Disregarding linguistics: A critical study of Google Translate's syntactic/semantic errors in rendering multiword units in English to Persian translations

by

Parnian Shafia

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Arts

in

Applied Linguistics and Discourse Studies

Carleton University
Ottawa, Ontario

Abstract

Translations rendered by machine translation have been found to be error-prone when translating medical communication language. Issues are especially frequent when Persian is either the source or target language. The source of errors has been attributed to machine translation's utilization of word-for-word translation and disregard for the formulaic nature of language. The aim of this research was, then, to investigate how multiword units (MWUs) of formulaic language (FL) are processed by a commonly used machine translator, Google Translate (GT). To do this, a medical-specific corpus was created to identify non-transparent MWUs. Adopting the framework by Simpson-Vlach and Ellis' (2010) *n*-gram criteria for MWUs, 20 frequently occurring MWUs were identified in the corpus. A comparison of GT's results with manual translations suggests that GT did not take FL into consideration in 50% of the data. These findings have implications for improving the accuracy of machine translation algorithms and reducing processing time.

Key Terms: Corpus Linguistics, Formulaic Language, Machine Translation

Acknowledgments

First and foremost, I would like to thank Eva Kartchava, my family, and friends (namely, Fariba Keihani and Alexandra Ross) for their advice preventing me from dropping out of the program when I could not readily see the overlap of my topic of interest at the time (i.e., examining machine translation from a theoretical linguistics perspective) with the Applied Linguistics and Discourse Studies program. I would also like to thank David Wood for his inspiring talk in 5001 leading to insights to early ideas of incorporating theory of formulaic language in my topic of interest. I would like to also thank Guillaume Gentil for his helpful comments on my fourth assignment, a mock thesis proposal, in ALDS 5002, that I used to finalize the proposal that started this thesis, as well as sharing one of his articles on a similar topic.

Upon the start of the newly discovered route to tackling machine translation from an applied linguistics perspective, my greatest appreciation goes to Michael Rodgers for accepting to supervise me for a directed reading course as well as continuing the project resulting in the current thesis paper. During the past 21 months, the core of the project has undergone both major and minor revisits. I am truly grateful for the extensive time Michael dedicated to all my drafts, as well as the constructive guidance and MA-level writing lessons that he provided during our meetings. Furthermore, I would like to thank his encouragement to presenting at Carleton University (*Theoretical Linguistics Group*, and *Talks and Tipples*) and conferences (*Association canadienne de linquistique appliquée–Canadian Association of Applied Linguistics* [ACLA–CAAL, The Congress], and *45th International Systemic Functional Linguistics*) where I received useful feedback on my topic. I am appreciative of his constant availability for providing feedback on the

preparation of abstract as well as the actual presentation slides/poster. Furthermore, I would like to thank him for granting me the opportunity to present with him in ALDS 5001 (Fall 2017), regarding how I got to where I was then, working on the thesis sharing with a future potential thesis-writers that there is light at the end of the tunnel if one keeps their eyes vigilant. It has truly been an honour working on the topic under his supervision.

Another scholar who has always been supportive of the project I would like to thank is Daniel Siddiqi. I appreciate his always-constructive criticisms towards the topic from a theoretical linguistics standpoint. Dan's points have played an important role in shaping some of the directions to be taken in future research on the topic. Furthermore, I am grateful and lucky to have had the opportunity to be his teaching assistant for three semesters in Morphology I (twice) and II where I enjoyed his teachings in lectures and stayed connected to some of the theoretical linguistics concepts that would be helpful in the future paths of my research in the study of machine translation.

My appreciation further goes to other faculty members Ida Toivonen, Kumiko Murasugi, and Raj Singh for sharing their perspectives on the theoretical linguistics aspects, specifically examining the structural composition of the formulaic sequences in the pilot study carried out in the directed reading paper which have been influential in some subsections of the thesis project.

I would like to thank my Father, Hamed Shafia, a former linguistics lecturer and the official translator of the thesis project, for his help with the correct Persian translations of the multiword units found in my corpora, as well as his useful preliminary lessons on programming and how databases work and interact with program codes.

I would also like to express my appreciation for the time Christina Dore took to edit my thesis in a timely manner with dedicated precision.

I would also like to thank my former classmate, now friend, Alexandra Ross, for her motivation and support not only to stay in the program, but also to have writing units with Flora App. I am grateful for our conversations on "*so, what is formulaic language?*". Furthermore, I appreciate her forwarding of job positions at Google, increasing my motivation.

Another friend, and sister, I would like to thank is Fariba Keihani for always keeping my motivation level up and keep on writing! Further, I would like to thank Rose Katagiri for the great conversations on formulaic language, writing, and thinking. Further, I would like to thank Daniela Henry who brought to my attention *Google's AI invents its own language* as I was working on the thesis in its early stages. I would also like to thank Marie-Catherine Allard for her kind help regarding double-checking Deep L's translation as a native French speaker.

I would like to express my heart-felt appreciation to Joan Grant for her administrative help and guidance, starting the directed reading through to the thesis. I appreciate her going over deadlines, balancing all credits out, always responding quickly and helpfully to the many questions about the administrative aspects of the thesis. I also would like to thank her for creating a calm mood just before my thesis defence by exchanging our Shanghai trip adventures.

Regarding my Shanghai trip, I would like to thank my uncle, Masoud Nazari, and his wife, who has become my friend, Sepideh Kakavand for their hospitality throughout

the trip and delicious chicken kabab barbeque as I finished writing chapter 4 (Results and Analysis).

Last but not least, I would like to thank Jaffer Sheyholislami for chairing my thesis defence. I would like to also thank my MA thesis defence committee, Arshia Asudeh [the external examiner], Daniel Siddiqi [the internal examiner], and Michael Rodgers [my supervisor] for taking the time to read through the thesis and to ask great questions leading to discussions that have helped shape some of the paths of future research. It is an honour having them as my chair and committee.

I would like to dedicate this paper to my loving Parents, Hamed Shafia and Eshrat Ahmadi, and my dearest brother, Pooria Shafia, who have encouraged me to dream big and work hard to reach it. This thesis project is only a beginning of this journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

GT	Google Translate
HMT	Hybrid machine translation
MT	Machine translation
MWU	Multiword units
NMT	Neural machine translation
SMT	Statistical machine translation

Chapter 1: Introduction

> As is often the case, when we review the information in an area, new ideas start to tumble all over one another as we stand back and reflect on what gaps still exist. This is how researchers tend to operate... read, think, reflect, and wait for the inspiration to design new projects and set out on new avenues of exploration.
>
> —Wood, 2015, p. 159

1.1. Machine Translation and the Rise of Formulaic Language Research

Translation is an important facilitator of communication globally for a wide range of needs. The ongoing demand for fast and effective translation has led to the rise of machine translation (MT) technologies. As MT continues to develop in scope and capacity, the question remains whether MT is able to (and does) accurately translate input material (i.e., source language), whether at the word level or entire bodies of texts.

A number of studies have focused on MT reliability in various fields (Balk, Chung, Patel, Winifred, Trikalinos, & Chang, 2012; Groves & Mundt, 2015; Mathers, Degenhardt, Wiessing, Hickman, Mattick, & Strathdee, 2010; Nguyen, Reide, & Yentis, 2009; Patil & Davies, 2014). Of the many MT outlets, Google Translate (GT) is one of the most commonly used and most thoroughly studied. For example, in the field of medical communication, GT translations have been found to be particularly error-prone (Mathers, Degenhardt, Wiessing, Hickman, Mattick, and Strathdee, 2010, Nguyen et al., 2009; Patil & Davies, 2014), especially when Persian is either the source or the target language. These errors have been attributed to GT's utilization of word-by-word translation and disregard for formulaic nature of language. In response to this issue, the present study investigates the role of multiword units (MWUs) in formulaic language in MT for English (i.e., source language) to Persian (i.e., target language) translations, with

a focus on medical communication. The rationale behind this focus in the medical communication area is the increased immigration movement of Persians to English-speaking nations, due to the current political turmoil in Iran, and the need for translation for individuals with less advanced English levels in doctor-patient communication. Many come from educated backgrounds and are, thus, more expected to be utilizing outlets such as GT. Furthermore, there has been a number of studies (Mathers et al., 2010; Nguyen et al., 2009; Patil & Davies, 2014) testing GT's reliability in the medical communication area; however, this has not been carried out for Persian.

A translation can be considered adequate when the intended (i.e., source) message or feeling of a statement or utterance is conveyed from the source language and received as-is in the target language (Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Lafferty, Mercer, & Roossin, 1990). In other words, the intentions deduced from the utterances in both the source and the target languages must be the same. This means that speakers of both of the languages have the same interpretation. For example, if a message is supposed to create a happy feeling in the recipient, such feeling must be experienced by both of the recipients despite their language difference. Sometimes an adequate translation can be achieved via a word-for-word approach; however, in more complex constructions, this approach is often insufficient. Formulaic language is an example of a difficult-to-translate complex construction.

Many researchers have examined formulaic language (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Bu, Zhu & Li, 2010; Choueka, 1998; Conklin & Schmitt, 2012; Cortes 2007; Hyland, 2008; Men, 2018; Myles, Hooper, & Mitchell, 1998; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Simpson-Vlach & Ellis, 2010;

Weinert, 1995; Wood, 2002; Wray, 2002, 2005, 2008; Wray & Perkins, 2000). According to Wray (2008), "[f]ormulaic language is a term used by many researchers to refer to the large units of processing–that is, lexical units that are more than one word long" (p. 3) where the overall intended meaning cannot be derived from the sum of its parts (Men, 2018). This notion of *semantic transparency* is common across formulaic language scholarship, regardless of field and purpose (e.g., second language learning, machine learning, language processing). In brief, semantic transparency primarily refers to the overall meaning of less transparent sequences, such as formulaic sequences, the meaning of which are less likely to be deduced through a word-for-word approach.

Conklin and Schmitt (2012) presented statistics from various studies (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Erman & Warren, 2000; Foster, 2001; Howarth, 1998; Oppenheim, 2000; Rayson, 2008; Sorhus, 1977) that drew attention to the existence of formulaic language, which constitutes one-third to one-half of a given discourse (p. 46). Given the importance and prevalence of formulaic sequences in everyday language use (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Cortes 2007; Cowie, 1998; Hyland, 2008; Nattinger & DeCarrico, 1992; Wood, 2002; Wray, 2002; Wray & Perkins, 2000), it is important to examine how MT performs in translating formulaic sequences (Wray, 2008).

Most current MT approaches are based on statistical models (Bahdanau, Cho, & Bengio, 2015; Bu et al., 2010; Choueka, 1998; Mariño, Banchs, Crego, Gispert, Lambert, Fonollosa, & Costa-jussà, 2006; Men, 2018; Quinn & Bederson, 2011). In brief, bilingual corpora form a database from which machine learning—in this case, learning to translate—takes place. The MT systems translate based on the probability of occurrence

of certain identified language patterns through a corpus search of its database. Such systems, based in statistics, do not consider linguistically-motivated rules of language (Patil & Davies, 2014), such as those involved in formulaic language. This thesis thus attempts to bring formulaic language more to the fore in the field of machine translation in order to render more accurate, more efficient, and more natural MT translations. Working towards this goal, the reliability of English to Persian GT translations of formulaic sequences in medical communications is tested through an evaluation of GT translation results by a manual translator. Such an analysis intends to identify patterns in adequate and inadequate MT translations.

1.2. Research Questions

The following research questions guide the present study:

1) What is the reliability of Google Translate (GT) when it comes to translating multiword units (MWUs) identified in medical corpora compared to manual translation?
2) What pattern forms can be observed in how GT approaches translating formulaic language?

1.3. Overview of Chapters

This chapter (Chapter 1, Introduction) has briefly introduced current MT systems and their reliability and provided a general sense of the theory of formulaic language. It has also argued for a focus on a rule-based approach to MT systems that considers theories of formulaic language.

Chapter 2 (Literature Review) reviews theoretical and empirical research about formulaic language and machine translation. The first half of the chapter defines relevant

terms in the formulaic language literature as well as approaches to identifying formulaic sequences. The second half of the chapter discusses current MT systems, such as the statistical system mentioned above, and evaluates their reliability based on previous scholarship.

Chapter 3 (Methodology) outlines the methodology of the present study. More specifically, the chapter addresses how the corpora used in this study were assembled, the basis on which specific formulaic sequences were chosen from the corpora, and the processes of translating and analyzing the selected formulaic sequences.

Chapter 4 (Results and Analysis) presents the accuracy results for the GT translations as well as examples of analysis of some of the translated data. Chapter 5 (Discussion) connects insights from the literature review to the results of Chapter 4.

The final chapter (Chapter 6, Conclusion) summarizes the key findings of the study and considers its limitations. The chapter also acknowledges potential future research projects that may stem from this work.

Chapter 2: Literature Review

Technological innovations, such as advanced computerized systems, and the increasing global need for social and professional communication between speakers of different languages have led to advanced machine translation (MT) technology. In an accurate translation, the intended meaning in a source language is conveyed, exactly, in a target language (Brown, et al., 1990). In other words, the segments in both the source and the target language contain and convey the same meaning, messages, and feelings to their respective recipients.

Accurate translations can be achieved through a word-for-word approach in cases where there is a one-to-one correspondence of words in a given data for source and target languages. When a word or string of words in a source language is not exactly reflected in the target language, however, an equivalency approach can be used. For example, the word-for-word translation of the English phrase *green and easy* into Persian 'sæbz va sadɛh'[1] does not reflect the intended meaning of the English phrase (i.e., "easy to deceive"). An equivalent Persian translation for this phrase would be 'sadɛh-dɛl væ sadɛh-loh', which evokes the same concept as in English.

The above example demonstrates how the intended meaning of whole strings of words cannot be deduced from the sum of its parts. These "strings of words" are referred to as *formulaic language* and form the basis of much of language use. Little research has investigated the particular role of formulaic language in machine translation; thus, this chapter examines theories of formulaic language and brings into perspective the role it might have in the field of machine translation.

[1] Throughout the thesis, every time something is translated, it appears in single quotations regardless of it being the source (i.e., English) or target (i.e., Persian) language.

The chapter is organized into two main sections. Section 2.1 reviews key terms and theories of formulaic language research (2.1.1), the role of compositionality and strength of association in defining formulaic sequences (2.1.2), approaches to studying formulaic language (2.1.3), and current applications of formulaic language (2.1.4). Section 2.2 focuses on machine translation, in particular three types of machine translation systems (statistical machine translation, hybrid machine translation, and neural machine translation; 2.2.1) and studies about MT reliability, such as those that use formulaic language (*n*-grams) as a way of measuring reliability (2.2.2).

2.1. Formulaic Language

2.1.1. Defining formulaic language.

According to Wray (2005), the term *formulaic language*[2] refers to "[w]ords and word strings which appear to be processed without recourse to their lowest level of composition" (p. 4). In Wray's (2008) view, formulaic sequences are processed in a similar manner to morphemes (p. 12). Such formulaic sequences are referred to as morpheme equivalent units (MEUs). This means that the overall meaning of sequences is not derived from the sum of the parts in an MEU. In other words, formulaic sequences are memorized chunks that have a holistic meaning, where such meaning is not derived by examining the internal lexical or grammatical composition (Wray, 2002, p. 116). Moreover, MEUs can have slots with variable elements. Formulaic sequences can have a

[2] Scholars use various terminology to refer to formulaic sequences of formulaic language. Upon discussion of their views, this thesis uses each scholar's terminology and notes that the term refers to formulaic sequences in parenthesis. The discussion of the distinction between terminology used to reference formulaic sequences is beyond the scope of this thesis as the aim is to observe MT performance with respect to formulaic language. Therefore, the label used to reference formulaic sequences is irrelevant here. It is in studies where the focus is on second language learning and speech fluency that such discussion becomes important.

wide range of idiomaticity. In other words, formulaic sequences can be less restricted sequences with less idiomaticity to pure idioms with strong idiomaticity.

A similar definition of formulaic language by Choueka (1998) states, "[a] Multiword Expression (MWE) (i.e., formulaic language) is a sequence of neighbouring words whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components" (as cited in Bu et al., 2010, p. 116). This means that the intended meaning in the overall proposition of the sentence cannot be achieved by adding up the meanings of the individual words in the expression. In other words, the meaning of the sequence as whole does not reflect the compositional meaning of the sequence. Bu et al. (2010) claim that the semantics intended by the expression *kick the bucket*, for example, cannot be derived from each of the words individually. Similarly, *in this case* can literally mean "in this [briefcase, computer, or legal] case" or it can be used with a more abstract, formulaic meaning, "in this situation", as is commonly used in written language or formal speech. Interestingly, according to Choueka (1998), MWEs (formulaic language) and single words are more or less the same in one's mental lexicon (as cited in Bu et al., 2010, p. 116). This point will be further discussed in Chapter 5.

According to formulaic language literature, *transparency* and *opacity* are key characteristics of formulaic language and are of primary consideration in phraseological studies (see section 2.1.3.2). These concepts are the base criteria for determining whether or not a sequence can be considered formulaic. Non-idiomatic sequences are referred to as *transparent* because "the meaning of the whole combination can be deduced from the meaning of the individual elements" (Men, 2018, p. 21). For example, the whole meaning

of *in the middle of* can be deduced from the individual parts. In contrast, idiomatic sequences, where "the semantics of the whole combination is... not made up of the sum of their constituents" (Men, 2018, p. 21), are considered *opaque*. For example, the whole meaning of *break the ice* (e.g., in a conversation with someone) cannot be deduced from the individual words in the sequence. This distinction helps with the classification and categorization of sequences, which can range from less restricted formulaic sequences (i.e., sequences with less formulaicity, such as *if* X, *then* Y) to idioms (i.e., sequences with high formulaicity, such as *kick the bucket*).

Formulaic language theory is relatively consistent across the literature; however, differences between individual scholars can be observed in terms of the means through which formulaic sequences are identified and categorized—for example, through a frequency-based (section 2.1.3.1) or phraseological (section 2.1.3.1) approach. In essence, "formulaic language is the sense that certain words have an especially strong relationship with each other in creating their meaning–usually because only that particular combination, and not synonyms, can be used" (Wray, 2008, p. 9). Wray (2008, p. 9) offers the example of *out of the question*, a formulaic sequence which cannot be rephrased as **external to the query*. In sum, all definitions by various scholars point to the notion that the more idiomatic a sequence, the less an intended meaning can be derived or deduced from a word-for-word consideration.

