀ MỘT SỰ	26	LÚC NÀO CŨNG	74	LÀ MỘT TRONG	45

LÀ MỘT TRONG	36	MỌI NGƯỜI ĐỀU	36	LÀM THẾ NÀO	35
LÚC NÀO CŨNG	44	MỘT TRONG NHỮNG	66	LÚC NÀO CŨNG	73
MỘT TRONG NHỮNG	39	NÀO CŨNG CÓ	42	MỘT TRONG NHỮNG	60
Ở ĐÂY LÀ	22	TẤT CẢ CÁC	58	NÀO CŨNG CÓ	36
QUAN TRỌNG NHẤT	26	TẤT CẢ MỌI	44	TẤT CẢ CÁC	36
TẤT CẢ CÁC	43	TẤT CẢ NHỮNG	36	TẤT CẢ NHỮNG	69
TẤT CẢ MỌI	22	THÌ CÓ THỂ	29	THẾ NÀO ĐỂ	29
TẤT CẢ NHỮNG	29	THÌ LÀM SAO	29	TRONG ĐÓ CÓ	29
TRONG ĐÓ CÓ	22	TRONG ĐÓ CÓ	30		

List 4: 4-Tiếng Function Class Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
CHÚNG TA CÓ THỂ	20	CHÚNG TA CÓ THỂ	12	BẤT CỨ LÚC NÀO	11
KHÔNG AI CÓ THỂ	10	KHÔNG AI CÓ THỂ	13	CHÚNG TA CÓ THỂ	9
LÀ MỘT TRONG NHỮNG	22	LÀ MỘT TRONG NHỮNG	41	CÓ CÁI GÌ ĐÓ	12
LÀM THẾ NÀO ĐỂ	10	MỌI NGƯỜI ĐỀU CÓ	23	CÓ THỂ XEM LÀ	9
NGƯỜI TA CÓ THỂ	17	MỘT NGÀY NÀO ĐÓ	10	HƠN BAO GIỜ HẾT	9
TẤT CẢ MỌI NGƯỜI	13	TẤT CẢ MỌI NGƯỜI	30	KHÔNG BAO GIỜ ĐƯỢC	9
		TẤT CẢ NHỮNG GÌ	10	LÀ MỘT TRONG NHỮNG	31
				LÀM THẾ NÀO ĐỂ	22
				LÚC NÀO CŨNG CÓ	10
				MỘT CÁI GÌ ĐÓ	11
				MỘT ĐIỀU GÌ ĐÓ	12
				NGƯỜI TA CÓ THỂ	11

				TẤT CẢ MỌI NGƯỜI	16
				TẤT CẢ NHỮNG ĐIỀU	10
				TẤT CẢ NHỮNG GÌ	22

List 5: 5-Tiếng Function Class Forms

AUS		US		VN	
Form	**Freq**	**Form**	**Freq**	**Form**	**Freq**
		CỦA TẤT CẢ MỌI NGƯỜI	6	KHÔNG PHẢI LÚC NÀO CŨNG	6
		LÀ MỘT TRONG NHỮNG NGƯỜI	6	VÂN VÂN VÀ VÂN VÂN	6
		TẤT CẢ MỌI NGƯỜI ĐỀU	6	VỚI TẤT CẢ MỌI NGƯỜI	6

APPENDIX X

3-CORPUS KEYNESS LISTS

List 1: 1-Tiếng Keyness List

AUS		US		VN	
Form	**Keyness**	**Form**	**Keyness**	**Form**	**Keyness**
TỚ	1,159.9	MÌNH	725.4	TÔI	1,241.8
#	1,075.4	TỚI	526.6	MỘT	643.9
BÁNH	851.9	MẸ	488.7	NHỮNG	572.9
BỘT	627.2	MAY	482.4	CÔ	470.7
ÚC	495.2	ĐI	450.2	YÊU	297.5
{Name}	487.0	HÔM	351.7	TRUYỆN	276.6
TÔI	410.4	MƯA	339.7	ẤY	274.3
TRỨNG	395.7	TÔI	304.5	MÌNH	272.7
KEM	380.9	VI	277.1	CHIẾC	244.1
VÀ	353.5	XE	273.1	M	225.1
{Name}	350.2	XONG	259.0	XE	210.5
MỘT	331.2	THÁNG	245.4	CẬU	208.4
K	329.5	CHO	244.7	ĐƯỜNG	195.6
VÀO	317.1	CHÀNG	238.2	LẶNG	182.2
ADORNO	288.3	LẠNH	237.8	QUÁN	180.4
MÌNH	264.9	TUẦN	231.9	ĐỀN	179.1
BRISBANE	258.9	MÙA	230.7	PHỐ	171.5
ĐUN	242.7	#	229.6	ĐI	165.8
HỖN	230.4	M	228.5	{Name}	165.0
DÙNG	230.0	CÁI	223.4	CUỐN	155.8
{Name}	229.3	NGÀY	223.0	NÓ	155.2
AND	212.6	LỄ	196.8	VIỆC	150.2
NGUỘI	200.7	BA	176.7	CƠN	146.3
VỊ	199.3	MẤY	174.0	NÚI	145.8
NƯỚNG	197.2	SÁNG	173.0	KHÁCH	145.6
TRỘN	193.3	CON	172.4	NẮNG	142.8
DỨA	187.9	MÁ	168.0	DUYÊN	142.1
KHI	178.5	CUỐI	165.5	CÂY	141.7