2.1.2. Key features in formulaic language studies.

There are two key features examined in formulaic language research: *compositionality* and the *strength of the relationship between elements in a sequence* (Wood, 2015). These features are further described in the sections that follow.

2.1.2.1. Compositionality.

In brief, formulaic sequences are noncompositional and there is a strong relationship among the elements, hence the high frequency of their occurrence in language (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Cortes, 2007; Cowie, 1998; Hyland, 2008; Nattinger & DeCarrico, 1992; Wood, 2002; Wray, 2002; Wray & Perkins, 2000). According to the principle of compositionality (Cann, Kempson, & Gregoromichelaki, 2009), the meanings of the words in a complex expression contribute to the overall meaning of the complex expression, in part due to the ways in which the words are combined together based on the syntactic rules of a language. This principle is particularly relevant in translation research, as it emphasizes that translation using a word-for-word approach is unlikely to be as accurate as if the meaning of the whole sequence is considered.

The compositionality of a complex word or sequence refers to whether the overall meaning is predictable or derived from its components (Lieber, 2010). This is comparable to the notion of opacity discussed earlier. In compositional sequences, the sum of the parts adds up to the overall meaning. In non-compositional expressions, the opposite view holds true, meaning that the overall meaning intended by the complex word or sequence cannot be deduced from the meaning produced by the sum of its parts (Wood, 2015; Wray, 2005). Non-compositional meanings are also referred to as having figurative or idiomatic meaning (Wood, 2015). Bu et al., (2010) provide additional insights on measuring the (non-)compositionality of formulaic sequences. Compositionality is further discussed in terms of its applicability to formulaic language in section 2.4.

2.1.2.2. Strength of the relationship between elements in a sequence.

In order to discuss the relationship between components that constitute a formulaic sequence, two terms must first be defined: *head* and its *dependent* elements. In theoretical linguistics, generally, "[t]he main element in each syntactic phrase is called its head... [and] ...[t]he other elements that combine with the head to become a phrase might be called the dependents of the head" (Lieber, 2010, p. 135). In other words, the head is the element in a compound with more semantic importance, while the other elements modifying it are its dependents (Haspelmath & Sims, 2013).

Relevant to the present thesis, many formulaic sequences can be considered syntactic phrases (i.e., stand-alone constituents[3]) and therefore can be examined for their head and dependent elements. For example, in the formulaic phrases *in this case* and *on the other hand*, the noun *case* in the first sequence and the noun *hand* in the second sequence are the semantic heads, as they are the least frequently occurring elements in the sequence compared to the words *in*, *this*, *on*, *other*, and *the*. *Case* and *hand* are also what grammarians call *content*[4] *words*, as they carry the primary semantic meaning of the phrase. Content words in a phrase are always accompanied by *function words*, which are dependent on the head (i.e., content word). Content words can be thought of as the meaning-carrying bricks of a building, and the function words as the mortar that binds and gives the bricks structure. In the above examples, *in*, *this*, *on*, *the*, and *other* are

[3] A constituent is "[a] linguistic unit which is an element of a larger construction" (Crystal, 1992, p. 80).

[4] It is important to mention that, although the focus here is on content words, there are also multiword units where *function* words may play as leading heads. It is limited to content words herein due to time and space constraints on the scope of this thesis project.

function[5] words and dependent on the content head words *case* and *hand*, as they serve grammatical purposes, but do not contribute meaning. Further linguistic analysis of these concepts is beyond the scope of this paper.

In formulaic sequences, there is a strong association between head words and function words occurring together. This means that their occurrence together is greater than chance as a formulaic sequence expresses a particular intended meaning. Together, the compositionality and strength of the relationship between elements in a formulaic sequence play a significant role in how an intended meaning is derived from the whole sequence as opposed to a word-for-word approach, i.e., from sum of its parts.

2.1.3. Formulaic language identification.

There are two broad categories of research methodologies for identifying formulaic sequences: *distributional* and *phraseological*. A distributional methodology identifies sequences via corpus analysis and frequency cut-offs, and classifies them as formulaic based on their discourse function (Wood, 2015). This method is hereafter referred to as *frequency-based*, as per Wood (2015). A phraseological methodology, in contrast, takes a top-down approach, where the criteria for classifying a sequence as formulaic are based on a given sequence's semantic or syntactic composition (Wood, 2015). Researchers employing this approach either study lists of particular formulaic sequences or study formulaic sequences obtained from texts based on predefined syntactic and semantic criteria (Wood, 2015). The two approaches are further described below.

---

[5] The function words (i.e., grammatical elements) *in*, *on*, *the*, and *of* are highly frequent words in English. In formulaic language studies (Wood, 2002), however, formulaic sequences with such highly frequently-occurring function words aid speech fluency in second language learners.

2.1.3.1. Frequency-based approach.

The fundamental assumption of a frequency-based approach is that high frequency sequences are more likely to be formulaic. This approach, often used in corpus linguistics studies, is considered effective with respect to speed and cleanliness (Wray, 2008). Nonetheless, there is consensus in the literature that frequency alone is not sufficient in determining the formulaicity of sequences. A high-frequency sequence requires "a unitary meaning or function, and perhaps a particular way of being mentally stored, retrieved, or produced as well" (Wood, 2015, p. 20). In other words, although frequency may be a useful measure for identifying formulaic sequences, it must be considered alongside the unitary meaning requirement. Both must be considered when developing sequence-identification methods.

There are three key factors in a frequency-based approach: *frequency cut-off*, i.e., number of times of occurrence of a sequence in a corpus; *range*, i.e., the number of text files in which a sequence occurs (Anthony, 2014); and *assignment of functional categories to the identified sequences* (Biber, 2006; Simpson-Vlach & Ellis, 2010; Wood, 2015). A frequency-based approach usually yields better results in large corpora with text files from specific language registers, genres, or disciplines (e.g., academic English). Of note, scholars may also use a percentage criterion—similar to the range criterion—whereby a sequence must occur in a certain percentage (i.e., 40%) of all texts in order to be considered a lexical bundle (i.e., formulaic sequences) (Hyland, 2008).

There are a number of criticisms of the frequency-based approach (Wood, 2015). One criticism is with respect to corpus size. Specifically, the threshold frequency cut-off may not be an appropriate setting for smaller corpora. For example, a frequency cut-off

of 10 may yield different results for a small corpus with 3 million words versus a large corpus of over 10 million words. In other words, the low frequency cut-off of 10 may be more suitable for the smaller corpus more than for the larger corpus. A frequency-based approach delimits the detection of any identified [structural] patterning in lower-frequency sequences (Wray, 2008, p.101-2). In other words, structural patterning in strings of words that are formulaic but have lower frequencies may not be included in analysis as they do not appear in the list of highly frequent sequences.

Nor would this approach be suitable for corpora of varying registers and genres, as the occurrence of sequences rendered as formulaic would be potentially reduced to one or two occurrences (i.e., idiosyncratic sequences). This means that, certain formulaic sequences may be technical (i.e., belonging to a certain register), and thus used specifically in a particular field. For example, the formulaic sequences[6] *center of the galaxy*, *existence of time-space*, *milky way*, and *solar system* are more likely to be used in fields that are related to the outer space (e.g., field of study as in astronomy; media such as movies that are related to the outer space, The Martian). Similarly, some formulaic sequences are only used in certain genres (e.g., written or spoken). Some examples of formulaic sequences in written genre include *in the course of, as a function of*, and *by a factor of*; and in spoken genre include *do you read me?*, *we're gonna have to*, and *we've got to*.

Furthermore, a pure frequency-based approach does not provide "information about the psycholinguistic validity of the formulas" (Wood, 2015, p. 21). In a task where

[6] These outer-space-related formulaic sequences are obtained from the present thesis' scholar's pilot project carried out in Winter 2017 under the supervision of Dr. Michael Rodgers, *In pursuit of formulaic language flags for computerized translation*.

participants were asked to reconstruct formulaic sequences from memory, the reconstructions observed were different among the participants (p. 21).

Simpson-Vlach and Ellis (2010) argue that for a frequency-based method to be successful it must demand other criteria to rule out meaningless sequences of words that appear in the corpus analysis results. The scholars employed criteria including mutual information scores and instructor intuitions to identify multiword units (MWUs; i.e., formulaic sequences) in ESL students' speech or writing, and excluded non-psycholinguistically salient sequences from their list of formulaic sequences (p. 9-14).

2.1.3.2. Phraseological approach.

Unlike a frequency-based approach, the phraseological approach classifies formulaic sequences according to a sequence's semantic or syntactic composition (Wood, 2015). In this approach, a researcher analyzes lists of particular formulaic sequences (Wood, 2015) or formulaic sequences that are extracted from text based on predefined syntactic and semantic criteria (Wood, 2015). This is a top-down approach.

A number of scholars employ a phraseological approach (Asomova, 1963; Cowie, 1994; Mel'čuk, 1998; Men, 2018; Vinogradov, 1947). Although each researcher has their own specific definitions and classifications, the common characteristic is the compositionality of elements in a string of words with respect to its level of semantic transparency. The level of semantic transparency varies with respect to the relationships between the elements in a given string of words. The remainder of this section provides further information regarding the different approaches to formulaic language from a phraseological standpoint (i.e., sequences with predefined syntactic and semantic criteria).

Because the term *collocation* (i.e., formulaic sequences) was used in early formulaic language studies, this thesis will continue to use this same terminology as Wood (2015) as an umbrella term for *formulaic language*[7]. Cowie (1994, as cited in Wood, 2015) does not have a restrictive perspective on collocations with respect to the length of a given sequence, and instead provides two classifications on a scale-basis: *composites* and *formulae*. Composites refer to sequences (i.e., collocations) where the elements at the sub-sentence level (i.e., in theoretical terms, at a constituent level) carry lexical or syntactic functions (e.g., *spill the beans*; Cowie, 1998, p. 213). Formulae concern a higher scope, i.e., sentence-level, and carry pragmatic functions (e.g., *a cat has nine lives*; Cowie, 1998, p. 219).

Similar to Cowie's (1994) formulae, Vinogradov (1947) and Amosova (1963) base their classification of collocations (i.e., formulaic sequences) "according to... [a] phraseological units['s]... semantic and pragmatic functions" (Wood, 2015, p. 39). There is overlap between the classifications of Vinogradov (1977) and those of Cowie (1994), with addition of one more category by the former. Vinogradov's (1977) categories including *opacity*, *literal/figurative meaning*, and s*tructural fixedness* are comparable to Cowie's (1994) *composites* and *formulae*. The additional category in Vinogradov's (1977) classification of collocations is *contextual boundaries*. Vinogradov's (1947) refers to such collocations as *phraseological combinations*, which Amosova (1963) later referred to as *phrasemes*. Amosova's (1963) approach "... outlin[es] specific parameters within a word combination" (Wood, 2015, p. 39) when classifying collocations.

[7] The term *multiword unit (MWU)* is a terminology adopted from Simpson-Vlach and Ellis (2010) to refer to formulaic sequences in the present study which is presented and further discussed in Chapter 3 (Methodology). In other words, from Chapter 3 onwards, formulaic sequences are referred to as *multiword units (MWUs)*.

It is worth noting that Vinogradov's (1947) early classification considers the noncompositionality and nonsubstitutability of formulaic sequences, where the former refers to morphological consideration (i.e., compositionality of morphemes), and the latter to syntactic consideration (i.e., tests for constituency). Phraseologists in Vinogradov's era recognized that one of the elements in a collocation is in a leading position. This is similar to the content word discussion in section 2.1.2.2.

Nearly 50 years later, Mel'čuk (1998) established the Meaning-Text Theory for the classification of formulaic sequences via their semantics and pragmatic functions (as cited in Wood, 2015, p. 39; Cowie, 1998). Unlike Vinogradov (1947), Mel'čuk (1998) highlighted the specific internal relations between elements in a collocation (i.e., formulaic sequence). Collocations under this view, "are not free and noncompositional, and… the specific relations between the words in a collocation cause it to be perceived as a single unit of meaning" (Wood, 2015, p. 39). This dynamic marks one of the elements in the formulaic sequence as the leading (i.e., head) item, and the other as its dependant. Thus, according to Mel'čuk's (1998) classification, the combination and participation of the leading and its dependant forms the whole meaning of a collocation.

Men (2018) describes collocations (i.e., formulaic sequences) as "close constructions… which represent a unit" as opposed to "free combinations… which are constructed on the basis **on** grammatical rules" (p. 20). Similar to previously discussed views, Men (2018) holds that semantic transparency and opacity play an important role in defining collocations. In non-transparent strings of words (i.e., collocations, formulaic sequences), the overall semantics of the intended proposition is not obtained by the semantics of the sum of the parts (Men, 2018, p. 21). For example, the overall meaning of

dying cannot be obtained from the sum of the components of *kick the bucket* (i.e., kick + the + bucket).

2.1.4. Applications of formulaic language research.

Scholars studying formulaic language have mainly focused on second or foreign language learning (Myles et al., 1998; Weinert, 1995). This tendency has impacted approaches to the classification and study of formulaic sequences, as research primarily focuses on second language learning and speech fluency (Wood, 2010). In formulaic language studies, there has been an "increasing awareness of the prevalence of ready-made memorized combinations in written and spoken language", and particularly with respect to their role in "first– and second–language and adult language production" (Pawley and Syder 1983; Peters 1983 as cited in Cowie, 1998). A notable finding from the literature is that formulaic sequences are said to be processed more quickly by second language learners compared to non-formulaic phrases (Conklin & Schmitt, 2008). Psycholinguists (Conklin & Schmitt, 2012; Gibbs & Gonzales, 1985; Schweigert, 1986) have conducted experiments with respect to processing time. Findings suggests that the more proficient a speaker, the more likely they process formulaic sequences in a similar manner to native speakers (Conklin & Schmitt, 2012). This finding is corroborated by Biber et al. (2004), Cortes (2007), Biber and Barbieri (2006), Hyland (2008), and Wood (2010), who also suggest that more proficient learners appear to make more use of lexical bundles (i.e., formulaic sequences).

Interestingly, formulaic sequences appear to have a lower processing time than non-formulaic sequences for both native and non-native speakers (Conklin & Schmitt, 2008). These findings support the perception that formulaic sequences "appear to be

prefabricated which means that they are stored and retrieved whole from memory at time of use, rather than being subject to generation or analysis by the language grammar" (Wray, 2005, p. 9). This implies that the elements in a given sequence where there is formulaicity are not processed on a word-for-word basis, and instead are stored in the mental lexicon (i.e., cognitive memory) as a whole.

Based on this view, a less proficient language learner may grasp the literal meaning of, for example, *in this case* (i.e., "in this briefcase, computer, or legal case"), but not the abstract meaning of *in this case* (i.e., "in this situation"). Likewise, sequences with both literal and figurative (idiomatic) meanings may cause difficulties in machine translation (MT). This is illustrated in the phrase *kick the bucket* (Gibbs, Nayak, & Cutting, 1989). Literal meaning refers to the meaning derived from a word-for-word decomposition of a sequence of words. *Kick the bucket* can have the literal meaning of an individual (or any animate being) kicking a bucket. In this case, word-for-word translation would be sufficient. Figurative or idiomatic meaning, on the other hand, refers to the deduced meaning of a sequence of words as a whole. If the idiomatic meaning of *kick the bucket* (i.e., dying) is intended, a word-for-word translation would be inadequate.

Additional MT difficulties can arise when formulaic sequences have non-formulaic linguistic equivalents—for example, *break the ice* versus *break the cup* or *pop the question* versus *ask the question* (Conklin & Schmitt, 2008). Even though the above paired phrases follow the same linguistic composition, one is considered as an idiom (the former) and the other is not (the latter). Such subtlety can be accounted for by a human translator with pragmatic competence and needs to be considered in machine translation.

2.2. Machine Translation

Hovy, King, and Popescu-belis (2000) observed, "more has been written about MT evaluation over the past 50 years than about MT itself!" (p. 43). Despite a lack of in-depth information regarding MT, the next section (2.2.1) presents a selection of published literature about the underlying mechanisms of three current MT systems (*statistical*, *hybrid*, and *neural*). The second section (2.2.2) investigates MT reliability.

2.2.1. Machine translation types.

There are three prominent machine translation systems: *statistical machine translation*, *hybrid machine translation*, and *neural machine translation*. All three systems are based in a statistical bilingual-corpora approach (Bahdanau et al., 2015; Mariño et al., 2006; Quinn & Bederson, 2011). This approach uses probability factors to identify language patterns by aligning segments of bilingual documents in their database (e.g., the Internet). For example, if there is a document in both English and Persian in the corpus, the machine translator compares and matches the two versions to identify matching patterns. This task is accomplished through data alignment (Brown et al., 1990; Brown et al., 1993), where every word in the source language is matched with a word in the target language. This is illustrated in the example in Figure 1:

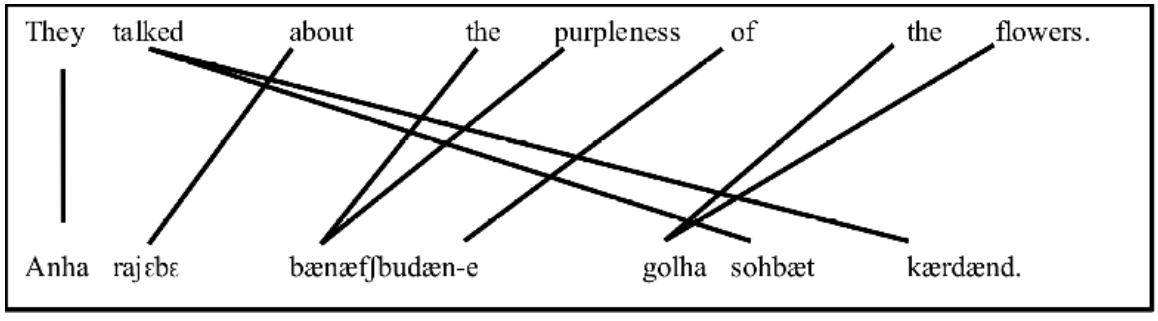


Figure 1. Alignment of words in English to Persian translation in statistical models.

In a many-to-many alignment model (notably used in the study carried out by Mariño et al., 2006, presented in section 2.2.1.1), alignment is done via a "stack search" process (Brown et al., 1990). In this case, a partial alignment hypothesis is utilized where a list of corresponding words is created as in the example below from Brown et al. (1990, p. 81):

(Jean aime Marie | John(1))*,
*(Jean aime Marie | *loves(2)*)*,
*(Jean aime Marie | *Mary(3)*)*,
(Jean aime Marie | Jeans(1))*.

Such approach seems to be practical for shorter-length data or longer-length data for languages of similar structures. This is because there would be higher chance of one-to-one (i.e., word-for-word) alignment of the elements in the data that would lead to accurate translation results.

Despite this similarity among the three machine translation systems, there are considerable differences observed among them, as outlined in the next three sections. The first section (2.2.1.1) presents the statistical machine translation system. The second section (2.2.1.2) discusses the hybrid machine translation system. Finally, the third section (2.2.1.3) examines the neural machine translation system.

2.2.1.1. Statistical machine translation.

Statistical machine translation (SMT) is the oldest type of machine translation. An element of the SMT approach that has remained intact over the years is that it translates based on bilingual corpora (Mariño et al., 2006), as explained in the previous section (2.2.1). According to Brown et al. (1990), prior to their study, statistical methods had

been primarily used in automatic speech recognition, lexicography, and natural language processing [widely referred to as NLP] (p. 79).

Historically, when statistical methods in machine translation were introduced scholars used a word-based model, such as that used in Brown et al. (1990) where sentences were treated as strings of words without structure (p. 84). Since then, it has evolved and been replaced by a phrase-based approach, such as that used in Mariño et al. (2006), where an *n*-gram-based model was used.

In Brown et al.'s (1990) study, the probability (represented by the symbol *Pr*) model employed in the statistical machine translation system states, "every sentence in one language [i.e., source language, *S*] is a possible translation of any sentence in the other [i.e., target language, *T*]" (p. 79). In this model, for every sentence pair (*S*, *T*), there exists a probability, $\Pr(T|S)$, whereby a translator produces *T* in the presence of *S*. For example, based on this model, Brown et al. (1990) anticipate that the $\Pr(T|S)$ for *Le matin je me brosse les dents*|*President Lincoln was a good lawyer* would be considerably small in comparison to $\Pr(T|S)$ for *Le president Lincoln était un bon avocat*|*President Lincoln was a good lawyer* (p. 79). In doing so, this model operates by employing Bayes' theorem which is as follows (p. 79):

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)}$$

To simplify this, Brown et al. (1990) suggest readers to think of the above equation as follows: "the translation probability… suggest[s] words from the source language that might have produced the words… observe[d] in the target sentence [whereby] the

language model probability… suggest[s] an order in which to place these source words" (p. 79). In other words, this equation helps in achieving the most probable translations on the basis of assigning values to the probability of every piece of language to be the translation of a number of words (i.e. a phrase or a sentence) and choosing the one that has the highest rank in the hierarchy of the said probability.

The following step to this model is the "stack search" described in the previous section (2.2.1). To restate, probable translation for each of the words in the source language is placed in the probable position in which it occurs. For each word, this process is presented in separate lines as in the example below from Brown et al. (1990, p. 81):

(Jean aime Marie | John(1))*,
*(Jean aime Marie | *loves(2)*)*,
*(Jean aime Marie | *Mary(3)*)*,
(Jean aime Marie | Jeans(1))*.

This means that, '*John*' is probable translation of *Jean* and occurs in first position, indicated by '*(1)*', and is followed by another element, marked by '*'. The second line shows that '*loves*' is the probable translation for *aime* which occurs in the second position, as indicated by '*(2)*', which follows and is followed by another word, as indicated by the '*'s before and after it.

Brown et al. (1990) use the term *n*-gram to refer to the number of individual words in a sentence, rather than using *n*-gram based approaches to refer to number of words in a formulaic sequence. In other words, Brown et al. (1990) use this term differently than in the study by Mariño *et al*. (2006), who use it to refer to words at the phrase-level.

In Mariño et al. (2006), the bilingual units, called *tuples*, "... are extracted from a word-by-word [i.e., word-for-word] aligned corpus in such a way that a unique

segmentation of the bilingual corpus is achieved" (p. 529). It is further explained that the extraction of such units is based on a many-to-many alignments method of sentence pairs. Although this method yields successful translations in similarly structured languages, it poses problems for languages with differing structures (Mariño et al., 2006).

Similar to formulaic language classifications and categorization via a frequency-based approach, the SMT model herein requires other feature functions to yield optimal results. Such feature functions, then, would be comparable to criteria in formulaic language research. The reader should refer to Mariño et al. (2006, p. 534-535) for a detailed explanation of proposed other feature functions (target-language model, word-bonus model, and two lexicon models—source-to-target and target-to-source) that may lead to more optimal SMT results. Such a discussion is beyond the scope of the present thesis.

2.2.1.2. Hybrid machine translation.

A hybrid machine translation (HMT) approach refers to a system that requires a combination of human computation and machine resources (Quinn & Bederson, 2011). Based on the arguments brought forth by Quinn and Bederson (2011), this model was essentially built to aid MT cost and quality—i.e., improve cost-efficiency through computer contribution and accuracy through human input, where the human translator would be paid less because of the involvement of computer technology. Quinn and Bederson (2011) suggest that, under this hybrid model, there is no requirement for professional bilingual human translators as the human tasks in their apparatus, as described below, does not demand high proficiency in bilingual translation skills (p. 3). For the same reason, there would be a lower-rate pay for the workers.

Quinn and Bederson (2011) introduce what they call *CrowdFlow*, a framework that runs on the basis of the hybrid model. Similar to SMT and neural machine translation (NMT; section 2.2.1.3) systems, HMT works based on statistical measures with bilingual corpora. Under this model, computer machines (CM) get trained through human input (i.e., machine learning). For a given task, there are three possibilities: (1) CM provides a correct answer, leaving the human with the simple task of reviewing and accepting; (2) CM provides a partially correct answer which the human will have to correct; or (3) CM provides a completely wrong answer which the human will need to completely replace (Quinn & Bederson, 2011, p. 2). Quinn and Bederson (2011) used Google Translate (GT) and Amazon Mechanical Turk[8] (AMT) workers to test their *CrowdFlow*, apparatus for Chinese-to-English translations.

In this model (Quinn & Bederson, 2011, p. 2-3), translation errors (i.e., problematic words or phrases) by GT are initially highlighted. The system, then, reproduced new sets of translations via detailed alignment of data. The new translations are paraphrased by human workers and provided to the system which substitutes them into the original source data. Similar to the discussion in previous section, there are many substitution possibilities. Next, a set of candidate translation results are produced by GT which are rated by the human workers.

They speculated that when the second possibility occurs (i.e., a partially correct answer corrected by a human translator) the problematic phrases that are paraphrased by the humans include formulaic language not identified as formulaic by GT.

[8] AMT is marketplace with the need for human intelligence, see: https://www.mturk.com

2.2.1.3. Neural machine translation.

Neural machine translation (NMT) is fairly new. This approach is unique for its comparability to the human mind due to its use of neural networking with encoded mathematical vectors. The neural networks have long- and short-term memory units (Bahdanau et al., 2015, p. 2) and work similarly to the neural networking in the human brain regarding deep learning and reinforcing abilities. Like neural activation in the brain, NMT uses an encoder–decoder system that has certain activations in the digital network. Bahdanau et al. (2015) describe the NMT encoder–decoder system deployed as follows:

> An encoder neural network reads and encodes a source sentence into a fixed-length vector. A decoder then outputs a translation from the encoded vector. The whole encoder–decoder system, which consists of the encoder and the decoder for a language pair, is jointly trained to maximize the probability of a correct translation given a source sentence. (p. 1)[9]

The encoder and decoder vectors are composed of mathematical functions. The system explained by Bahdanau et al. (2015) means that certain patterns activate certain other patterns in the encoder–decoder layers, creating a network which seems to be similar to the neural network in the human brain. For example, a 'sequence to sequence learning' model presented by Sutskever, Vinyals, and Le (2014)–utilized by Google–, phrases are organized by meaning into clusters (e.g., word order, p. 6). This concept is similar to the *word association* concept in psycholinguistics where meaning plays an important role in associating words into groups. For example, if the words *nurse* and *doctor* are presented together, subjects have a shorter reaction time, rather than *nurse* and *butter* (Church & Hanks, 1990). Such strong association and co-occurrence of words is similar to ideas of

[9] See this video for a visual representation of the neural network system: https://www.youtube.com/watch?v=aircAruvnKk

formulaic language theory were elements in a formulaic sequence have stronger associations.

In an NMT approach, increased sentence length can pose problems, as all of the source language information is compressed into a fixed-length vector in the neural network model (Bahdanau et al., 2015, p. 1). To account for this shortcoming, Bahdanau et al. (2015) propose an extended version of this method: adding to the statistical-based approach, an extended system can re-rank candidates via computation of "score of source and target phrases" (p. 9). The ultimate goal of NMT is to detach it from statistical approaches altogether to avoid the above issues. This is further discussed in chapter 6 (Conclusion).

Due to the relative similarity of NMT to the neural networking of the human brain, it is speculated that such system may be able to account for formulaic language similarly to a human mind. This means that, in theory, the machine could learn to distinguish the nuances of literal and non-literal meanings (i.e., intentions & propositions) based on context. Furthermore, because it is said that NMT has short- and long-term memory units, it is speculated that identified formulaic sequences could be stored in its long-term memory. These arguments are similar to those brought forth by Conklin and Schmitt (2012), Pawley and Syder (1983), and Wray (2005), who argue that the human mind stores and retrieves pre-fabricated language chunks. It is speculated that translation results in an NMT system would be more reliable than other statistical-based approaches.

2.2.2. Machine translation reliability.

Despite the advantage of cost-effectiveness compared to human translation, machine translation—most notably Google Translate (GT)—lack consistency in

producing precise, well-formed, and understandable translations (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014). GT translations have questionable reliability and require verification (Groves & Mundt, 2015, p. 114). Patil and Davies (2014) claim that GT's inconsistency is due to its approach to translation, which seems to be deficient in making use of some necessary linguistic concepts, such as syntax (sentential structures in a language) and semantics (concerned with meaning in a language).

As mentioned previously, current machine translators rely on statistical approaches, which means that patterns of structures are gathered (i.e., "learned") from bilingual corpora. This approach does not make use of linguistic rules and concepts required for an accurate translation. In support of this point, Patil and Davies (2014) argue that "[G]oogle Translate uses statistical matching to translate rather than a dictionary/grammar rules approach, which leaves it open to nonsensical results" (p. 1). Thus, it requires implementation of linguistic rules within its system for a better translation quality.

Despite fundamental problems with its translation methods, GT has proved useful in situations where a rough translation is sufficient to render an approximate understanding of texts for certain daily uses. GT may also be sufficient, to some extent, in certain fields and areas. Generally, there has been little research on GT's reliability and uses; however, there are studies evaluating GT's reliability in medical communication, a selection of which is presented below.

2.2.2.1. Google Translate in medical communication.

One area in which use of GT seems to be popular is in medical communications between doctors and patients when the two parties speak two different languages, or

when researchers examine medical documents in languages other than English (Mathers, Degenhardt, Wiessing, Hickman, Mattick, and Strathdee, 2010; Nguyen et al., 2009; Patil and Davies, 2014). In Mathers et al. (2010), GT was used as a translation tool for non-English websites, as well as for documents that focused on HIV prevention, treatment, and care services (p. 1016). Two studies in particular assessed GT's reliability by either exploring this matter specifically or using GT in their study. They are described below. In another study conducted by Nguyen et al. (2009), GT was used for the translation of prevalent anaesthetic questions (p. 96). A major finding regarding reliability was that GT had a better performance rate for common Western languages like Spanish, German, and French (Nguyen et al., 2009, p. 97).

Patil and Davies (2014), meanwhile, focused on the evaluation and accuracy of GT when translating medical statements[10], noting, "many translations were completely wrong" (p. 1). They advised that GT should not be used/relied on for major medical decisions such as "… consent for surgery, procedures, or research from patients or relatives unless all avenues to find a human translator have been exhausted, and the procedure is clinically urgent" (p. 2). Similar to Nguyen et al. (2009), Patil and Davies (2014) found a noticeable accuracy for Western European languages; nonetheless, in total, only 57.7% of all translations were correct (p. 1).

The above studies illustrate the performance rate of GT, in particular its overall inadequacy in translating medical communication between English and a range of other languages. Neither of the studies mention examining formulaic language when testing for

[10] Patil and Davies' (2014) study stems from an interaction with a child's parents who did not speak English and the urgency to communicate with the parents regarding their very sick child in pediatric intensive care unit (p. 1).

GT's reliability. To date, research has not been carried out on GT's performance with English to Persian translations in the medical communication area, especially with respect to formulaic language theory, a gap the present thesis work aims to address.

Chapter 3: Methodology

The rationale motivating the current thesis' focus on medical communication stems from the increased immigration movement of Persians to English-speaking nations that has increased the need for doctor-patient communication translation—in particular in instances where the doctor speaks English and the patient speaks Persian and translation is required. Moreover, the Persian immigrants often have an educated background and are, thus, more expected to be using GT as a means of communication.

To carry out the present study, medical-specific written and spoken corpora were created to identify multiword units based on a protocol addressed in a later section (3.2.2.1). This research used AntConc (Anthony, 2016), and adopts a framework by Simpson-Vlach and Ellis (2010) for the identification of *n*-grams of the multiword units (MWUs) in each of the sub-corpora (i.e., written and spoken). Moreover, this thesis also adopts the terminology by Simpson-Vlach and Ellis (2010), i.e., *multiword unit* (MWU), to refer to formulaic sequences identified in the present study. Other terminology used often have specific definitions that are particular to formulaic language studies in particular areas (e.g., Wood (2010) uses lexical bundles for English for Academic Purposes, i.e., EAP classes). Multiword unit seems to be a terminology that can represent formulaic language more generally, and thus suitable to adopt to refer to formulaic strings of words in an MT study. In addition to frequency measures, the present study adopts a protocol as guiding criteria in the identification of MWUs in the sequences extracted from the corpus search.

A total of 20 non-transparent MWUs were identified and translated for decontextualized analyses (3.2.2), and 51 sentential examples were identified for

contextualized analyses (3.2.3). Decontextualized refers to MWUs on their own (e.g., *a great deal*), and contextualized means that the MWUs were given context by appearing in full sentences (e.g., *there's a great deal of interest in osteoporosis associated fractures*). The identified MWUs were translated with GT and the results categorized by error types. Both the decontextualized and the contextualized translation data by GT were reviewed by the present study's manual translator for accuracy.

The first section below (3.1) discusses how each corpus of the medical corpora (i.e., written and spoken) was created. The second section (3.2) presents information on how the data was collection, including the settings for n-gram, frequency, range and formatting AntConc, the selection of MWUs, and sentential examples for each MWU. The third section of this chapter (3.3) discusses the translation process for GT and the review process by the manual translator. The final section of this chapter (3.4) outlines the analysis procedure for the GT translation results.

3.1. Building the Corpora

To build the corpora, both written and spoken texts were compiled. For the written corpus, journal articles that were related to medical communication were obtained through a Google Scholar search with the journal name as the search string. Also, articles from the BioMed Central (BMC) corpus were obtained by scrolling down the website's main page. Further medical articles were obtained from Medical News Today, Microbiology Research, Scientific America, The Guardian, and Web MD.

For the spoken corpus, text files were collected from four sources. First, texts from the British Academic Spoken English (BASE) corpus were obtained from "Life and Medical Sciences" in "BASE Files" under "Search". Second, texts from the Michigan

Corpus of Academic Spoken English (MICASE) were selected. Under the "Browse" section, "Native speaker, American English" for the status of speakers was chosen. The academic division was set to "Biological and Health Sciences" and the academic discipline to "Biology", "Chemistry", and "Nursing", which are related to the field of medicine. Third, texts were obtained from Technology, Entertainment, and Design (TED) Talks, where the primary selection criteria were that the talks had transcripts and were related to the medical area. Finally, subtitles of episodes from two medical TV shows, *Grey's Anatomy* and *House MD*, were obtained. In the case of the TED talks subtitles, all the lectures that met the selection criteria were included.

The next section (3.1.1) presents the preparation process for the written and spoken corpora (i.e., the source of texts to be analyzed, how they were organized into single files, as well as information regarding word count illustrated in Table 1). This is followed by a section (3.1.2) with an explanation of the preliminary measures taken in cleaning the data in text files in preparation for corpus analysis.

3.1.1. Preparing the Corpora

The composition of the corpora built for this study is presented in Table 1. The text files for each item in each corpus (e.g., Scientific America, Web MD, The Guardian) were condensed into one text file to aid processing by AntConc, as AntConc can malfunction with an increased number of text files. For example, the content of the 12 text files for BMC Medicine were compiled into one single text file. For *Grey's Anatomy*, each season was compiled into one text file per season, with a total of 12 seasons. The 12 seasons were then compiled into one single text file. The above compilation of several files into a single text was done for all the components in the written and spoken corpora.

Ultimately, the written corpus consisted of seven text files and the spoken corpus consisted of five text files.

The two corpora individually utilized the settings described in section 3.2.1. to identify MWUs, as the *n*-gram setting did not result in valuable MWUs when the two corpora were combined. In other words, because the two corpora belong to two different genres (i.e., written and spoken), the list of MWUs obtained from the corpus search through AntConc was not valuable. The number of words and the sub-corpus distribution were not equal between the corpora; this was not of concern in this project, as the corpus was utilized to extract MWU constructions to aid machine translation, and not for an intra/inter-disciplinary comparative study of formulaic language that would have required a more balanced corpora with respect to word count within the written and spoken corpora or between the two.

Table 1.

Composition of the Corpora

Written			**Spoken**		
Sub-corpus	**No. of Text Files**	**Tokens**	**Sub-corpus**	**No. of Text Files**	**Tokens**
BMC Medicine	12	59, 215	BASE	29	232, 378
Medical News Today	2	2, 666	Grey's Anatomy	12	1, 490, 853
Microbiology Research	45	390, 297	House	7	723, 676
Science Blogs	20	14, 017	MICASE	3	35, 445
Scientific America	2	8, 585	TED Talk	22	45, 599
The Guardian	11	7, 004			
Web MD	18	11, 760			
Total	110	493, 544	**Total**	73	2, 527, 951
Corpora Total			3, 021, 495		

3.1.2. Cleaning the Data

Data were cleaned for any elements that might interfere with a corpus search for MWUs. Tables of contents from text files in the written corpus were erased to exclude unwanted symbolic data; time-stamps from text files in the spoken corpus were erased for the same reason. Although the data could have been cleaned manually, it would have been a long and tedious process. Instead, cleaning was accomplished automatically via "Find and Replace" searches.