PHÚT	175.6	GIÁNG	164.8	ĐÊM	137.5
CHÚNG	174.5	TUYẾT	161.6	THỞ	136.9
BƠ	165.0	NGỦ	156.9	SÁCH	132.7
MÓN	163.4	ĐỨA	155.9	GIÁC	127.5
KHOẢNG	159.7	{Name}	154.8	PHIM	124.4
VỚI	158.9	CHỒNG	154.3	BÓNG	122.3
THÌA	152.8	CHƠI	152.6	ÁNH	122.2
NGỌT	151.7	CHÚT	152.3	SUỐI	121.8
VANILLA	148.4	NĂM	150.9	BỜ	111.8
M	147.1	RỒI	145.9	CHIỀU	111.1
MT	144.0	ĐƯỜNG	141.8	ĐÃ	110.7
XOÀI	143.5	XUỐNG	138.5	NỘI	109.6
SYRUP	140.3	OBAMA	138.2	TỈNH	109.5
VỎ	137.6	NGỒI	137.9	NƠI	107.8
CUP	136.6	NÀY	136.1	ĐẾN	106.0
HỢP	133.6	NAY	135.6	THOẠI	103.7
MUỐI	133.5	ĂN	134.9	BÌA	102.6
CHOCOLATE	130.9	ĐÈN	134.0	TRÔI	100.7
THE	130.3	LÀM	133.8	NGÀY	100.7
KHÔ	129.6	TỐI	132.2	TÌNH	100.0
WATSON	129.4	NOEL	129.9	VỀ	99.2
LÒ	128.7	LÁI	129.0	MIỄU	98.8
LỰU	126.2	BÊN	128.6	THỂ	98.7
HABIBI	125.9	TRỜI	127.5	CHỢT	96.9
MACARONS	124.8	VỪA	125.3	TỐI	95.9
ĐĨA	123.0	LẠI	123.5	BUỔI	95.3
CHANH	121.6	TỎI	121.7	YÊN	92.6
BOSWELL	121.1	VÒNG	121.0	MẮT	92.0
THỨC	120.8	ĐỒ	118.6	SAIGON	91.6
PHẦN	116.9	PHÒNG	116.3	VẬT	91.1
KHUÔN	116.2	SHOW	115.8	SỢI	90.4
FOREX	114.9	NHỎ	115.7	XANH	89.2
BẠN	113.4	CÔ	115.6	THẦY	89.1
DỪA	111.3	QUÀ	113.1	MƯA	89.1
OF	109.1	MÀU	112.2	KANU	88.6
KHAY	108.2	CHIỀU	110.0	CUỘC	88.5
THỂ	105.2	NÀNG	109.6	KHIẾN	88.3
TRẦM	104.0	MẠ	109.0	GIÓ	88.3
JAZZ	103.8	CHỤP	107.0	NGỒI	87.0
SÔI	102.9	MIẾNG	106.8	DIỀU	86.0

SỮA	102.1	DC	104.9	LINH	85.1
LỚP	101.0	HÃNG	102.4	HẢO	85.0
TREVOR	100.7	CHẠY	101.6	ĐÁ	85.0
TRONG	99.4	QUA	99.9	CƯỜI	84.6
MỊN	99.3	PHÚT	99.4	BÁT	83.3
TNS	98.6	CHRISTMAS	99.0	ĐÔI	82.6
RÂY	98.4	MUSHU	98.3	TRẦN	80.6
LOẠI	97.8	NÓ	95.6	KHANG	80.4
HOLMES	97.1	LÁ	93.1	KHU	79.6
QUEENSLAND	96.0	VÀI	92.5	HÀ	79.5
TSP	94.5	LẦN	92.4	SÔNG	78.7
BẾP	93.3	TRƯA	92.2	NGÃ	78.7
PANNA	93.2	XMAS	90.9	NGẮN	76.3
GIÁ	92.7	THẰNG	89.6	SÀI	76.2
COTTA	91.1	TIỆM	89.3	GÒN	75.3
MAPLE	90.1	BAY	88.4	TƯỢNG	75.1
MÌ	88.9	SOE	86.2	MẶT	75.1
GR	88.8	CỬA	85.8	KHÁ	74.4
KHUẤY	86.9	CÂY	84.9	CHÂN	74.0
CÀ	86.7	ÁO	84.8	CHUYẾN	73.6
XAYABURI	84.7	VIRGINIA	84.5	BẾN	73.4
TÚ	84.5	NỆM	83.9	BÃI	73.1
FREUD	83.9	RA	83.8	TRỞ	72.9
TBS	83.8	CHIÊN	83.6	AYUTTHAYA	72.7
BÔNG	82.7	NẤM	81.4	ĐẦU	72.0
KHỎANG	81.4	VÔ	78.7	VỞ	71.9
MẠCH	79.9	GỎI	78.3	PHÍA	71.8
PIZZA	79.5	HAI	77.3	LẠI	71.7
SAU	78.5	CHIẾC	77.2	XA	71.6
NATALIE	78.3	MUA	76.8	THIỀN	71.3
SẼ	77.9	NHÀ	76.5	NGỦ	71.2
OG	77.7	ĐÊM	76.5	TẾT	70.6