The first step was pattern identification. The timestamps, such as **00:00:04,377 --> 00:00:06,242**, followed a pattern of **#:#:#,# --> #:#:#,#** (note that there is one space before the arrow and one space after it; it is important to include the spaces as well). To obtain the pattern in the "Find" space, the magnifying glass was selected, followed by "Insert Pattern". The desired element was chosen and then in "Find" the pattern was inserted. In "Replace", nothing was added. Clicking "All" replaced all of the timestamps in the text files where there were time stamps. This entire procedure was then applied to the remaining text files with the same timestamp pattern. Once another text file was opened, "Find and Replace" automatically transferred the identified pattern to the "Find" space. Every time the input in the "Find" space changed, the new form was transferred to the new text files; therefore, the same input was applied to all of the text files before cleaning other elements in the text file.

Empty spaces and line numbers were also excluded through "Find and Replace" searches. To do this, the empty spaces above each line number as well as the line number were selected and copied into the "Find" space. Then, the specific number with a "# Digits" was changed to generalize it for a greater pattern. In "Replace", nothing was

added. Clicking “All” replaced all of the instances of this pattern in the text file with empty spaces, deleting them from the text file.

Another element that was removed from the data was one-word units in square brackets. To do this, the whole unit, including brackets, was copied into the “Find” space and the word inside the brackets was changed to “Word Any Word Character” via the “Insert Pattern” function. Any unwanted space gaps or other elements (e.g., , :, - dash with a space after it) were deleted with the copy-and-paste function in the “Find” space. Within the same file, this function keeps recent searches in case there are instances left that were not previously removed. Once the file is closed, this feature is no longer accessible.

In the text files, some words were attached to each other, as in *itas* as opposed to *it as* (more examples included *wayto* and *checkfor*). The separation was done manually by inserting a space between them. Beyond impacting the word count, this separation was important for the identification of MWUs during the AntConc search.

3.2. Preparing for Corpus Analysis

In this section, preliminary information on data collection is presented. The first section (3.2.1) describes the setting requirements for MWUs (n-gram, frequency, and range) and formatting for AntConc. The second section (3.2.2) discusses the selection of the MWUs; here, the protocol principles (3.2.2.1) as well as the protocol steps (3.2.2.2) that were taken into account in selecting the MWUs are outlined. Moreover, this section also focuses on the selection of sentential examples for each MWU (3.2.2.3).

3.2.1. N-gram, frequency, range, and formatting AntConc.

This study adopts the framework by Simpson-Vlach and Ellis (2010) for *n*-grams (i.e., number of words in a sequence), frequency (i.e., number of time an MWU occurred in the corpus), and range (i.e., distribution of MWUs across text files in each corpus). Simpson-Vlach and Ellis (2010) set their *n*-gram size to 3-gram, 4-gram, and 5-gram; frequency to a minimum of 10; and range to a minimum of 3. An example of a 3-gram sequence is *in this case*, a 4-gram sequence is *on the other hand*, and a 5-gram sequence is *would you like me to*. The settings required a 3-gram, 4-gram, or 5-gram MWU to occur 10 times and across 3 different text files. This type of setting yields more practical results than if, for example, the n-gram is set to a minimum of 2 words because often 2-gram sequences (i.e., collocations) are contained within 3-gram sequences.

A pilot search with 183 files (total for spoken and written corpora) showed that a 2-gram setting resulted in a number of sequences such as *and the*, *it is*, etc.—types which are impractical as they are sequences that do not conform to the definition of MWUs as described in section 2.1. To recap, MWUs carry functional meanings (e.g., referential bundles, stance expressions, and discourse organizers). Collocations like *and the* and *it is* do not have a specific function and thus do not belong to a functional category, as they are meaningless in isolation. Compare this to *and also* which is categorized as "'relators' [under]… macro-organizers that serve the… function of signaling high-level relations in a discourse" (Nattinger & DeCarrico, 1992, p. 102). Such sequences are not suitable examples to be included in the data analysis. To put it differently, they may be thought of as "by-products" of corpus results. Moreover, as Simpson-Vlach and Ellis (2010) point out, such collocations occur frequently and are often contained within 3-gram or 4-gram

sequences. Therefore, because of their impracticality and the likelihood of 2-gram sequences being contained within 3-gram, 4-gram, and 5-gram sequences, 2-grams were excluded from the current search.

Similar to Simpson-Vlach and Ellis' (2010) frequency criteria, the lowest cutoff of 10 is chosen as other criteria are taken into account in selection of the final list of MWUs to be analyzed. The difference here lies in the measures where they use statistical measures, whereas, the current thesis proposed a protocol (3.2.2.1) as a measure to narrow down the large list of sequences extracted from each corpus.

The combination of a frequency-based approach *and* a criteria-approached would yield in formula that would serve Simpson-Vlach and Ellis' study's purpose in a desired practicality aspect. In other words, in the Simpson-Vlach and Ellis (2010) study, psycholinguistic saliency and pedagogical relevance was of importance, and they argue that "... a formula [being] above a certain threshold and distributional range does not necessarily imply either..." (p. 7). Similarly, a purely frequency-based approach would only lead to a number of highly frequently occurring sequences that are either linguistically meaningless, or formulaically functionless in which case, a word-for-word translation would result in the correct forms in a target language and thus would not need to be classified as MWUs by the established definition in this project (D. Siddiqi[11] & H. Shafia[12], personal communication, Fall 2017). The formulaic sequences (MWUs) herein are chosen to serve the specific function of aiding a more natural salient translation, and are thus narrowed than based on certain criteria discussed in the consequent sub-section.

---

[11] Associate Professor of Linguistics, Cognitive Science, and English, & Assistant Director of SLaLS (Linguistics) at Carleton University

[12] See section 4.3.4 for further information regarding the educational background of the programmer for this project, as he is also the Official Translator.

Distributional range (i.e., occurrence of sequences in more than one text file) across text files was also another important factor in the compilation of MWUs, as it "ensures that the formulas on the list are not attributable to the idiosyncrasies of particular speakers or speech events" (Simpson-Vlach & Ellis, 2010, p. 11). This aspect was taken into consideration in the present study. What is excluded in this research was the calculation of the mutual information (MI) score, which is "a statistical measure commonly used in the field of information science designed to assess the degree to which the words in a phrase occur together more frequently than would be expected by chance" (Simpson-Vlach & Ellis, 2010, p. 11). Simpson-Vlach & Ellis (2010) employed this scale to examine its usefulness in ranking the formulas in their list, as it extracts formulas in the corpus that seem more suitable for pedagogical objectives. However, the present study does not require such ranking of formulas (i.e., MWUs), hence the exclusion of this statistical measure.

Formatting AntConc settings was also required prior to analysis to fix punctuation. This was accomplished by checking off "Punctuation" through the following steps: Global Setting, Token Definition, Punctuation Token Classes. These steps fixed issues regarding contractions—for example, *i*[13] *don t* vs. *i don't*, where the former mistakenly counts as a 3-gram sequence, and the latter correctly as a 2-gram sequence. See Appendix A for more examples from a visualization in AntConc.

3.2.2. Selection of multiword units.

Simpson-Vlach and Ellis (2010) present Biber et al. (2004) as a contrast to their study by arguing that reliance on frequency-based approach "... alone generates too many

[13] The reason why *i* is not capital is because this is how it shows up on AntConc.

items of undifferentiated value" (p. 29). Of importance, Simpson-Vlach and Ellis (2010) mention that "[s]election criteria that allow for intuitive weeding of purely frequency-based lists, yield much shorter lists of expressions..., but they are methodologically tricky and open to claims of subjectivity" (p. 7). This is a key consideration when it comes to writing the protocol principles/steps. This is because it must be as specific and linguistic-logically driven as possible so it reduces subjectivity. In other words, the MWUs are not selected at random; rather, they are selected because of their degree of transparency that may cause issues for word-for-word translations.

The MWUs analyzed in the present study were selected based on their polysemous and/or non-compositional readings. A total of 20 non-transparent MWUs (Table 2, in section 3.2.2.2) were selected based on the protocol principles and steps outlined in the next two sections (3.2.2.1. and 3.2.2.2). The 20 non-transparent MWUs are the decontextualized data (i.e., none of the 20 MWUs appeared in a sentence; e.g., *do you mind*, see section 3.2.3. for an example of a contextualized MWU). As discussed in the previous sub-section, a wholly frequency-based approach was considered insufficient in fulfilling the objectives of this project. Instead, in a given sequence it was deemed important to identify the leading or head element (i.e., content word) and the depending elements (i.e., usually function words) and pay close attention to the relationship between them. To do so, a protocol was created based on four principles, presented below.

3.2.2.1. Protocol principles.

A protocol of four principles was proposed for the present study to guide the selection of MWUs from the sequences in the corpus search in AntConc. The four principles are explained in this section.

Principle One.

MWUs must be selected based on their content words, preferably the head word, such as the noun or main verb, as function words are too limited in number of types[14] and are too frequently used in the majority of MWUs. Checking every function word to determine whether they participate in MWUs or not would likely bring a computer code to a halt. Therefore, it is important to designate the word or words in a multiword phrase that appear(s) in language less frequently, i.e., content words. This step will tremendously reduce the processing time for computer codes and also mitigate possible mistakes by reducing the number of occurrence of such words in a string of words, i.e., phrases. For example, in the sequence *on the other hand*, the head word *hand* is the least frequent word in comparison to the other elements in the sequence and yet the most significant part of the phrase, and would be the best choice to raise the flag for the code to look for the multiword unit in which the word "hand" is found along with the rest of the words in the unit.

Principle Two.

The MWUs that can be translated correctly into their equivalents in the target language word-for-word must be identified and excluded from the list. For example, the collocational phrase *Royal Garden* in *Cyrus walked the men to the Royal Gardens,* would not be problematic when being translated from English to Persian or vice versa because the equivalent Persian words for the words *Royal* [*Sœltænœti*] and *Garden* [*Bagh*] individually are exactly the same as the ones used as the translation of the phrase as a

[14] There are several function words in comparison to content words in which there is an endless number of words.

collocation. In this case, the translation *Bagh Saltanat*[15]*i* works perfectly. Therefore, such MWUs should not be included in the list of MWUs from the medical corpus to be analyzed.

Principle Three.

In machine translation, the location of the head word (or any word that has been designated as the flag to notify the computer code to look for multiword sequences) must be identified. A human brain is able to go back and forth in a sentence and identify a MWU as formulaic language; however, a computer code needs to be instructed as to how many words it should go back and/or forth to find a possible multiword unit in order to find the correct translation for it. For example, the word *fly* in the sequences *on the fly* and *when pigs fly* occurs at the end of the sequence, but the number of words before this flag word is two for the first sequence and three for the second. Therefore, for the above phrases the computer code needs to be instructed to go back two words and three words, respectively, in order to put the words together to form a MWU in its memory. This problem takes into account the time required to select MWUs. For the computer code, a short routine can be written to go back and forth and put the words together into an MWU sequence.

Principle Four.

In certain cases, words might occur in a string of words that is an MWU, but also in an identical string that is not a formulaic segment. Consider the following examples:

In this case, I will have to study harder.
In this case, the accused is exonerated.

[15] In Persian, adjectives appear after nouns. This can be taken into account in programming by providing syntactic rules.

In the first sentence, the phrase *in this case* is an MWU and should be translated into Persian as *dær in sooræt*, which is a multiword unit in Persian as well. Looking more closely at this example, the word *sooræt* at the end of this phrase is not a congruent meaning for the word *case*; rather, the word *sooræt* as a part of the phrase *dar in sooræt* renders the exact equivalent for the phrase *in this case* in the first sentence.

In the second sentence, *in this case, the accused is exonerated*, the phrase is not an MWU and the correct Persian translation for it would be *dar in parvændɛh*, which is a word-for-word translation of the English phrase that renders the literal meaning of a legal case. These sequences must be identified and the flag words therein categorized as dual-use words. This is because in some cases, such sequences are considered as MWUs, while in other cases, they are not. Probably the most perplexing items for machine translation are such units and the routines for processing them are the most complicated sections of the code.

The linguistic flags to identify such multiword units and find the appropriate translation for them must be based on other signifiers in the context where these units occur. In many cases, semantic and pragmatic and circumstantial signs can be identified in the preceding and/or succeeding words and/or sentences. Although it would not be an adequate theory to be utilized for devising algorithms for machine translation, the approach adopted by Fillmore (1976) may be illuminating regarding what signifiers might be looked at for context to help the computer code decide which meaning it should pick in a sequence containing dual-use words. It might be worth noting that the same problem exists for the case of polysemy. Identification of such signifiers would resolve the choice of correct translation polysemy word as well.

3.2.2.2. Protocol steps.

Based on the principles outlined above, four steps have been designed in the identification of the multiword units: (1) find the content word in the sequence; (2) identify whether or not it is a polysemous word—if yes, choose as a multiword unit in DB; (3) if the content word identified in (2) is not polysemous, identify whether the sequence as a whole, i.e., with its function words have a contextual meaning that is different than the word-by-word meaning; and (4) check the identified MWU in (3) to see if it has dual-use.

Table 2, on the following page, illustrates the 20 non-transparent MWUs identified for analysis. The table is organized on an alphabetical-order basis, rather than by frequency count. This is because a frequency-based approach is not the focus of the present study, rather it is the principles in the proposed protocol that have influenced their selection.

Table 2.

20 Non-Transparent MWUs

Corpus & *n*-gram		MWUs	Frequency	Range
Spoken				
3-gram	**1**	a clinical trial	12	4
	2	a great deal	13	3
	3	comes down to	17	3
	4	do you mind	37	3
	5	get away with	12	3
	6	how could you	38	3
	7	in this case	29	4
	8	let you know	40	3
	9	to figure out	90	3
	10	would you mind	16	3
	11	would you please	11	3
	12	we're dealing with	12	3
4-gram	**13**	a hell of a	27	3
	14	do you know what	71	4
	15	i have no idea	43/41	3/4
5-gram	**16**	to come out of hiding (House MD)	1	1
	17	would you like me to	13	3
Written				
4-gram	**18**	in the absence of	104	3
	19	in the case of	94	3
	20	at the same time	19	4

3.2.2.3. Selection of sentential examples for each multiword unit.

In this step, the MWUs were put into context (i.e., contextualized) of a sentence from the data in the medical corpus. There were a total of 51 sentential examples selected from the text files in the medical corpus containing the 20 non-transparent MWUs. Each of the MWUs was entered in the word search box in AntConc. The search showed in which text files each multiword unit appeared. An example from each text file was gathered on an Excel spreadsheet. The number of text files in which the 20 non-

transparent MWUs appeared added to a total of 51 sentences. This is the reason why the total number of sentences is 51. For example, a MWU only appeared in 3 of the text files (*Grey's Anatomy*, *House MD*, and MICASE) in the spoken corpus out of the 5 text files (i.e., the MWU did not appear in BASE and TED Talks) refer to Table 1 in section 3.1.1.). The sentence example for the contextualized data was usually the first sentence in the text file in which the MWU occurred. Sometimes a sentence was incomplete, in which case another was chosen.

The data on the spreadsheet were organized into two columns: *decontextualized* (i.e., only the MWUs) and *contextualized* (i.e., MWUs in the original sentences in which they appeared; e.g., *do you mind if I scrub in?*, where the underlined is the MWU). Each column was further divided into *English* (i.e., source language data) and *Persian* ([translated] target language data).

3.3. Translation Procedure

There are four steps for the translation procedure. In the first step, each MWU in the decontextualized column was translated through GT. The same was done for the sentences in the contextualized column containing the MWUs. In the second step, the translation results by GT of the decontextualized and the contextualized data were presented to the official translator (3.3.1) and evaluated for accuracy of translation. Third, if the translation was accurate, there were no changes made. Fourth, if the translation was inaccurate, the translator provided the accurate equivalent translation.

3.3.1. The official translator

An official translator reviewed and evaluated GT's translation results and provide revisions for any incorrect data. Hamed Shafia[16] was selected as the official translator of the current study due to his educational background and work experience as a translator. Highlights of his professional experience are listed below:

1. BA in Translation, MA in ESL Teaching Methodology, Tehran Azad University – Central Campus
2. Years of experience as a university professor, teaching linguistics and translation courses at the Faculty of Graduate Studies, Tehran Azad University – Central Campus
3. Oral translator (interpreter) in UNIIMOG (United Nations Iran-Iraq Military Observers' Group)
4. Translator and interpreter of international sports events
5. Law (LLB & LLM), University of Tehran
6. International Contract Lawyer and Canadian Arbitrator

3.4. Analysis Procedure

To analyze the data, the decontextualized (i.e., individual MWUs; e.g., *do you mind*) and the contextualized (i.e., MWUs appearing in a sentence; e.g., *Do you mind if I scrub in?*) translations by GT were presented to the manual translator who evaluated for accuracy. Any translation results of MWUs, for both decontextualized and contextualized data, by GT that were accurately translated, the manual translator left intact. For the decontextualized data, any MWU that was inaccurately translated, the manual translator

[16] He is also the programming consultant for this thesis project.

provided the accurate translation. Next, based on the accurate translations of the MWUs (whether it was observed in GT translations or provided by the manual translator), the GT translations of the contextualized data was evaluated to test its performance for accuracy when the MWUs are contextualized.

The following step was to investigate the inaccurate translations for the source of errors. In other words, the task was to see what pattern forms were observed in the translation errors (e.g., whether the errors stemmed from a word-for-word translation approach of the MWUs). The accurate translations of the MWUs were also examined for any observed method they were translated (i.e., via a word-for-word or translation-by-equivalency approach). It is expected that such task provides insights into awareness, or lack thereof, of GT with respect to formulaic language.

Chapter 4: Results and Analysis

In this chapter, results (i.e., statistical reports on percentage of translation accuracy of the data) of the present study are presented, followed by analysis of the data obtained, reflecting on the research questions. For the results section, statistics for the decontextualized and contextualized MWUs are presented separately. Similar presentation holds for the analysis section where separate tables are used for the decontextualized and contextualized data used as examples. Each table includes both manual and GT translation results.

4.1. Results for Translation Accuracy

A comparison between GT and manual translations indicates that GT shows awareness of formulaic language in some cases, while in others the results were noticeably incorrect. The central issue with the GT's inaccurate results stem from a word-for-word approach (i.e., literal meaning) instead of translating the MWUs based on a translation-by-equivalency approach (i.e., formulaic, overall intended meaning), providing the equivalents of each of the multiword units.

Overall, 50% (10 of 20 MWUs) of the decontextualized MWU data and 52% (27 of 51 MWUs appearing in sentential examples) of the contextualized MWU data were incorrectly translated. The rest of this section elaborates on the results by providing empirical details for each MWU in decontextualized and contextualized data.

Of the incorrect contextualized data, six out of the 10 MWUs—used in a total of 17 sentences—had zero correct translations. In other words, none of the sentences that contained the six multiword units was correctly translated. The six multiword units are as follows: *would you mind*, *do you mind*, *to come out of hiding*, *comes down to*, *we're*

dealing with, and *a hell of a*. Importantly, a comparison of the manual translations with the translations generated by GT for these six multiword units shows that none were correctly translated in the decontextualized data, either.

In contrast, there was one instance of an inaccurate translation by GT for the decontextualized MWU *a clinical trial*; however, the MWU resulted in accurate translations in the four sentences where it appeared in the contextualized data. Furthermore, there was only one inaccurate translation by GT in the contextualized data in three out of 10 MWUs (i.e., one incorrect sentence for each, thus a total of three incorrect sentences out of total nine sentences for the three multiword units). The three multiword units include, *would you like me to*, *in the absence of*, *a great deal*. A comparison of the manual translations with the translations generated by GT shows that only the decontextualized translation for *in the absence of* was correctly translated.

Additionally, for three of the contextualized data, there were four instances of partial incorrectness for the translations generated by GT. The three multiword units include, *how could you*, *do you know what*, and *at the same time*. For these three multiword units, GT had generated correct translations for the decontextualized MWUs.

For the multiword unit *get away with*, two out of the three contextualized data were not correctly translated. GT generated an incorrect translation for the decontextualized MWU.

Contrary to the abovementioned MWUs, four multiword units were both decontextualizedly and contextualizedly correctly translated (total of 9/9 correct translations for the contextualized data).

4.2 Analysis

Twenty multiword units (MWUs) were extracted from the corpora. Each was translated both manually and with GT in decontextualized as well as contextualized forms. For the contextualized data, one sentence from each of the sub-corpora text files where the multiword units appeared was identified (a total of 51 sentences; see section 3.2.2.3 for data collection methods). GT translations were evaluated by the manual translator who provided accurate translations for any inaccurately translated MWU in both decontextualized and contextualized data. Pattern forms were identified for both the accurate and inaccurate data regarding MT's consideration of the formulaic nature of language. In other words, the nature of the errors, or lack thereof, was observed to classify any recurring patterns. This means that the translation approach (i.e., equivalency-based or word-for-word) in GT translation results were examined with respect to translating MWUs of formulaic language theory. This is further investigated in chapter 5 (discussion).

Results indicate that translation errors observed in GT results appear to stem from a word-for-word translation approach. This is illustrated in the examples in the following section (4.2.1) of four multiword units that were incorrectly translated by GT through a word-for-word translation mechanism. Both decontextualized and contextualized data are presented in a table for each example. For GT translations, the first line is translation in Persian, the second line is the inter-linear gloss (word-for-word correspondence), and the third line is the English gloss. The English gloss for some data may be missing, as it was impossible to formulate one due to the severity of errors made by GT. The acronym *DOM*, used in some tables in the interlinear glosses represents a direct object marker in

Persian, *ra*. The acronym *FUT* is used to mark future tense. In Persian, *aja* is used as a question marker.

4.2.1. Data analysis.

Table 3 shows the manual and GT translations for the multiword unit *do you mind*. The translation provided by GT is based on a word-for-word translation of the terms *do*, *you*, and *mind*. In this case, an error observed is the use of the wrong question marker (i.e., *Q-marker* in Table 4)—*aja* instead of *ægɛ*. Furthermore, in the manual Persian translation, the pronoun *you* is absent. In other words, when the MWU is translated by equivalency into Persian, the pronoun *you* is not a part of the translation. In the translation result by GT, the term *mind* is translated as a noun (as in "brain", "intelligence", "wits") rather than its verb (as in "care", "object", "be bothered") form.

Here, the word *mind* is the leading element (i.e., the head). The presence of *mind* in the MWU in the grammatical construction *do* + *you* + *mind* appears to act as an important feature, indicating the formulaicity of the multiword unit. In other words, the presence of such feature results in the consideration of such sequence of words to be rendered as a formulaic segment which would have to be translated as a whole segment rather than word-for-word. The reason *mind* is chosen as the base to process and look for the phrase *do you mind* in a translation lexical database[17] inside a software is that the head words, especially the least frequently used one, make it much easier and faster for a computer code to find the phrase in the database and replace it with its equivalent translation.

[17] The term *lexical database* is a term I have invented for the ultimate goal of my machine translation software project in the future. See Chapter 6 (section 6.4.2.) for more details regarding the argument presented here.

Table 3.

Decontextualized Translation of the MWU do you mind

English	**Persian**
	Manual Translation
do you mind	ægɛ ɛʃkali nædarɛ
	Google Translate
	*tō zɛhnæm dari you mind have-2SG 'do you have (a) mind?'

A similar approach is taken by GT when translating *do you mind* when contextualized in the interrogative sentence *do you mind if I scrub in?* (Table 4). Similar to the decontextualized form, the error with the Q-marker, by GT, is observed in the contextualized translation data, as well. An error by GT is the mistranslation of the second person singular *you*, translating it as first person singular *I* in Persian (i.e., it translates *mæn* as opposed to the correct translation *tō*). Moreover, as mentioned earlier, the second person pronoun, *you*, is not a part of the equivalency translation observed in the manual translation.

Table 4.

Contextualized Translation of the MWU do you mind

English	**Persian**
	Manual Translation
do you mind if I scrub in?	ɛʃkali nædarɛ (agɛ) mæn zɛdɛufuni konæm? Do you mind (if) (I) scrub in-I
	Google Translate
	aja mæn fɛk mikon-æm ægær mæn dær xæraʃidɛgi kon-æm? Q-marker 1SG think-1SG if I in scrub do-1SG '? I think if I in scrub do'

Similar to the above example, there is a word-for-word translation by GT for the multiword unit *would you like me to*, as illustrated in Table 5. As observed in the interlinear-gloss, GT translated each word individually without factoring in the formulaic nature of the MWU. Specifically, the term *like* loses its meaning as affection, which is the meaning observed in GT's word-for-word translation. In this MWU, the elements surrounding *like* are of importance, indicating the formulaicity of such construction. More specifically, of the surrounding elements, the infinite *to* plays the most crucial role in the prevention of such word-for-word translation. This is because here the verb is *like to* despite the presence of *me* between the two elements.

Table 5.

Decontextualized Translation of the MWU would you like me to

English	**Persian**
	Manual Translation
would you like me to	aja mixahid kɛ mæn
	Google Translate
	aja mæn ra doost darid? Q-marker I DOM like-2PL(formal) '? you like me?'

Table 6 demonstrates translations of the MWU *would you like me to* when contextualized in the interrogative sentence *would you like me to remove the vent?*. The interlinear gloss for the translation by GT indicates that its engines have not identified the string of words *would*, *you*, *like*, *me*, and *to* as an MWU. In other words, instead of treating this sequence as a formulaic unit and translating it as a chunk, GT translated the

words individually on a word-for-word basis. Further error is made with the incorrect translation of the term *vent* as “wastewater”.

Table 6.

Contextualized Translation of the MWU would you like me to

English	**Persian**						
	Manual Translation						
Would you like me to remove the vent?	aja Q-marker	mixahid kɛ mæn would-2PL like me to		havakɛʃ vent	ro DOM	bærdaræm? remove	
	Google Translate						
	aja Q-marker	mixahi would-2SG	mæn 1SG	ra DOM	bɛ to	fazɛlab wastewater	bærdaræm? remove

The MWU *comes down to*, illustrated in Table 7, further exemplifies how GT’s word-for-word translations are incorrect. In this case, the issue seems to stem from the lack of detecting *comes down* as a phrasal verb plus the infinitive *to*. Due to this misidentification, *down*, a prepositional part-of-speech, has been literally translated as “downward [direction]” into the Persian ‘pajin’. This prepositional aspect is absent in the manual translation, which is the correct translation.

Table 7.

Decontextualized Translation of the MWU comes down to

English	**Persian**
	Manual Translation
comes down to	æslɛ mozoo' in æst kɛ
	Google Translate
	bɛ pajin miajæd to down comes-2SG '2SG >>literally<< comes down'

The GT result for the contextualized data *if it comes down to it, just let me go* is shown in Table 8. The GT result for the contextualized data is not translated based on a word-for-word approach as in the decontextualized data (Table 7). It seems to show some awareness of formulaic language; however, it is not accurate. Instead of translating the MWU *comes down to* as 'kar bɛ anja rɛsid', it translates it as 'bɛ an mirɛrɛsæd'.

Table 8.

Contextualized Translation of the MWU comes down to

English	**Persian**
	Manual Translation
If it comes down to it, just let me go.	ægar kar bɛ anja rɛsid, fæGæt bogzar mæn bɛravam. If it comes down to it, just let me go
	Google Translate
	ægær bɛ an mirɛrɛsæd, fæGæt ɛʤazɛ dæhid mæn bɛrævid. if to it reach-2SG, just let- FUT 1SG go-2SG 'if it reaches it, just let me go-2SG'

The example in Table 9 demonstrates that the MWU *a great deal* was also incorrectly translated by GT's word-for-word approach. However, it is also an example of context-sensitivity, as the translation provided by GT may be considered correct in some specific contexts, e.g., business deals. In other words, it is the elements in the rest of a full sentence that would aid triggering the correct choice of the manual or GT translation. In other words, if the formulaic aspect of the string of words *a great deal* is considered with respect to the context in a sentence (as in data in Table 10), then the manual translation must be chosen as results. However, if it is a business deal that is the context of a particular sentence, then the translation by GT would be considered accurate (as in the car company example in the following paragraph).

Table 9.

Decontextualized Translation of the MWU a great deal

English	**Persian**
	Manual Translation
a great deal	ta hædɛ zijadi
	Google Translate
	jɛk mo'amɛlɛ bozorg a deal(business) big 'a big deal'

In the case of *a great deal*, GT's result could be a suitable translation in the context of a grand business deal. For example, in the sentence *His purchase of the car company was a great deal in the company's history*, the verb *purchase* can act as an important element in triggering a more accurate translation. In contrast to the inaccurately translated decontextualized example in Table 9, in the contextualized data in Table 10, *a*

great deal in the sentence *There's a great deal of interest in osteoporosis associated fractures* is indeed translated by GT similarly to the manual translation. In sum, when contextualized, the MWU *a great deal* is correctly translated by GT, suggesting that GT is context-sensitive, which leads to a more correct translation. The only element missing in the GT translation is the term 'bɛsijar', which reflects the English equivalent *great*.

Table 10.

Contextualized Translation of the Multiword Unit a great deal

English	**Persian**				
	Manual Translation				
There's a great deal of interest in osteoporosis associated fractures.	ælaɢɛ ([or] tævæjohɛ) bɛsijar ziadi a great deal of interest	bɛ ʃɛkæstɛgihajɛ to fractures	naʃi æz associated	pukiɛ ostoxan osteoporosis	vodʒud daræd. there's
	Google Translate				
	ælaɢɛ ziadi a (great) deal of interest	bɛ ʃɛkæstɛgihajɛ to fractures	mærbut ba associated	pukiɛ ostoxan osteoporosis	vodʒud darad. there's

Chapter 5: Discussion

This chapter discusses the results of analysis presented in Chapter 4. The first section (5.1) addresses the first research question, *What is the reliability of Google Translate (GT) when it comes to translating multiword units (MWUs) identified in medical corpora compared to manual translation?*. The second section (5.2) addresses the second research question, *What pattern forms can be observed in how GT approaches translating formulaic language?*

5.1. Research Question 1: Machine Translation Reliability

Research question 1 was regarding the translation reliability of English to Persian translations by Google Translate (GT). Translation results by GT suggest that it is unreliable when it comes to translations to Persian both for decontextualized data, where only 50% (10 of 20 MWUs) of the data were accurately translated, and contextualized data, where only 52% (27 of 51 MWUs appearing in sentential examples) of the data were accurately translated. This is in line with previous findings (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014). For example, Patil and Davies (2014) found that only 57.7% of all GT translations for medical phrases were correct.

Also similar to other studies (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014), the inaccurate translation results in the current study are imprecise, poorly formed, and unintelligible. This is an unsurprising finding, as GT operates with a statistical-based approach with bilingual corpora, where texts in English are paired with Persian translations as available on the Internet and patterns are identified. In this approach, there is a lack of linguistic—i.e., syntactic and semantic—considerations (Patil & Davies, 2014).

Another notable finding was that the MWUs identified in the corpora were not necessarily medically-related (i.e., not technical medical language). However, this does not seem to be a major issue for translation, as the ultimate goal of a translation machine is to translate language. In other words, the MWUs still reflect the nature of the language used in medical communication between doctors and patients. This means that the MWUs identified may also be found in language from other fields. Despite this lack of field-specificity, the current study has in common with other studies (e.g., Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014) the fact that GT shows inadequate performance in translating data and cannot be relied upon for consistently accurate English-to-Persian translations.

5.2. Research Question 2: Pattern Forms Observed in GT Translation Results

The second research question sought to identify patterns in GT translations compared to manual translations. Results suggest that there are two patterns, one for the accurate translations, and one for inaccurate translations. For accurate translations, it appears that a word-for-word translation of MWUs can lead to correct results. In some cases, it appears that GT shows an awareness of formulaic language. For example, GT accurately translated *in the case of* as 'dær ʃærajɛti kɛ' and *in this case* as 'dær in morɛd', treating each as multiword unit. In the first example, the word *case* is translated as 'ʃærajɛti'; whereas, in the second example, it is translated as 'morɛd'. This shows that GT's mechanism of translation appears to have identified the two strings of sequences as formulaic, hence, the accurate translations.

For the inaccurate translations, the inaccuracy seems to stem from a word-for-word translation of the MWUs. Moreover, such inaccuracy seems to stem from GT

overlooking MWUs as formulaic. These two points—word-for-word translation & formulaic language (un)awareness—mean that although GT shows awareness of formulaic language in some cases, this factor does not seem to be systematized in its algorithms. In other words, it seems that GT can find MWUs statistically if the English and Persian MWU counterparts are correctly aligned. This means that each segment of the MWU is paired with an element in the Persian translation. However, this perfect word-for-word alignment does not exist for all MWUs. For example, there is no one-to-one alignment for the MWU *do you mind* and its Persian equivalent *ægɛ ɛfkali nædarɛ* (manual translation). Therefore, GT inaccurately translates it as:

*tō	zɛhnæm	dari
you	mind	have-2SG

‘do you have (a) mind?’

It seems that such issue stems from the lack of a rule-based approach to translation.

For MWUs with less idiomaticity (e.g., *in this case*), the overall meaning of its elements may be deduced from the sum of its components. In other words, compositional meaning may lead to the intended meaning. Hence, for the accurate translations GT’s word-for-word approach was sufficient, as translation-by-equivalency is achieved through a word-for-word translation. For such cases, it is speculated that the English phrases are properly aligned with the Persian translations within GT’s bilingual database. Thus, at times, the statistical approaches (be it neural or statistical machine translation systems) that GT utilizes appear to be sufficient.

In MWUs with stronger idiomaticity (e.g., *do you mind*), a compositional meaning does not equal the overall intended meaning, i.e., noncompositional meaning. The second observed pattern form showed that the overall intended meaning of MWUs

was not obtained by GT's word-for-word translation. In these cases, translation should have been approached by finding the equivalent counterparts (i.e., translations) of the MWUs in the target language, as the intended meaning and connotation cannot be deduced from the sum of its elements (Choueka 1998 as cited in Bu et al., 2010). For example, the MWU *do you mind* was translated by GT as "do you have a mind?", thus incorrectly representing the intended meaning and connotation which is '*ægɛ ɛʃkali nædarɛ*'. In contrast, a human translator is able to correctly translate this expression due to pragmatic competence. The human brain is capable of accessing the interface between various linguistic components such as syntax–semantics, or semantics–pragmatics; a machine translator, however, may lack such a component and not account for such subtleties. Therefore, consideration of such component (i.e., syntax–semantics or semantics–pragmatics) needs to be provided to computer software through routines (computer code segments) to address the syntax–semantics or semantics–pragmatics interfaces and thereby lead to a translation-by-equivalency approach. Thus, it is important to consider and add formulaic language theory to machine translation studies.

Because current machine translation approaches are based on statistical systems with bilingual corpora as their databases, inaccuracies in translations may stem from inadequately translated documents in the corpora, which in turn may stem from errors by human translators. For example, a translator who originally translated a document from English to Persian may have incorrectly translated an MWU, which cause it to be wrongly aligned in the statistical system, thus causing issues in the identification of patterns. If this is the case, then reliability test might be needed for the translated documents in the corpora—more so than on the end result, i.e., machine-generated

translation results. Reliability test might be carried out by adopting a similar approach as Quinn and Bederson's (2011) test apparatus, where proficient translators reviewed translated documents for quality, correcting any errors in translation. However, this would be a time-consuming and costly process.

For the above reasons, and considering possible issues with translated documents in the bilingual corpora of statistical systems, of the three machine translation (MT) approaches—statistical, hybrid, and neural—the test apparatus by Quinn and Bederson (2011) using a hybrid approach appears to be the best option, as there are human workers available to correct any errors made by machine engines. On the other hand, the comparability of the neural networking model to the human mind in the neural approach seems more promising overall (Bahdanau et al., 2015). If neural MT becomes a reality in the world of machine translation, then it is speculated that machine engines may be able to account for formulaic language similarly to the human mind—MWUs may be lexicalized as chunks (i.e., strings of words) and retrieved as whole, similar to a human brain's storing of the MWUs. It is speculated that the MT system (i.e., artificial intelligence in the area of translation) would become sensitive to pragmatic competence. *Pragmatic competence*, herein, is adopted from Nattinger and DeCarrico (1992) who adopt it to discuss formulaic language which they refer to as lexical phrases. Pragmatic competence deals with interaction of linguistic competence and pragmalinguistic competence, i.e., semantic subtleties related to context. This shift would lead to a translation by equivalency approach.