List 2: 2-Tiếng Keyness List

AUS		**US**		**VN**	
Form	**Keyness**	**Form**	**Keyness**	**Form**	**Keyness**
CÔNG THỨC	365.6	HÔM NAY	175.3	HÀ NỘI	225.3
HỖN HỢP	319.4	BẢO HIỂM	166.9	TIỂU THUYẾT	212.6
NƯỚC DÙNG	225.2	NĂM NAY	166.3	TÌNH YÊU	188.6

ĐẾN KHI	219.4	MỸ GỐC	141.3	TRUYỆN NGẮN	156.2
# PHÚT	217.0	TƯỜNG VI	140.0	CHÚNG TÔI	154.0
CHÚNG TA	215.1	ĐI LÀM	134.2	CẢM GIÁC	131.2
KHOẢNG #	189.7	VI #	129.8	NHÀ VĂN	129.9
CHIẾC BÁNH	164.0	GIÁNG SINH	126.9	CÔNG VIỆC	123.2
# ĐƯỜNG	155.5	LÁI XE	125.6	AI ĐÓ	114.7
NGUYÊN LIỆU	153.1	DOANH NGHIỆP	121.8	CÁI BÁT	112.7
{Name} {Name}	142.2	HÔM QUA	115.2	PHỤ CẤP	105.8
MÓN ĂN	142.0	THÌ MÌNH	114.9	CUỘC SỐNG	105.5
YẾN MẠCH	139.7	# GIỜ	113.3	TÔI ĐÃ	99.9
ĐỂ NGUỘI	134.1	KIM CƯƠNG	105.0	NHÂN VẬT	97.2
CÓ THỂ	132.8	CHỤP HÌNH	103.5	NHẬN RA	92.7
# THÌA	126.7	NGƯỜI MỸ	92.6	ĐIỀU GÌ	92.1
K CÓ	116.7	HAI ĐỨA	92.1	NHÀ GIÁO	89.9
MÌNH K	114.7	SÁNG NAY	91.1	DU LỊCH	85.1
MÌNH NHỚ	114.7	NƯỚC BROTH	82.8	NHỮNG NGÀY	84.0
# QUẢ	107.6	ĐỒ ĂN	81.9	BỘ PHIM	83.4
CÀ CHUA	105.9	VỀ LẠI	79.8	SÀI GÒN	79.4
ĐẦU CƠ	105.8	CHỪNG #	79.1	VÉ SỐ	78.9
THỊ TRƯỜNG	105.6	CÂY THÔNG	76.5	TRÁI TIM	78.8
TRẦM TƯỞNG	104.3	CẢM CÚM	76.0	MỘT NGƯỜI	78.3
# CUP	103.4	RẤT LÀ	75.1	DU KHÁCH	77.5
CUNG TRẦM	103.1	# PHÚT	74.6	CUỐN SÁCH	75.4
CỔ PHIẾU	102.9	Ở NHÀ	73.4	ĐIỆN THOẠI	73.1
BÁNH QUY	101.0	DON HỒ	72.5	CẢM XÚC	72.4
Ở ÚC	96.3	LIỆU AMP	69.9	# THĂM	71.7
NHÀ LÃNH	91.3	ĐỒNG HỒ	69.4	VĂN MIẾU	71.5
PANNA COTTA	91.2	THẮNG PP	68.9	CON ĐƯỜNG	69.3
DÂU TÂY	90.2	II CÁCH	67.7	TRUYỆN TRANH	69.3
# GR	88.8	ĐẠI TỶ	67.5	R CARVER	68.3
# BỘT	88.6	MÌNH KHÔNG	65.9	DỄ CHỊU	66.6
KHỎANG #	86.1	ĐIỀU #	65.5	TRƯNG BÀY	66.2
VỊ NGỌT	85.7	DAY #	64.9	BẮT ĐẦU	65.5
HỖN HỢP	84.5	SAN FRANCISCO	63.2	TÁC PHẨM	63.4
NƯỚC CỐT	84.0	VỀ NHÀ	63.0	ĐAM MÊ	62.7
MAPLE SYRUP	83.8	CÁC DOANH	62.8	Ý ĐIÊN	61.8
SAU ĐÓ	83.6	CON MAY	62.4	CẤP THÂM	61.7
CÁCH LÀM	83.3	NƯỚC MẮM	61.4	NHƯ MỘT	61.1
PHÔ MAI	83.0	CỦA VI	61.0	BỐ TÔI	60.5
# TSP	82.6	VỢ CHỒNG	60.0	NHỮNG CƠN	60.4

HUN KHÓI	81.4	TUỲ Ý	59.9	MÌNH ĐÃ	60.4
CHỒNG DÍNH	81.0	CUỐI THÁNG	58.5	LAO RÙA	59.5
HOA CHUỐI	80.9	VIETNAM AIRLINES	58.2	VỀ PHÍA	59.2
CHẾ BIẾN	79.9	MÙA THU	57.7	GƯƠNG MẶT	58.4
MẬT ONG	79.2	RỬA SẠCH	57.6	XE ÔM	58.3
ÁNH SÁNG	79.1	THỨ SÁU	57.1	VŨNG TÀU	57.9
BÁNH ĐÚC	77.7	HOA TÌNH	56.8	HÀNG QUÀ	57.4
# TBS	77.6	CHO CHÚT	56.3	BẢO TÀNG	57.1
TỚ THÍCH	76.9	LÀ MÌNH	56.0	ẤU TẨU	57.0
BỘT MÌ	76.1	CÔ NHỎ	55.8	NHỮNG CUỐN	56.0
NỀN CÔNG	75.9	CHÚT DẦU	55.4	XE MÁY	55.9
SỮA TƯƠI	75.8	TẤM HÌNH	54.6	DI TÍCH	55.4
VÀO HỖN	75.2	MÌNH CÓ	54.5	CHUYẾN ĐI	54.3
NHÂN QUYỀN	74.8	THỨ BẢY	54.5	TÔI THÍCH	53.5
TUỔI THỌ	73.9	# TUẦN	54.5	KHOẢNH KHẮC	53.1
BẠN CÓ	73.8	# MUỐNG	54.3	TIN NHẮN	53.1
QUẢ TRỨNG	73.6	BA MẸ	53.7	CỦA TÔI	52.2
HÀNH TÂY	72.4	CHO MẸ	53.1	NHỮNG THỨ	51.9
BỮA SÁNG	71.0	GIA ĐÌNH	51.5	PHÚ THỌ	50.9
THÌA CÀ	70.2	CA SĨ	51.5	MỘT QUÁN	50.8
CỦA ADORNO	70.2	NGÀY MAI	51.3	TỪ LÁY	50.7
DÂN BIỂU	69.7	NGÀY CUỐI	50.2	THÀNH PHỐ	50.2
ĐỒNG ÚC	69.0	XONG THÌ	50.2	NGÃ TƯ	50.1
VÀO TỦ	68.4	BỘT NÊM	49.8	CÙ LAO	50.0
TỦ LẠNH	68.1	LÚN GHẺ	49.2	TÔI KHÔNG	49.6
VÀO NỒI	68.1	CHÉP VỤN	49.1	CÔ GÁI	49.5
THÁI LÁT	67.1	CHO CON	48.8	ĐỒNG NAI	49.3
# BƠ	65.9	TRỜI MƯA	48.8	TÀ LÀI	48.6
CÀ RỐT	65.6	LỰA CHỌN	48.5	TRỞ VỀ	48.3
LEMON CURD	65.2	ĐI BỘ	48.4	THU TRANG	47.6
# NƯỚC	64.8	MỘT NGÀY	48.2	THÂM NIÊN	47.6
ĐUN SÔI	64.5	GỐC VIỆT	48.2	NGƯỜI ẤY	46.9
CÁ HỒI	64.2	BÙI GIÁNG	48.1	MẸ TÍ	46.9
KHUẤY ĐỀU	64.0	MÙA ĐÔNG	48.0	{Last} {Middle}	46.6
CUỘC ĐỐI	63.7	XE FOOD	47.9	CUỘC ĐỜI	46.5
VIỆT LUẬN	63.7	HÀNH LÁ	47.9	ANH CSGT	46.4
TNS BOSWELL	63.5	CHA CON	47.5	CAN LỘC	46.4
BỘT NỞ	63.5	CÁI ÁO	47.4	VỚI TÔI	45.9
MUỐI #	62.8	THẰNG CON	47.3	HÀNH TRÌNH	45.9
KEM TƯƠI	62.4	ĂN TRƯA	46.8	NHÌN THẤY	45.6