Chapter 6: Conclusion

The first section (6.1) of this chapter summarizes the findings of the present study. The remainder of the chapter addresses the implications of this thesis (6.2), the limitations (6.3) of the study, and future research directions (6.4).

6.1. Summary of Findings

The present study aimed to address two research questions:

1. *What is the reliability of Google Translate (GT) when it comes to translating multiword units (MWUs) identified in medical corpora compared to manual translation?*
2. *What pattern forms can be observed in how GT approaches translating formulaic language? (5.2).*

To summarize, similar to previous studies on machine translation reliability (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014), this thesis found that GT is unreliable for English-to-Persian translations in written and spoken medical communication texts. Of this study's decontextualized data (i.e., the MWUs in isolation, such as *do you mind*), only 50% (10 of 20 MWUs) was translated accurately. Of the contextualized data (i.e., the MWUs in a sentence, such as *Do you mind if I scrub in?*), only 48%[18] (25 0f 51 MWUs appearing in sentential examples) was translated accurately.

The findings in the presents study are in accordance to the study carried out by Patil and Davies (2014) who found that only 57.7% of all GT translations for medical phrases were accurate. The issues with the inaccurate GT translation results seem to stem

[18] In section 5.1, it was said that 52% (27 of 51 MWUs appearing in sentential examples) were inaccurately translated. Because the focus here was on the MWUs that were *accurately* translated in the contextualized data, the percentages and numbers are different in comparison to the discussion in section 5.1.

from lack of precision, well-formedness, and therefore intelligibility of the MWUs for both the decontextualized and contextualized data (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014). This is a foreseeable finding, because GT functions based on a statistical approach with bilingual corpora as its database. In statistical models, texts in English are paired with Persian translations as available on the Internet, then, patterns are identified lacking any syntactic or semantic considerations as suggested by Patil and Davies (2014).

In sum, GT's translation mechanisms are insufficient and, similar to other studies (Groves & Mundt, 2015; Nguyen et al., 2009; Patil & Davies, 2014), it is unreliable in providing accurate English-to-Persian translations. This said, an MT system whereby, for example, 60% of the translation results are accurate might be considered more efficient than the percentages presented above. The key aspect is to assess the translation results with respect to context (i.e., field, area, or setting). For example, a 60% accuracy rate where errors pose issues in communicating important messages that put a patient at risk would make the MT system insufficient. However, in a different circumstance, for example a casual conversation, a 60% accuracy rate in an MT system may be sufficient. In other words, although translation results by GT has been described to be insufficient in the medical domain in this thesis, translation results in the less serious circumstances may be sufficient despite minor errors.

Regarding the second research question, two patterns were observed: (1) GT translation through a word-for-word approach, in some cases, yielded accurate translation results, while showing some awareness of formulaic language; and (2) the lack of

awareness of formulaic language by GT due to a word-for-word translation approach, in other cases, led to inaccurate translation results.

As described in detail in Chapter 5 (Discussion), the cases where GT translated via a word-for-word approach and the translation results were accurate, the MWUs were less idiomatic. An example of this is the MWU *in this case*. In other cases where GT's translation results were inaccurate because of a word-for-word approach to translation, the MWUs were more idiomatic. An example of this is the MWU *do you mind*. This means that in the former, GT showed awareness of formulaic language, whereas, in the later, it lacked awareness of the formulaic nature of the MWU.

6.2. Implications

The results of the present study have two major implications: (1) improving the accuracy of machine translation (MT) algorithms and (2) reducing processing time.

It is suggested that the theory of formulaic language be considered and utilized in MT. This would improve the accuracy of MT, as the translation process would be based on an equivalency rather than word-for-word approach. In an equivalency-based approach, MWUs would be translated as whole units of language leading to accurate translation results. This means that the intended meaning of an MWU would be transferred from the source to the target language (Brown et al., 1990). In the current statistical approaches to MT, this approach seems to be absent for many cases of translation of MWUs whereby an accurate translation cannot be derived via a word-for-word approach.

In a word-for-word approach, the translation is rendered through a substitution of words with their equivalent words which may or may not result in an accurate transfer of

meaning from the source language to the target language (Brown et al., 1990). Moreover, syntactic rules may or may not be reflected in the translated results. For example, the MWU *comes down to* was inaccurately translated by GT, on a word-for-word basis, as:

bɛ pajin miajæd
to down comes-2SG
'2SG >>literally<< comes down'

The inaccurate translation result neither makes any sense nor follows any correct syntactic order. An accurate translation for the above MWU is provided by the manual translator as 'æslɛ mozoo in æst kɛ'.

By implementing the theory of formulaic language, machine translation approaches could, unlike statistical approaches, become more systematized and rule-governed as MWUs would be looked up in a lexical database, find the equivalent MWUs in the target language, and replace the source MWUs with the target MWUs. In this approach, both semantic and syntactic aspects are addressed adequately leading to accurate translation results. To this end, an MWU is identified, looked up in a database in order to find its equivalent translation which conveys the semantic aspect thereof to ensure that the rendered translation is understood by the native speakers of the target language as by those of the source language. This approach deals with an MWU as a single language unit for translation and disregards the elements (i.e. words) within such units. This will in turn make the process of machine translation easier and more accurate.

Furthermore, by taking formulaic language into consideration, it is speculated that processing time would be reduced because MWUs could be stored and retrieved as chunks in the long-term memory unit of the MT database, rather than looking for individual words in a multiword unit which would be scattered in the lexical database.

When an MT is provided with a sentence, it breaks down the sentence into its components (i.e., words) and looks up every single word in its database (e.g., corpora), which contains hundreds of thousands of entries[19]. Then, it finds and picks the equivalent for each word. It is noteworthy that MT looks up for each single word by breaking it into individual characters (i.e., letters, numbers, hyphens, and other symbols) and searches for each character one by one to find the word. This means that literally millions of searches and comparisons between the characters of a word and the entries existing in a database (corpora and lexical databases) are performed. This process is done for every single word. This is a time-consuming process, even if techniques, such as binary search[20], are implemented. However, if formulaic language units are identified and looked up in a database, the processing time would be tremendously mitigated.

In this approach, when a computer encounters a content word (i.e., head word) which is flagged as an MWU participant in the database, it looks for the MWUs under this word and checks to find out which MWU is an exact match to the word and its surrounding (i.e., dependent) and eventually replacing the source MWU with the target MWU. Through such approach, the computer would not require to search through

---

[19] For example, see *Computer Programming Concept and Visual Basics* (1999, 4th *ed.*, Chapter 1, p. 3-13), by David I. Schneider. Also see >http://guyhaas.com/bfoit/itp/Programming.html<. or see *Fundamentals of C++ Programming* by Richard L. Halterman (2017).

[20] Binary search is a computer programming technique in which the first character in an entry is searched in a database by means of dividing the database into two halves and see in which half the entry exists and then divide that half into two halves in turn and look for the half which contains the entry and, again, divide this very half into two halves and so on. When the first character is found in the first record, it begins looking for the second character in that half via the same process. In this technique, the computer software cuts the number of searches to almost half times if such a technique were not implemented.

hundreds of thousands of entries in a lengthy process to locate words one by one, and most likely leading to inaccurate translations.

Thus, application of formulaic language theory is expected to yield more accurate results and lowered processing times.

6.3. Limitations

In the present research, there were a number of limitations which may be observed, worked on and avoided in future studies. Some of the limitations appear to be more difficult to address (e.g., access to current MT underlying algorithms), while others seem to be less difficult to account for (e.g., methodological approaches to extracting MWUs).

One of the limitations of the present study was a lack of access to the current underlying algorithms of GT, i.e., the systems of the statistical approaches. Moreover, there was also a lack of direct access to the neural MT approach; therefore, it is unclear how neural machine translation[21] (NMT) hopes to change its database from bilingual corpora of statistical systems.

Likewise, there was a lack of knowledge about possible translation errors in the bilingual corpora documents by the human translator who originally translated them. Regardless, it is suggested that approaching MT via statistical models is inadequate; instead, there should be more focus on systems that are enriched with linguistic rules—i.e., rules from both theoretical (syntax and semantics) and applied linguistics (i.e., formulaic language).

---

[21] Stuskever, Vinyals, and Le (2014) work at Google and have written an article discussing the neural network approach in GT.

Further shortcomings of this study are observed in the methodology, specifically in the use of the corpora used to identify MWUs. Firstly, the data preparation process was time-consuming, as all the text files require cleaning for interfering symbols and/or time stamps. Also, spacing between words was lost when PDF files were converted to text files. Secondly, the number of *n*-grams was problematic. MWUs of 2-gram length are often observed in 3-gram sequences, and sometimes 3-gram sequences in 4-gram sequences. For this reason, it was more logical to begin searching for MWUs at the 3-gram level. However, MWUs of 2-gram length may have been missed if they did not appear in sequences of higher *n*-gram lengths. This means that collocations that are strictly 2-gram sequences, such as *you know* (filler phrase), *right now*, *figure out* (spoken corpus only), *fatty acids*, *immune response*, *public health* (written corpus only), were not identified.

A third methodological issue was that, as discussed by Wood (2015), frequency cut-offs may pose problems for smaller corpora such as that used in the present study because idiosyncratic sequences with low frequencies may be excluded. Related to this point, the current study agrees with Simpson-Vlach and Ellis' (2010) criticism of the need for additional criteria for ruling out meaningless sequences obtained from a purely frequency-based approach. An important issue to discuss is that, although it is a useful tool, AntConc tends to crash with large number of text files. This, in turn, seems to directly influence frequency and range, as sequences identified in the corpus search reflect only a part of the bigger corpus(ora).

It is also important to note that corpus search for formulaic sequences may yield many false positives. In other words, there may be strings of words which would seem to

be formulaic, however, a word-for-word translation of them would lead to accurate translation results (e.g., *in the middle of*). This means that such phrases would not be considered as formulaic from an MT perspective. Importantly, the elements in these sequences are high frequency words in English, and so for speech fluency they may be considered formulaic. However, in an MT view they would not be considered formulaic in the same way as the MWUs *do you mind* or *a great deal* are considered formulaic.

Another issue is that not all formulaic sequences contain content words. For example, *is made to* is comprised of function words and is formulaic, and therefore considered as an MWU. The present approach did not cover this important factor.

6.4. Future Research

As an original contribution, this thesis offers a list of semantically and syntactically non-transparent MWUs that would be useful for application of formulaic language in MT. There are many opportunities for future research as a result of this study. The first part of this section (6.4.1) discusses the possibility of using a phraseological approach to identify MWUs, while the second part (6.4.2) discusses an alternative to current MT systems' software that utilizes a statistically detached approach.

6.4.1. Phraseologically-influenced approach for identifying multiword units.

Based on the results of this study, it is suggested that an identification-through-text approach would be valuable for extracting MWUs where they are extracted directly from examining texts, rather than corpus search. This approach for the identification of MWUs is done by phraseologists who study formulaic sequences identified based on syntactic and semantic criteria from texts (Wood, 2015, p. 4). It is speculated that such an approach may aid an MT system to "learn" much like in the neural approach, with the

speculated hope of reconfiguring the underlying mechanisms of MT databases (i.e., detaching it from statistical approaches, as described in section 6.4.2).

In addition to this approach to tackling MT issues with formulaic language, it is also suggests that psycholinguistic studies (Conklin & Schmitt, 2012; Gibbs & Gonzales, 1985; Schweigert, 1986) may be advantageous for identifying formulaic sequences, as they examine sequences with nuances (e.g., *break the ice* vs. *break the table*).

Finally, although there are many definitions of formulaic language, each seems to be specific to a particular purpose (e.g., speech fluency), where its criteria are valuable only for that specific purpose. Similarly, MT studies' definition of formulaic language needs to be reformulated and narrowed down to suit the specific need of accounting for formulaic language in MT. The common factor in all definitions of formulaic language is *semantic opacity* (i.e., non-compositionality or non-transparency); thus, in the case of MT, formulaic language may be defined as sequences with semantic and syntactic opacity which cannot be translated on a word-for-word basis.

6.4.2. A statistically-detached approach.

Although the main focus of the present study has been on the role of formulaic language in machine translation, namely GT, this thesis also attempts to call for a new approach to machine translation that is completely detached from statistical approaches. Because of such intention for the future of the present study, the protocol principles of the selection of the MWUs had been influenced by the underlying mechanisms of the new approach to machine translation. In other words, the selection of the MWUs was influenced by the newly defined lexical database which does not comprise of bilingual corpora. Instead, the design of the lexical database, containing lexicalized words and

MWUs, can be comparable to a human mind. This lexical database can be provided to programmers to be used to define commands and routines for finding the best equivalent MWUs and also devising routines for syntactic aspects of a language at sentential levels. Such model is further elaborated in the following section (6.3.2.1) presenting preliminary information on what is meant by a lexical database as well as computer code.

6.3.2.1 How typical translation software work

Computer translation software typically consist two main parts: the computer code (referred to as the code) and the database. The information provided here regarding computer programming, are the ABCs of programming found in the opening chapter of any computer programming book[22].

The computer code contains a number of commands, usually clustered into "procedures" or "routines". These commands are written in a high-level[23] programming language, saved in a file named the "source code" and eventually compiled into a low-level machine language which is the format used by the computer processor and other segments to implement a series of actions intended by the programmer. For example, some of these commands order the computer to get data input from the user through input devices such as a keyboard, store them in certain cells in the memory as "variables", retrieve data from the lexical database, compare the input data with the retrieved data,

---

[22] For example, see *Computer Programming Concept and Visual Basics* (1999, 4th *ed.*, Chapter 1, p. 3-13), by David I. Schneider. Also see >http://guyhaas.com/bfoit/itp/Programming.html<. or see *Fundamentals of C++ Programming* by Richard L. Halterman (2017).

[23] *High-level* language refers to programming languages such as C++, Fortran, PHP, Javascript, and older ones like Pascal and Foxpro. The commands in these languages are closer to human language whereas the low-level languages such as Assembly look like binary computer codes. High-level languages need to be "compiled", i.e. transformed, into low-level languages which are understandable and used by computer hardware.

make certain processing and decision-makings, and eventually display the result on an input device such as the monitor or even store the result in an output file.

In this case, a user runs the program and then inputs a sentence into the appropriate place prompted on the screen. One or more routines execute designated lines of commands in the code to store the input (in this case “words”) in variables. Then, the following command lines retrieve data from the database and compare them with the input and choose the appropriate data entry, put them in a syntactic order as already defined and improvised in the source code, and finally send the result to a defined output device.

The above information is illustrated in Figure 2 on the following page:

Routines

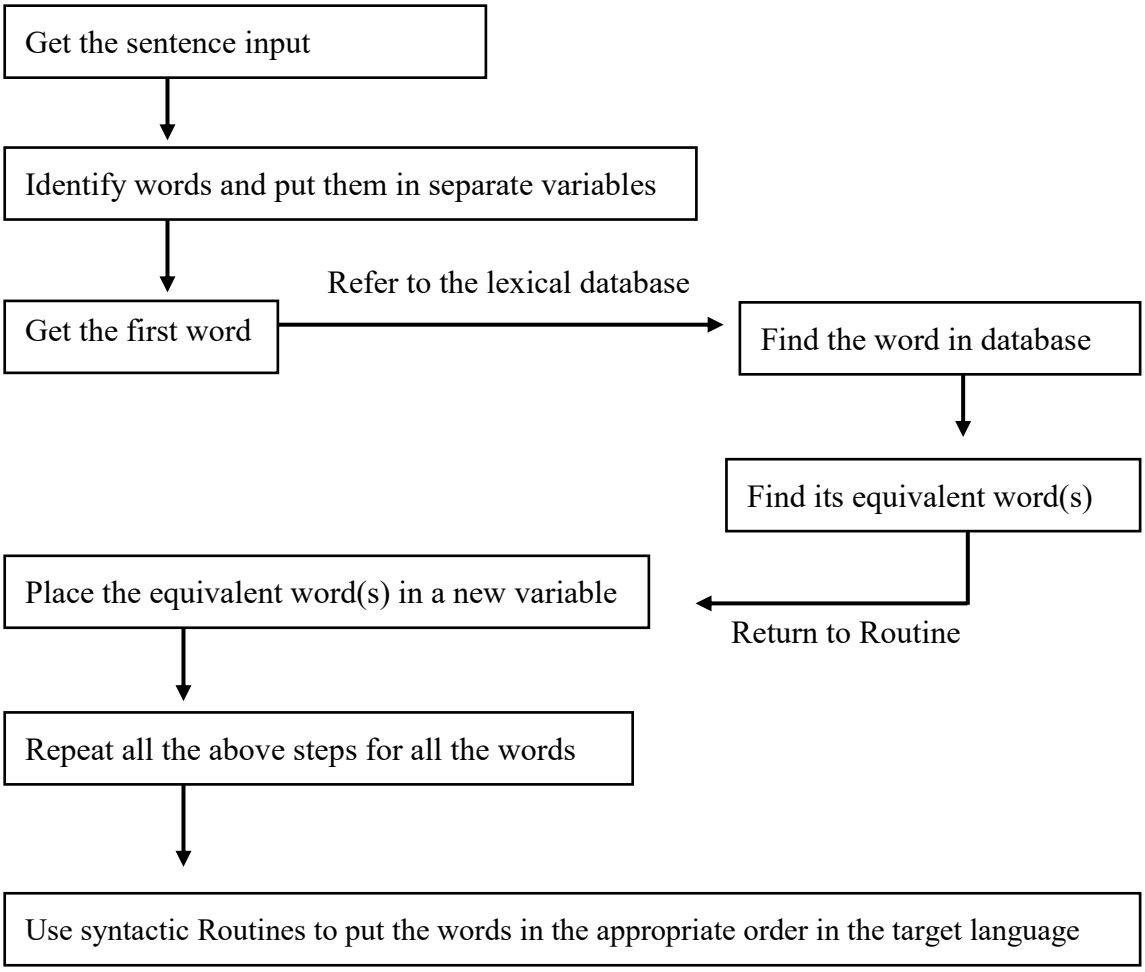


Figure 2. Programming information related to Routines.