CHO VÀO	62.3	LÀM VỀ	46.8	CƠN MƯA	45.4
LÝ TÍNH	61.6	CHO TỚI	46.7	VƯỜN THƠ	45.2
LỚP #	61.6	CHUẨN BỊ	46.5	ĐẾN MỨC	45.2
ĐẦU TƯ	61.4	CHỊ TA	46.3	LỘ #	45.1
CŨNG K	60.6	ĐẦU TIÊN	46.3	NGÔI NHÀ	45.0
CHỤP ẢNH	60.5	NƯỚC MỸ	46.2	HOÀN THÀNH	44.9
NƯỚC CHANH	60.5	TUYỆT TÌNH	46.2	NGƯỜI BẠN	44.3
HỢP BỘT	60.5	KÉO NHAU	46.1	ĐỘC THÂN	44.2
BÁNH MÌ	60.0	CÔ GIÁO	45.9	TRÊN ĐƯỜNG	44.2
GIÁ CẢ	59.9	VỀ TỚI	45.9	MỈM CƯỜI	44.2
BẠN {Name}	59.0	CÁC SOE	45.9	THUYẾT CHIẾN	44.2
LÁ XOĂN	59.0	I NGUYÊN	45.9	MỘT MÌNH	44.1
VÀO KHUÔN	58.0	NHÀ TÔI	45.6	TRỞ NÊN	44.1
BƯỚC #	57.8	VÒNG VÒNG	45.6	CẢM HỨNG	43.9
ĐÁNH TRỨNG	57.4	HÀNG KHÔNG	45.5	ĐỂ HÔN	43.8
NGHIỆP VĂN	57.2	QUẢNG CÁO	45.4	CHUNG CƯ	43.4
BẠN SẼ	56.6	TỐI QUA	45.3	CÀ PHÊ	43.3

List 3: 3-Tiếng Keyness List

AUS		US		VN	
Form	**KN**	**Form**	**KN**	**Form**	**KN**
CUNG TRẦM TƯỞNG	104.3	TƯỜNG VI #	143.0	CẤP THÂM NIÊN	61.7
BẠN CÓ THỂ	99.3	NGƯỜI MỸ GỐC	111.3	PHỤ CẤP THÂM	61.7
NHÀ LÃNH ĐẠO	92.7	MỸ GỐC VIỆT	94.5	CÙ LAO RÙA	59.5
VÀO HỖN HỢP	75.2	CÁC DOANH NGHIỆP	67.5	GIỜ LÀ THÁNG	49.6
KHOẢNG # PHÚT	74.1	II CÁCH LÀM	65.5	TRẦN THU TRANG	47.9
NGHIỆP VĂN HÓA	74.0	CHỪNG # PHÚT	55.7	HÔN EM LẦN	44.4
NỀN CÔNG NGHIỆP	74.0	NGUYÊN LIỆU AMP	50.2	ĐỂ HÔN EM	44.4
CÔNG NGHIỆP VĂN	74.0	GHI CHÉP VỤN	49.1	THUYẾT CHIẾN TRANH	44.2
VỀ NHÂN QUYỀN	73.6	ĐI LÀM VỀ	47.6	TIỂU THUYẾT CHIẾN	44.2
THÌA CÀ PHÊ	70.2	I NGUYÊN LIỆU	45.9	ĐIỀU GÌ ĐÓ	43.8
# QUẢ TRỨNG	67.6	CON CẢM CÚM	45.8	EM LẦN NỮA	43.5
CUỘC ĐỐI THOẠI	65.4	BIỂN ĐÔNG NAM	45.8	MÃ PÍ LÈNG	43.0
NHÂN QUYỀN Ở	65.3	HIỂM SỨC KHỎE	43.7	TAM GIÁC VÀNG	43.0
CÁC CUỘC ĐỐI	61.4	ĐẠI TỶ TUI	43.7	VĂN MIẾU TRẤN	41.9
CẢI LÁ XOĂN	59.0	BẢO HIỂM SỨC	43.7	MIẾU TRẤN BIÊN	41.9
HỖN HỢP BỘT	54.3	XE FOOD TRUCK	43.6	ĐỖ HỒNG NGỌC	40.1
# THÌA CÀ	51.5	TUYỆT TÌNH CỐC	42.8	NGƯỠNG THỜ MẪU	39.7

NHAN NHƯ NGỌC	51.2	TỜ NGƯỜI VIỆT	42.5	VƯỜN THƠ VIỆT	39.7
# ĐỘ C	51.2	MUA BẢO HIỂM	41.0	TÍN NGƯỠNG THỜ	39.7
# THÌA CANH	49.7	TRẠI TỊ NẠN	40.5	LÀM CHUYỆN ẤY	39.7
CHO ĐẾN KHI	49.6	MUỖNG CÀ PHÊ	39.7	TÔI NHẬN RA	37.9
THOẠI VỀ NHÂN	49.2	GAI HOA TÌNH	39.3	HÀNG QUÀ RONG	35.3
CÔNG THỨC NÀY	48.6	ĐỀU CÓ QUYỀN	39.3	BẢO TÀNG PHỤ	35.3
ĐỐI THOẠI VỀ	48.1	THỦ ĐÔ DC	38.0	TÀNG PHỤ NỮ	35.3
VÀO TỦ LẠNH	47.3	MÙA GIÁNG SINH	37.8	J L BORGES	35.3
# ĐƯỜNG CÁT	46.6	CÁC NGÂN HÀNG	36.7	CUỐN TIỂU THUYẾT	33.9
HỖN HỢP TRỨNG	46.6	# MUỖNG CÀ	36.4	LÀ THÁNG #	33.4
ĐẢNG LAO ĐỘNG	45.6	QUẢNG CÁO CỦA	36.3	NGƯỜI NHÀ QUÊ	33.2
HÓA BÌNH DÂN	45.3	TỔNG THỐNG OBAMA	36.1	THẲNG CHÂN RƯỠI	33.1
NHIỆT ĐỘ PHÒNG	45.3	CHO CHÚT DẦU	36.0	QUỐC LỘ #	31.9
VĂN HÓA BÌNH	45.3	EM CHÚNG TÔI	35.3	NHỮNG CHI TIẾT	31.9
ĐA VĂN HÓA	45.0	CỦA ÔNG BÀ	35.2	NHỮNG CƠN MƯA	31.3
CỦ HÀNH TÂY	44.2	TÒA BẠCH ỐC	35.0	CẮT CHO TA	30.9
QUYỀN Ở VIỆT	43.5	TIẾNG ĐỒNG HỒ	34.2	CHẾ ĐỘ PHỤ	30.9
# RESULTS #	43.5	TỊ NẠN BATAAN	33.8	BỘ PHẬN KHÔNG	30.5
BEN TRẦN #	43.5	NAM Á CHÂU	32.8	PHẬN KHÔNG NHỎ	30.5
NHÀ QUẢN LÝ	42.9	NỮ CA SĨ	32.6	HỘI NHÀ VĂN	30.5
{Name} {Name} BRISBANE	42.7	CÔ GIÁO TRẺ	32.5	BÁN VÉ SỐ	30.4
Ở NHIỆT ĐỘ	42.4	THÁNG MƯỜI HAI	31.9	VỀ CHIẾN TRANH	30.3
BLOG {Name} {Name}	41.5	LẦN ĐẦU TIÊN	31.8	VIẾT VỀ CHIẾN	29.8
# BỘT MÌ	41.1	NGƯỜI ĐỀU CÓ	31.7	ĐỘ PHỤ CẤP	29.8
NGHIÊN CỨU SINH	40.7	CUỘC DIỄN HÀNH	31.6	CHIẾC ĐIỆN THOẠI	29.7
CHÁO YẾN MẠCH	40.4	CHÀNG HỌA SĨ	31.6	TRÁI TIM MÌNH	28.3
THỊ TRƯỜNG FOREX	40.4	VÀI TẤM HÌNH	31.0	{Last} {Middle} {First}	27.6
# {Name} {Name}	40.4	CAO PHÁP VIỆN	30.9	{Last} {Middle} {First}	27.6
ĐỒNG NGƯỜI VIỆT	40.4	TỐI CAO PHÁP	30.9	ĐI ĂN ỐC	27.5
ĐÀO TẠO TIẾN	39.0	III GHI CHÚ	30.6	TẬP TRUYỆN NGẮN	27.1
TẠO TIẾN SĨ	39.0	VỤ NỔ SÚNG	30.6	NHỮNG CÂU CHUYỆN	27.1
CÓ THỂ THAY	37.3	VÀ BẠN TÔI	30.5	BẠN CÓ THỂ	27.1
ĐỘ DÀI TALOMERE	37.3	BẢN TUYÊN NGÔN	30.1	KHU DI TÍCH	27.0
BÁNH CÀ RỐT	37.3	ĐẦU TƯ CÔNG	30.1	NGƯỜI ĐÀN ÔNG	26.6
DỊCH CỦA TÚ	37.3	ANH EM CHÚNG	29.9	HOA HƯỚNG DƯƠNG	26.6