The above schematic diagram is a simplified algorithm[24] followed to process and translate a sentence using a lexical database instead of a corpora database used in statistical models. The lexical database is a file that contains an array of data organized in certain structures. These data can be arranged and stored in two or more dimensional structures like tables. A two dimensional table of data is sufficient to meet the needs of a machine translation software. A simple table with rows and columns, in cells of which

[24] This thesis has created Diagram 1, however, similar information regarding the steps can be found in any programming book.

words, the meaning(s), and other linguistic information are stored for a lexical database. The first column contains word entries and the rest of the columns contain a series of information attached to each entry. These columns are called “fields”.

A computer program puts a whole sentence into an array of memory allocations, or variables, in that each word is put in a separate variable. In fact, the code starts getting the input characters one by one from the beginning of the input sentence, and as soon as it reaches certain special characters such as a space, it puts those characters in the first variable (Var1), and continues to process the rest of the string of characters in the same way until it hits the end of the sentence. At the end of the process, there would be a number of variables (i.e., Var1, Var2, … Var_n) temporarily kept in the memory.

Then, the code looks up each and every one of the contents of these variables (i.e. words) in its lexical database. In the above example, it would not be that difficult to find equivalents for the words as each of the words would exist as an independent entry for which there is an equivalent word in the target language, stored in the relevant “field” in the lexical database. In other words, the computer would find each word from its English-Persian lexical database. This process is done by the software for all the words. These target-language equivalents are in turn put in a number of other memory allocations, or variables.

The next step would be to put these variables in an order dictated by the syntactic rules of the target language. The final product of this process is rendered and displayed on the computer screen or stored in a file as the translation of the above sentence. This depiction of the computer translation is a simplified picture of what actually happens in the background.

In a number of cases, finding equivalent meanings for the words in a database would be simple. The code searches the word among the entries in the database, and after locating it, retrieves its meaning stored in the relevant field in that database. As an example, a word-for-word approach is sufficient for the MWU *in this case* as there is a one-to-one correspondence for the element in the sequence (i.e., in + this + case) in Persian. However, a computer software would face challenges like distinguishing the non-compositional meanings of MWUs. For example, a word-for-word translation of the MWU *do you mind* would result in an inaccurate translation as explained in previous sections (4.2. & 5.2.). For such cases, the equivalent translation of the MWU in Persian (i.e., *ægɛ ɛʃkali nædarɛ*) would have to be entered in a field in the lexical database. This means that every time the MWU *do you mind* is encountered, instead of translating it by a word-for-word approach, the computer software translates it by equivalency.

In order to identify the sequence *do you mind* as an MWU, the sequence must be flagged on its head word. The head word is the least frequently occurring element in the sequence[25]. The least frequently occurring element in this MWU is the word *mind*. Therefore, after finding the word *mind*, as an entry in the lexical database, the software would check the field assigned for the flag (True|False) showing that the word is an MWU participant to see if this word occurs in a collocation. For this entry, the response returned from the database is "T" (True) meaning that the word *mind* does occur in a collocation. Then, the software looks in the collocation field(s) and checks back the input phrase. If, and when, it finds an exact match, the software would pick the collocation's

[25] In some cases, such element may also be polysemous.

equivalent translation (i.e., stored in the appropriate field in the lexical database) and puts it in a variable for future use in the final translation of the whole sentence.

A challenge in the above algorithm would be the categorization and classification of such collocations and multiword units of language to make sure that the software knows how many words to include in its search for formulaic sequences. It is also important to define towards which directions (before and/or after a certain word in a sequence) to look onto. Moreover, there would be a need to have a routine for the cases where the words in a multiword unit are possibly separated into segments occurring in different positions in a sentence (e.g. double-word verbs, triple-word verbs, etc. – e.g. *is made to*). In other words, it remains a challenge to find clues, or "flags" to use its programming technical word, for a computer translation software to decide which meaning to pick. Such flags could be syntactic clues and/or semantic environments (i.e., preceding and succeeding words).

In both cases, application of formulaic language theory in order to identify and categorize formulaic sequences and any flags to be used for such identifications would be a big leap in the development of MT methodology and software thereof.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ICLR 2015*. Retrieved from https://arxiv.org/pdf/1409.0473.pdf.

Balk, E. M., Chung, M., Hadar, N., Patel, K., Winifred, W. Y., Trikalinos, T. A., & Chang, L. K. W. (2012). Accuracy of data extraction of non-English language trials with Google Translate. *Agency for Healthcare Research and Quality*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK95238/pdf/Bookshelf_NBK95238.pdf.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., ... & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263-311.

Bu, F., Zhu, X., & Li, M. (2010, August). Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 116-124). Association for Computational Linguistics.

Cann, R., Kempson, R., & Gregoromichelaki, E. (2009). Semantics: An introduction to meaning in language. Cambridge University Press.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO 88:(Recherche d'Information Assistée par Ordinateur). Conference* (pp. 609-623).

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Linguistics and Language Behaviour Abstract (LLBA)*, *32*, 45-61.

Cowie, A. P. (Ed.). (1998). Introduction & Phraseological Dictionaries: Some East–West Comparisons. *Phraseology: Theory, analysis, and applications*. 1-23, 209-228. Oxford, UK: Oxford University Press.

Gibbs Jr, R. W., & Gonzales, G. P. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition*, *20*(3), 243-259.

Gibbs, R. W., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28*(5), 576-593.

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, *37*, 112-121.

Haspelmath, M., & Sims, A. (2013). *Understanding morphology*. Routledge: London.

Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. *Machine Translation*, *17*(1), 43-75.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21.

Lieber, R. (2010). *Introducing Morphology*. Cambridge University Press.

Marino, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., & Costa-Jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics*, *32*(4), 527-549.

Mathers, B. M., Degenhardt, L., Ali, H., Wiessing, L., Hickman, M., Mattick, R. P., & Strathdee, S. A. (2010). HIV prevention, treatment, and care services for people who inject drugs: a systematic review of global, regional, and national coverage. *The Lancet*, *375*(9719), 1014-1028.

Mel'čuk, I. (1998). Collocations and lexical functions. *Phraseology. Theory, Analysis, and Applications*, 23-53. Oxford, UK: Oxford University Press.

Men, H. (2018). The notion of collocation. In *Vocabulary Increase and Collocation Learning* (pp. 9-33). Springer: Singapore.

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, *48*(3), 323-364.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford, UK: Oxford University Press.

Nguyen-Lu, N., Reide, P., & Yentis, S. M. (2010). 'Do you have a stick in your mouth?'–use of Google Translate as an aid to anaesthetic pre-assessment. *Anaesthesia*, *65*(1), 96-97.

Patil, S., & Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. *BMJ*.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, *191*, 225.

Quinn, A. J., & Bederson, B. B. (2011). Human-machine hybrid computation. In *Position paper for CHI 2011 Workshop On Crowdsourcing And Human Computation*.

Schweigert, W. A. (1986). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research*, *15*(1), 33-45.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487-512.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, *16*(2), 180-205.

Wood, D. (2002). Formulaic language in acquisition and production: implications for teaching. *Linguistics and Language Behaviour Abstract (LLBA), 20*(1), 1-15.

Wood, D. (2010). *Formulaic language and second language speech fluency*. New York, NY: Continuum International Publishing Group.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. New York, NY: Bloomsbury Publishing.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2005). *Formulaic language and the lexicon*. New ed. Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford; New York: Oxford University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*.

Further Reading

Formulaic Language

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97-116.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72-89.

Gentil, G. (2014). Will ESL writing teachers lose their jobs in the age of highly accurate computer-assisted translation? *Contact Magazine* (TESOL Ontario Newsletter)*, 40*(4),31-34.

Machine Translation: Approaches

Ambati, V., Vogel, S., & Carbonell, J. G. (2010). Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 2169-2174).

Amtrup, J. W., Rad, H. M., Megerdoomian, K., & Zajac, R. (2000). Persian-English machine translation: An overview of the Shiraz project. *Memoranda in Computer and Cognitive Science*.

Bezrukov, A. (2013, September). Flexible learning model for computer-aided technical translation. In *Interactive Collaborative Learning (ICL), 2013 International Conference on* (pp. 673-675). IEEE.

Gornostay, T. (2008). Machine translation evaluation. *NGSLT Machine Translation Course*.

Koehn, P. (2004, July). Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP* (pp. 388-395).

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, *43*(2), 393-427.

Leusch, G., Ueffing, N., & Ney, H. (2003, September). A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX* (pp. 240-247).

Megerdoomian, K. (2004, May). Developing a Persian part of speech tagger. In *Proceedings of the 1st Workshop on Persian Language and Computer* (pp. 99-105).

Nesson, R., Rush, A., & Shieber, S. (2006). Induction of probabilistic synchronous tree-insertion grammars for machine translation. *Association for Machine Translation in the Americas*.

Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 160-167). Association for Computational Linguistics.

Quinn, A. J., & Bederson, B. B. (2011). Human-machine hybrid computation. In *Position paper for CHI 2011 Workshop on Crowdsourcing And Human Computation*.

Sakir, E. Z., & Petrik, S. Machine Translation Evaluation.

Shieber, S. M. (1996). A call for collaborative interfaces. *ACM Computing Surveys (CSUR)*, *28*(4es), 143.

Shieber, S. M. (2007, April). Probabilistic synchronous tree-adjoining grammars for machine translation: The argument from bilingual dictionaries. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 88-95). Association for Computational Linguistics.

Shieber, S., & Kulesza, A. (2004). A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 75-84).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

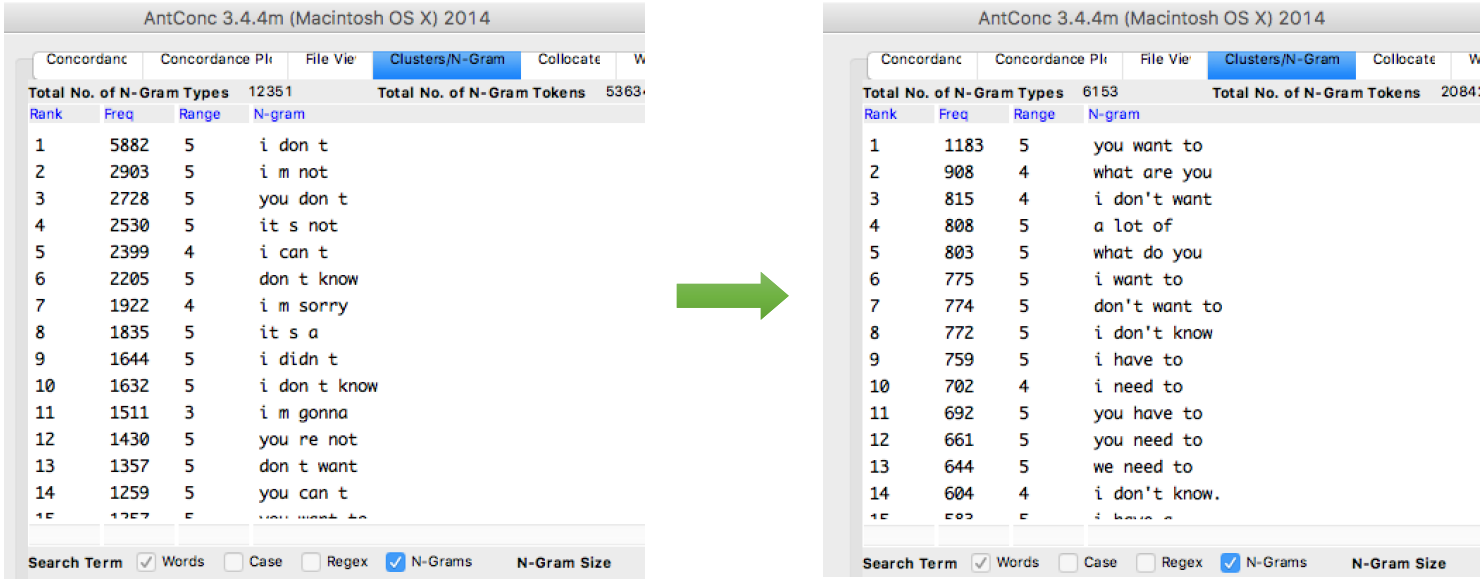
Tang, J., Wang, X., Gao, H., Hu, X., & Liu, H. (2012). Enriching short text representation in microblog for clustering. *Frontiers of Computer Science in China*, *6*(1), 88-101.

Tucker Jr, A. B. (1984). A perspective on machine translation: Theory and practice. *Communications of the ACM*, *27*(4), 322-329.

Appendices

Appendix A

Formatting for AntConc: Punctuation



Appendix B

Spoken Corpus: Full List of 3-gram Sequences

Spoken Corpus
3-gram MWUs

#Total No. of N-Gram Types: 5258
#Total No. of N-Gram Tokens: 186356

#	Freq.	Ran.	MWU
1	1183	5	you want to
2	908	4	what are you
3	815	4	i don't want
4	808	5	a lot of
5	803	5	what do you
6	775	5	i want to
7	774	5	don't want to
8	772	5	i don't know
9	759	5	i have to
10	702	4	i need to
11	692	5	you have to
12	661	5	you need to
13	644	5	we need to
14	604	4	i don't know.
15	583	5	i have a
16	531	5	you have a
17	473	3	do you think
18	468	3	need you to
19	452	5	this is a
20	446	5	we have to
21	421	5	i'm going to
22	420	5	i don't think
23	417	5	one of the
24	391	4	want me to
25	390	4	you don't have
26	389	4	going to be
27	384	5	do you want
28	381	5	be able to
29	360	3	you know what?
30	359	5	to be a
31	358	5	it was a

32	357	4	do you know
33	357	4	want to be
34	354	5	this is the
35	352	3	want you to
36	342	5	don't have to
37	341	4	i don't have
38	339	5	to talk to
39	335	4	you want me
40	324	5	you know what
41	321	5	out of the
42	313	5	the rest of
43	308	3	you don't want
44	293	3	don't know what
45	283	3	i'm not gonna
46	278	4	do you have
47	276	5	not going to
48	275	3	supposed to be
49	272	4	if you want
50	270	4	take care of
51	268	5	have to do
52	266	5	a couple of
53	266	5	and i don't
54	266	4	i thought you
55	265	4	i think i
56	264	3	are you doing
57	264	5	the fact that
58	263	5	need to get
59	256	5	how do you
60	255	5	this is not
61	252	3	i want you
62	249	4	get out of
63	249	4	i am not
64	249	5	it's not a
65	248	5	a little bit
66	248	5	we have a
67	244	4	i need a
68	242	5	if you don't
69	241	3	what did you
70	239	5	have to be
71	238	3	i know you
72	238	3	you know, i
73	236	5	i wanted to
74	235	5	you're going to
75	233	5	to talk about

76	226	4	i had to
77	226	5	you can do
78	220	4	i think it's
79	217	4	i'm trying to
80	217	4	want to go
81	216	4	as soon as
82	215	3	i just want
83	213	3	why do you
84	213	3	why don't you
85	212	4	you don't know
86	209	5	and this is
87	209	5	you can see
88	207	4	i was just
89	204	4	in front of
90	204	4	want to do
91	203	4	i had a
92	200	4	i can do
93	200	5	i told you
94	200	5	the end of
95	196	4	don't know how
96	196	5	to have a
97	195	4	you know how
98	194	4	go to the
99	194	5	need to be
100	193	3	give me a
101	190	5	there is a
102	188	5	i'd like to
103	188	3	just want to
104	187	4	and i have
105	187	4	is going to
106	187	5	look at the
107	187	4	to go to
108	185	5	there was a
109	185	4	when i was
110	184	3	out of my
111	182	3	you think i
112	181	4	we're going to
113	180	5	we don't have
114	180	4	what do we
115	179	3	want to know
116	178	5	to do with
117	177	5	to make sure
118	176	4	need to know
119	175	4	i think you

120	172	3	i know what
121	172	4	it could be
122	170	4	i have no
123	170	4	what's wrong with
124	169	4	i know that
125	168	4	go back to
126	164	3	didn't want to
127	164	5	know how to
128	162	3	i just need
129	162	5	part of the
130	161	4	and then you
131	161	4	there is no
132	160	4	for you to
133	159	5	is not a
134	159	4	the middle of
135	159	5	to be the
136	158	3	it's gonna be
137	156	3	i was gonna
138	155	4	and i think
139	155	5	i thought i
140	155	4	want to talk
141	154	5	have to go
142	154	5	i think we
143	154	4	this is my
144	154	3	you're not gonna
145	153	4	are going to
146	153	4	i was a
147	153	4	in the middle
148	153	5	it would be
149	153	3	just need to
150	152	4	to do a
151	151	4	and then i
152	151	5	don't have a
153	151	3	i got a
154	149	5	if you have
155	148	3	i don't care
156	148	3	what does that
157	147	4	and if you
158	147	4	what kind of
159	145	4	but i don't
160	144	4	and i know
161	144	4	back to the
162	144	4	to get a
163	143	4	have to get

164	142	3	just wanted to
165	142	4	you don't get
166	141	5	on the other
167	141	3	you got a
168	141	4	you have any
169	140	4	just trying to
170	139	4	i'm not going
171	139	4	want to see
172	137	5	has to be
173	137	3	thank you for
174	136	3	which is why
175	135	4	could be a
176	135	4	i know it's
177	135	4	i'm not sure
178	134	4	did you get
179	134	3	don't know. i
180	133	5	and you can
181	133	3	what happened to
182	133	4	you have no
183	132	3	how are you
184	132	5	some of the
185	131	5	so you can
186	131	4	to do it
187	131	5	we can do
188	130	4	in terms of
189	130	4	you and i
190	129	4	as long as
191	128	4	you should be
192	127	4	a bunch of
193	127	5	a look at
194	127	5	end of the
195	127	3	i don't see
196	127	3	tell me what
197	127	3	want to get
198	126	3	and i want
199	126	4	don't need to
200	126	5	need to do
201	126	5	to look at
202	126	5	to take a
203	126	3	you are not
204	125	5	it was the
205	125	4	talk to you
206	124	3	and i am
207	124	3	i just wanted