JUAN RAMÓN JIMÉNEZ	37.3	VỀ TỚI NHÀ	29.2	TRẤN TORTILLA FLAT	26.5
CỦA {Name} {Name}	36.0	LỰA CHỌN THAY	28.4	THỊ TRẤN TORTILLA	26.5
HỮU NỮ NHAN	36.0	ĐỒNG TIỀN VÀNG	28.4	CHÈ BỘT LỌC	26.5
NỮ NHAN NHƯ	36.0	CHỌN THAY THẾ	28.4	ÚT NHÁI BẦU	26.5
ALL THE WAY	36.0	CHO VỪA ĂN	27.3	NHỮNG NGÀY NÀY	26.3
KHI HỖN HỢP	36.0	BẢO HIỂM Y	27.3	THIẾT KẾ BÌA	25.4
PHONG LỮ THẢO	36.0	HIỂM Y TẾ	27.3	TỈNH PHÚ THỌ	25.3
TÌNH TRẠNG NHÂN	35.8	BẾN PHÀ NÀY	27.3	TIN NHẮN CỦA	25.3
LÒNG TRẮNG TRỨNG	35.5	CÁC PHÁT MINH	27.3	MỘT AI ĐÓ	25.3
MÁY ĐÁNH TRỨNG	35.1	DOANH NGHIỆP NHÀ	27.2	PHỤ NỮ VIỆT	24.7
BẤM VÀO LINK	35.1	LỄ TẠ ƠN	26.9	NGHỆ QUÂN ĐỘI	24.3
BÀ XÃ TÔI	35.1	MÌNH KHÔNG CÓ	26.7	VĂN NGHỆ QUÂN	24.3
BẰNG CHỮ NÔM	34.9	GIA ĐÌNH MÌNH	26.5	TÔI BẮT ĐẦU	24.1
ĐẾN KHI HỖN	34.9	# TIẾNG ĐỒNG	26.4		
CỦA THỊ TRƯỜNG	34.3	NATIONAL GALLERY OF	26.3		
MÓN ĂN NÀY	34.2	VOICE OF VIETNAMESE	26.2		
NHỮNG CHIẾC BÁNH	34.2	CHÚ CÁ CON	26.2		
MÌNH VỚI MH	34.2	NÓI NGƯỜI MỸ	26.2		
HOA TỬ ĐẰNG	34.2	OF VIETNAMESE AMERICANS	26.2		
CỦA NỀN CÔNG	34.2	TÔI VÀ BẠN	26.2		
NGUYÊN LIỆU #	34.2	MẤY NGÀY NAY	26.1		
{Last} {Middle} {First}	34.2	MỘT TẤM HÌNH	25.6		
Ở VIỆT NAM	34.1	NGHIỆP NHÀ NƯỚC	25.4		
NƯỚNG KHOẢNG #	33.9	CHIẾC LÁ CUỐI	25.4		
TRẠNG NHÂN QUYỀN	33.8	CẦU NGUYỆN CHO	25.3		
BÀI PHÁT BIỂU	33.4	CÁC TIỂU BANG	25.3		
ÔNG TƯ NGHIÊM	32.9	BÔNG ĐIÊN ĐIỂN	25.1		
GIẤY CHỐNG DÍNH	32.9	TIẾNG NÓI NGƯỜI	25.1		
HỒI HUN KHÓI	32.9	HAI CHA CON	24.8		
LĂN VÀO BẾP	32.1	VÔ GIA CƯ	24.7		
CHIẾN BINH ÚC	32.1	GALLERY OF ART	24.2		
LÚC ĐÓ MÌNH	32.0	MIỀN NAM CALIFORNIA	24.1		
NHỮNG VÌ SAO	31.9	TU TU TU	24.0		
THỂ THAY BẰNG	31.8	DÙNG NÓNG VỚI	24.0		
TRỨNG GÀ #	31.8	PHI VÀ VIỆT	24.0		
MÌNH K CÓ	31.8	NGƯỜI CÔ GIÁO	24.0		
SIDE OF THE	31.8	MỤC SƯ KING	24.0		
VÀO LÒ NƯỚNG	31.1	MẮM CHANH TỎI	24.0		

MỘT CHÚT MUỐI	31.1	CHÍNH QUYỀN OBAMA	24.0		
CÁC CỔ PHIẾU	31.1	CHANH TỎI ỚT	24.0		
VỚI # NƯỚC	30.8				
CHO VÀO TỦ	30.7				
QUẢ TRỨNG GÀ	29.8				
# # B	29.8				
CÁ HỒI HUN	29.8				
LÀ CÔNG THỨC	29.8				
TRƯỜNG CHỨNG KHOÁN	29.8				
HÓA ĐẠI CHÚNG	29.7				

List 4: 4-Tiếng Keyness List

AUS		US		VN	
Form	**KN**	**Form**	**KN**	**Form**	**KN**
NỀN CÔNG NGHIỆP VĂN	76.4	NGƯỜI MỸ GỐC VIỆT	71.5	PHỤ CẤP THÂM NIÊN	61.7
CÔNG NGHIỆP VĂN HÓA	74.0	I NGUYÊN LIỆU AMP	45.9	BÂY GIỜ LÀ THÁNG	49.6
CÁC CUỘC ĐỐI THOẠI	61.4	BIỂN ĐÔNG NAM Á	45.8	ĐỂ HÔN EM LẦN	44.4
# THÌA CÀ PHÊ	51.5	BẢO HIỂM SỨC KHỎE	43.7	HÔN EM LẦN NỮA	44.4
ĐỐI THOẠI VỀ NHÂN	50.4	NGƯỜI ĐỀU CÓ QUYỀN	42.5	TIỂU THUYẾT CHIẾN TRANH	44.2
THOẠI VỀ NHÂN QUYỀN	49.2	MỌI NGƯỜI ĐỀU CÓ	40.1	VĂN MIẾU TRẤN BIÊN	41.9
NHÂN QUYỀN Ở VIỆT	46.7	# MUỖNG CÀ PHÊ	36.4	TÍN NGƯỠNG THỜ MẪU	39.7
VĂN HÓA BÌNH DÂN	45.3	TRẠI TỊ NẠN BATAAN	33.8	GIỜ LÀ THÁNG #	38.6
QUYỀN Ở VIỆT NAM	43.5	ĐÔNG NAM Á CHÂU	32.8	BẢO TÀNG PHỤ NỮ	35.3
CỘNG ĐỒNG NGƯỜI VIỆT	40.4	ANH EM CHÚNG TÔI	32.7	TÀNG PHỤ NỮ VIỆT	33.1
ĐÀO TẠO TIẾN SĨ	39.0	TÔI VÀ BẠN TÔI	31.6	BỘ PHẬN KHÔNG NHỎ	30.5
CUỘC ĐỐI THOẠI VỀ	38.0	TỐI CAO PHÁP VIỆN	30.9	VIẾT VỀ CHIẾN TRANH	29.8
DỊCH CỦA TÚ TRINH	37.3	LỰA CHỌN THAY THẾ	28.4	CHẾ ĐỘ PHỤ CẤP	29.8
BẢN DỊCH CỦA {Name}	37.3	BẢO HIỂM Y TẾ	27.3	PHỤ NỮ VIỆT NAM	27.7
# CỦ HÀNH TÂY	36.9	ANH CÚN VÀ CHUỘT	26.3	THỊ TRẤN TORTILLA FLAT	26.5
ĐẾN KHI HỖN HỢP	36.0	NÓI NGƯỜI MỸ GỐC	26.2	VĂN NGHỆ QUÂN ĐỘI	24.3
NỮ NHAN NHƯ NGỌC	36.0	TIẾNG NÓI NGƯỜI	26.2		

		MỸ			
HỮU NỮ NHAN NHƯ	36.0	VOICE OF VIETNAMESE AMERICANS	26.2		
TRẠNG NHÂN QUYỀN Ở	34.9	DOANH NGHIỆP NHÀ NƯỚC	25.4		
CỦA NỀN CÔNG NGHIỆP	34.2	# TIẾNG ĐỒNG HỒ	25.2		
TÌNH TRẠNG NHÂN QUYỀN	33.8	NATIONAL GALLERY OF ART	24.2		
CÓ THỂ THAY BẰNG	31.8	NGƯỜI CỔ GIÁO TRẺ	24.0		
BẠN CÓ THỂ THAY	31.1				
CỰU CHIẾN BINH ÚC	30.0				
CÁ HỒI HUN KHÓI	29.8				
Ở NHIỆT ĐỘ PHÒNG	29.8				
# QUẢ TRỨNG GÀ	29.8				
THỊ TRƯỜNG CHỨNG KHOÁN	29.8				
VĂN HÓA ĐẠI CHÚNG	29.7				
WRONG SIDE OF THE	27.9				
# BỘT MÌ #	27.9				
SIDE OF THE WORLD	27.9				
THE WRONG SIDE OF	27.9				
ẤN PHẨM KHOA HỌC	26.8				
BẠN CÓ THỂ DÙNG	26.7				
DÂN BIỂU CHRIS HAYES	26.7				
HABIBI YA NOUR EL	26.7				
CÁC MẠNG LƯỚI TRUYỀN	25.7				
MỘT NHÀ LÃNH ĐẠO	25.4				
XEM PHẦN # Ở	24.8				
VÀO LINK SAU ĐÂY	24.8				
ỦY BAN ĐỀ NGHỊ	24.8				
BÁC SĨ VÀ Y	24.8				
ÁNH SÁNG NHÂN TẠO	24.8				
# TSP MUỐI #	24.8				
SĨ VÀ Y TÁ	24.8				
# ĐƯỜNG CÁT #	24.8				
# Ở ĐÂY NHÉ	24.8				
CÁC CHÍNH TRỊ GIA	24.8				
CHO VÀO TỦ LẠNH	24.6				
LƯỚI TRUYỀN THÔNG XÃ	24.0				

List 5: 5-Tiếng Keyness List

AUS		US		VN	
Form	**KN**	**Form**	**KN**	**Form**	**KN**
NỀN CÔNG NGHIỆP VĂN HÓA	76.4	MỌI NGƯỜI ĐỀU CÓ QUYỀN	38.1	ĐỂ HÔN EM LẦN NỮA	44.4
ĐỐI THOẠI VỀ NHÂN QUYỀN	50.4	ANH {Name} VÀ {Name} {Name}	26.3	BÂY GIỜ LÀ THÁNG #	38.6
NHÂN QUYỀN Ở VIỆT NAM	46.7	TIẾNG NÓI NGƯỜI MỸ GỐC	26.2	TÀNG PHỤ NỮ VIỆT NAM	33.1
CUỘC ĐỐI THOẠI VỀ NHÂN	38.0	NÓI NGƯỜI MỸ GỐC VIỆT	26.2	BẢO TÀNG PHỤ NỮ VIỆT	33.1
BẢN DỊCH CỦA {Name} {Name}	37.3	BIỂN ĐÔNG NAM Á CHÂU	26.2		
HỮU NỮ NHAN NHƯ NGỌC	36.0				
TÌNH TRẠNG NHÂN QUYỀN Ở	34.9				
CỦA NỀN CÔNG NGHIỆP VĂN	34.2				
THE WRONG SIDE OF THE	27.9				
WRONG SIDE OF THE WORLD	27.9				
CÁC MẠNG LƯỚI TRUYỀN THÔNG	25.7				
BÁC SĨ VÀ Y TÁ	24.8				
PHẦN # Ở ĐÂY NHÉ	24.8				
LƯỚI TRUYỀN THÔNG XÃ HỘI	24.0				
MẠNG LƯỚI TRUYỀN THÔNG XÃ	24.0				