208	124	3	to tell me
209	124	3	you need a
210	123	5	it's just a
211	123	5	so if you
212	122	4	know what i
213	122	5	the only thing
214	121	4	i'm not a
215	121	3	so i can
216	120	5	that's what i
217	120	3	you think you
218	119	4	and i are
219	119	3	in love with
220	119	5	it is a
221	119	5	it looks like
222	118	4	at the end
223	118	4	she has a
224	118	4	to be in
225	118	4	to do the
226	118	3	why would you
227	118	3	you don't need
228	118	5	you in the
229	117	4	and i was
230	117	4	and i will
231	115	5	i think that
232	114	3	a chance to
233	114	4	and i'm not
234	114	5	know what to
235	114	4	there's no way
236	114	4	to tell you
237	113	3	i know you're
238	113	4	i've got a
239	113	4	that would be
240	112	4	are you talking
241	112	3	gonna be a
242	112	3	i'm gonna go
243	112	4	if we don't
244	112	3	out of your
245	112	4	you get a
246	111	4	but i think
247	111	5	this is what
248	111	3	you think i'm
249	110	5	we don't know
250	109	4	because of the
251	109	3	don't get to

252	109	3	give me the
253	109	5	just have to
254	109	5	take a look
255	108	3	and i can't
256	108	3	don't even know
257	108	4	don't know if
258	108	3	he has a
259	108	4	i know i
260	108	3	was supposed to
261	108	3	you can get
262	108	3	you wanted to
263	107	3	the one who
264	107	4	trying to get
265	106	4	but if you
266	106	3	you are a
267	106	3	you to be
268	105	3	do you need
269	105	4	get back to
270	105	4	have to tell
271	105	3	i can get
272	105	5	there was no
273	105	5	to make a
274	104	4	got to be
275	104	4	if you need
276	104	3	need to talk
277	104	3	what are we
278	103	4	and you know
279	103	5	make sure that
280	103	4	that's why i
281	102	5	give you a
282	102	3	he's got a
283	102	3	you've got a
284	101	4	some kind of
285	101	5	that is not
286	101	4	that you can
287	100	3	are you sure
288	99	4	have no idea
289	99	3	i don't like
290	99	3	i think you're
291	99	5	is that a
292	99	3	to do it.
293	99	5	want to make
294	99	3	you tell me
295	98	3	i'm supposed to

296	98	5	that was a
297	97	5	if you were
298	97	4	thought it was
299	97	4	to get to
300	97	4	you get to
301	97	4	you had to
302	96	4	going to do
303	96	5	so this is
304	96	5	the kind of
305	96	4	up to the
306	96	3	was trying to
307	96	4	you know that
308	95	3	i went to
309	95	4	it's not like
310	95	5	need to go
311	95	4	the two of
312	95	5	would like to
313	94	3	don't think i
314	94	3	he had a
315	94	3	i was in
316	94	5	in order to
317	94	5	one of those
318	94	5	the first time
319	94	3	was in the
320	93	3	don't want you
321	93	4	for me to
322	93	4	i have been
323	93	3	i used to
324	93	4	if it was
325	93	4	if you can
326	93	4	if you could
327	93	3	is gonna be
328	93	5	to get the
329	93	5	up in the
330	93	3	you going to
331	92	4	get rid of
332	92	4	going to have
333	92	5	how do we
334	92	4	just a little
335	92	4	what do i
336	91	4	and it's not
337	91	3	this is your
338	91	5	to do is
339	91	4	to see if

340	90	4	i know this
341	90	5	i think i'm
342	90	5	it in the
343	90	3	one of you
344	90	4	so we can
345	90	3	to figure out
346	90	4	you were a
347	89	4	a piece of
348	89	3	but i can't
349	89	4	how did you
350	89	3	i can't get
351	89	4	the only way
352	89	5	we want to
353	89	5	what's going on
354	88	3	and you don't
355	88	3	i made a
356	88	4	it's going to
357	88	3	she had a
358	88	4	the only one
359	88	4	used to be
360	88	3	what am i
361	88	3	you think that
362	87	5	i think the
363	87	3	it's not the
364	86	4	a little more
365	86	4	and you have
366	86	3	but i'm not
367	86	4	for the rest
368	86	5	that is a
369	86	3	what about the
370	86	4	you can't do
371	85	5	a bit of
372	85	4	don't have time
373	85	4	get to the
374	85	3	i'll be right
375	85	4	it was just
376	85	5	to you about
377	85	4	you get the
378	85	4	you had a
379	84	5	come up with
380	84	3	he wants to
381	84	4	i wish i
382	84	4	in the first
383	84	5	needs to be

384	84	4	to know what
385	84	3	want to have
386	84	3	you out of
387	83	5	be in the
388	83	5	i would like
389	83	5	is that the
390	83	3	thank you. thank
391	83	4	was going to
392	83	4	you know the
393	82	3	he doesn't have
394	82	3	i thought it
395	82	4	we need a
396	81	3	are you going
397	81	4	i think it
398	81	3	i'm just gonna
399	81	5	that you were
400	81	4	the last time
401	81	5	to go back
402	81	3	you to take
403	81	3	you were in
404	81	4	you're trying to
405	80	3	do i have
406	80	4	i can see
407	80	3	i have the
408	80	3	might as well
409	80	4	that's not what
410	80	3	want to take
411	80	4	we can get
412	79	4	a problem with
413	79	5	all the way
414	79	4	i do not
415	79	3	i will be
416	79	3	i'm not the
417	79	4	if you think
418	79	5	in a few
419	79	4	is there a
420	79	4	it's hard to
421	79	4	that is the
422	79	4	what i was
423	78	4	a way to
424	78	4	and it was
425	78	3	i should have
426	78	5	most of the
427	78	3	that's what you

428	78	5	to do this
429	77	3	and you are
430	77	3	is not the
431	77	4	that's a good
432	77	5	to give you
433	77	4	to make it
434	77	3	to see you
435	77	3	to take the
436	77	5	you have the
437	76	4	if i had
438	76	3	said it was
439	76	5	that kind of
440	76	3	you talking about?
441	76	3	you think it's
442	75	4	and i just
443	75	4	asked me to
444	75	3	but this is
445	75	4	but you don't
446	75	4	going to get
447	75	3	have you been
448	75	5	it has to
449	75	4	need to take
450	75	3	of course i
451	75	4	rest of the
452	75	4	some sort of
453	75	4	that we can
454	75	3	to get her
455	75	4	you can tell
456	74	5	a very good
457	74	4	going to the
458	74	4	i would have
459	74	3	it is not
460	74	4	tell me about
461	74	5	tell you that
462	74	5	think this is
463	74	4	this is why
464	74	3	to deal with
465	74	3	what if i
466	73	5	all over the
467	73	3	at the moment
468	73	4	give it to
469	73	3	going to talk
470	73	5	if you had
471	73	3	is wrong with

472	73	5	it might be
473	73	4	it's a good
474	73	4	middle of the
475	73	3	that i was
476	73	4	think we should
477	73	3	what did i
478	73	3	you so much.
479	72	3	he was a
480	72	3	i tried to
481	72	3	if it's not
482	72	5	is one of
483	72	4	really want to
484	72	5	to find a
485	72	4	to have to
486	72	4	you can go
487	72	3	you talk to
488	71	5	a long time
489	71	4	i know the
490	71	3	is a good
491	71	5	little bit of
492	71	3	me in the
493	71	4	nothing to do
494	71	3	you can't just
495	71	5	you look at
496	71	3	you. thank you.
497	70	3	and i'm gonna
498	70	4	don't know why
499	70	4	have time to
500	70	3	i know it
501	70	4	looks like a
502	70	5	now i have
503	70	3	talk to the
504	70	5	that you have
505	70	5	there are no
506	70	3	you to do
507	69	4	a few more
508	69	3	a long time.
509	69	5	have to take
510	69	3	i should be
511	69	3	i told her
512	69	3	is that what
513	69	3	know this is
514	69	4	not a good
515	69	4	thank you so

516	69	3	that's why you
517	69	3	you can't be
518	68	4	but it's not
519	68	3	got to get
520	68	5	is a very
521	68	5	out of a
522	68	5	side of the
523	68	5	up on the
524	68	3	you guys are
525	67	4	and there are
526	67	5	as far as
527	67	3	if you've got
528	67	5	it's not just
529	67	4	not in the
530	67	3	to get you
531	66	3	i guess i
532	66	3	i thought we
533	66	4	know what it
534	66	3	not supposed to
535	66	4	so what do
536	66	5	that was the
537	66	4	think you can
538	66	3	to do that.
539	66	4	up with a
540	66	3	when you get
541	66	4	when you were
542	66	4	would you like
543	66	5	you're not going
544	65	3	and if i
545	65	3	but i have
546	65	5	have a lot
547	65	3	have you ever
548	65	4	i didn't think
549	65	3	i got the
550	65	3	it was an
551	65	4	it's been a
552	65	4	kind of a
553	65	4	no sign of
554	65	4	one of your
555	65	3	to get back
556	65	5	to see the
557	64	5	a number of
558	64	5	all of the
559	64	4	at the same

560	64	3	can i have
561	64	3	for a few
562	64	3	give her a
563	64	3	i get a
564	64	5	in the same
565	64	3	is why i
566	64	5	it's the only
567	64	5	on top of
568	64	5	this is an
569	64	4	to know that
570	64	4	to try and
571	64	5	would be a
572	64	4	you know i
573	63	3	and i can
574	63	4	and you will
575	63	4	i like to
576	63	5	i thought that
577	63	4	if i can
578	63	4	it's not my
579	63	5	out in the
580	63	3	that's not a
581	63	3	the way you
582	63	3	want to tell
583	63	4	you are the
584	63	4	you can have
585	63	5	you could have
586	62	3	as much as
587	62	3	how to do
588	62	3	i can't tell
589	62	4	i think that's
590	62	4	i was going
591	62	4	if you do
592	62	4	it to the
593	62	4	looking for a
594	62	4	one of my
595	62	5	to be able
596	62	5	to find out
597	62	5	to try to
598	62	4	you have an
599	62	3	you know who
600	61	4	by the time
601	61	4	don't have any
602	61	3	don't worry about
603	61	4	i didn't have

604	61	3	i think he
605	61	5	that i have
606	61	3	told me to
607	61	4	why are we
608	61	3	you really want
609	60	4	how can you
610	60	4	i hope you
611	60	3	i think she
612	60	4	i'm pretty sure
613	60	4	if you're not
614	60	4	that's not the
615	60	4	this is just
616	60	4	to get out
617	60	3	two of you
618	60	5	we had a
619	60	3	you should have
620	60	3	you've got to
621	59	3	am i gonna
622	59	4	and that is
623	59	3	can you get
624	59	3	going on with
625	59	4	have to say
626	59	4	i didn't do
627	59	4	i talk to
628	59	3	i want a
629	59	5	i'm not saying
630	59	4	i'm talking about
631	59	4	in the last
632	59	5	is in the
633	59	4	it's not that
634	59	3	need to see
635	59	4	no idea what
636	59	3	not gonna do
637	59	3	one of us
638	59	3	put in a
639	59	4	so you don't
640	59	4	they want to
641	59	4	trying to make
642	59	3	we got to
643	59	4	what are the
644	58	3	even if you
645	58	3	have no idea.
646	58	3	might want to
647	58	3	of course you

648	58	5	the back of
649	58	5	the way that
650	58	4	we've got a
651	58	5	what is the
652	58	3	would have been
653	57	4	and he was
654	57	5	and then we
655	57	5	as well as
656	57	4	back on the
657	57	4	don't have the
658	57	3	down to the
659	57	4	going to take
660	57	5	i'm about to
661	57	4	it can be
662	57	4	it to be
663	57	3	she's got a
664	57	5	talk about the
665	57	4	the one that
666	57	3	the only reason
667	57	5	they have a
668	57	3	this is about
669	57	3	to get in
670	57	3	we're not gonna
671	56	3	are you all
672	56	4	be a good
673	56	4	be a little
674	56	3	because i don't
675	56	4	by the way,
676	56	3	get her to
677	56	3	going to need
678	56	5	how do i
679	56	4	if we can
680	56	3	just let me
681	56	4	not talking about
682	56	3	put him on
683	56	5	that i can
684	56	4	that i don't
685	56	4	that this is
686	56	4	there are a
687	56	4	this is all
688	56	3	we got a
689	56	4	we had to
690	55	4	and then the
691	55	4	away from the

692	55	4	can see that
693	55	3	can you do
694	55	4	had to be
695	55	5	i was thinking
696	55	4	if i have
697	55	4	in there and
698	55	4	is there any
699	55	3	it was my
700	55	3	know that you
701	55	3	no matter how
702	55	3	that i am
703	55	4	that you don't
704	55	3	you might want
705	55	3	you're not a
706	54	4	a lot to
707	54	3	are you still
708	54	3	back in the
709	54	3	did you hear
710	54	3	do you really
711	54	5	have to have
712	54	3	he was in
713	54	4	it's kind of
714	54	3	it's not an
715	54	3	me to be
716	54	3	me to do
717	54	3	out of here
718	54	4	out of his
719	54	3	soon as i
720	54	4	there's a lot
721	54	4	to do an
722	54	3	to give me
723	54	3	to know if
724	54	4	to make the
725	54	3	trying to save
726	54	3	was a good
727	54	5	we're talking about
728	54	4	what would you
729	54	3	why don't we
730	54	3	you can be
731	54	3	you can take
732	54	3	you don't like
733	54	3	you must be
734	54	3	you think this
735	53	4	all the time

736	53	4	and it is
737	53	3	blood in the
738	53	4	going to go
739	53	3	him out of
740	53	3	him to the
741	53	3	i could do
742	53	4	i'll give you
743	53	5	just a few
744	53	4	make sure you
745	53	4	need to make
746	53	5	out of this
747	53	5	that if you
748	53	5	the way to
749	53	5	think it's a
750	53	4	this isn't a
751	53	5	to do that
752	53	3	was just a
753	53	4	we are not
754	53	3	went to the
755	53	5	were able to
756	53	3	where do you
757	53	3	you know where
758	53	3	you want a
759	53	3	you will be
760	52	3	by the way.
761	52	3	can be a
762	52	5	for the first
763	52	3	for the next
764	52	3	have a good
765	52	3	i'm sure you
766	52	4	just going to
767	52	3	know what i'm
768	52	5	may not be
769	52	3	she's going to
770	52	3	the most important
771	52	3	the patient is
772	52	4	the size of
773	52	4	there are some
774	52	4	to remove the
775	52	3	won't be able
776	51	3	can i help
777	51	4	he's going to
778	51	3	i tell you
779	51	3	if you get

780	51	3	just tell me
781	51	4	need to tell
782	51	5	talking about the
783	51	4	there are other
784	51	3	to take it
785	51	4	to think about
786	51	4	we should be
787	51	4	which is a
788	51	3	you for a
789	51	3	you know, i'm
790	51	5	you think about
791	50	3	ask you to
792	50	3	but it is
793	50	3	do we have
794	50	3	don't care about
795	50	4	don't think it's
796	50	3	how am i
797	50	3	i was the
798	50	3	i'm a little
799	50	5	in the world
800	50	3	in the world.
801	50	4	it doesn't matter
802	50	4	know how much
803	50	5	they have to
804	50	5	to do something
805	50	5	to say that
806	50	3	you to tell
807	49	4	and i thought
808	49	3	as you can
809	49	4	come back to
810	49	3	i've never seen
811	49	4	is that you
812	49	4	it out of
813	49	5	of the most
814	49	5	one of these
815	49	3	said you were
816	49	4	she was in
817	49	5	thank you very
818	49	3	the immune system
819	49	3	told you i
820	49	4	top of the
821	49	4	try not to
822	49	4	which is the
823	49	3	you give me

824	49	5	you go to
825	49	3	you like to
826	48	3	and i got
827	48	4	because if you
828	48	4	but there are
829	48	3	but you can't
830	48	3	do you do
831	48	3	don't do that.
832	48	3	don't think you
833	48	3	figure out how
834	48	5	for a long
835	48	4	i am trying
836	48	4	i'm sorry i
837	48	5	is on the
838	48	3	is there anything
839	48	4	it's just that
840	48	4	so it's not
841	48	4	the top of
842	48	4	there will be
843	48	5	these are the
844	48	5	this kind of
845	48	3	thought we were
846	48	3	to ask you
847	48	3	we have no
848	48	3	we're gonna do
849	48	4	when she was
850	48	3	you do not
851	48	3	you think the
852	48	3	you're a good
853	47	4	a lot more
854	47	5	a part of
855	47	3	but you have
856	47	3	don't know what's
857	47	3	don't need a
858	47	3	give us a
859	47	3	have to make
860	47	3	hell of a
861	47	4	how can i
862	47	3	i asked you
863	47	3	i would be
864	47	3	i'm sorry about
865	47	4	if this is
866	47	4	is the best
867	47	3	look like a

868	47	4	only way to
869	47	4	rest of your
870	47	3	she was a
871	47	3	taking care of
872	47	4	then you can
873	47	3	want to give
874	47	3	we get to
875	47	5	we're trying to
876	47	5	whether or not
877	47	4	you are going
878	47	3	you in a
879	47	4	you said that
880	47	4	you want the
881	46	3	asked you to
882	46	4	can we get
883	46	3	do what you
884	46	3	get away from
885	46	4	get to know
886	46	5	have a very
887	46	5	he has to
888	46	3	i could have
889	46	3	i don't mean
890	46	3	in a couple
891	46	5	is the only
892	46	5	it's a very
893	46	3	let me get
894	46	4	like you to
895	46	5	look at this
896	46	5	looking at the
897	46	5	might be a
898	46	3	need me to
899	46	4	of all the
900	46	4	so i don't
901	46	4	that it was
902	46	5	they're going to
903	46	3	to be on
904	46	3	to know the
905	46	4	up in a
906	46	4	what i want
907	46	4	you know you
908	46	3	you think it
909	45	4	am trying to
910	45	3	but i was
911	45	4	did you say

912	45	4	do we do
913	45	3	i kind of
914	45	4	i think this
915	45	4	i'm looking for
916	45	4	if it's a
917	45	4	if there's a
918	45	5	in the next
919	45	5	is what i
920	45	3	know how many
921	45	5	really need to
922	45	3	she doesn't have
923	45	5	the first thing
924	45	3	the good news
925	45	3	the same thing.
926	45	3	there it is.
927	45	4	to have the
928	45	3	trying to figure
929	45	3	we should do
930	45	4	we've got to
931	45	3	what does it
932	45	3	what is wrong
933	45	3	you could just
934	45	4	you do the
935	45	3	you not to
936	44	4	but you can
937	44	3	but you know
938	44	3	do you see
939	44	3	doesn't have a
940	44	3	get in there
941	44	3	give him a
942	44	4	have to give
943	44	3	i've got to
944	44	3	it wasn't a
945	44	4	it's a little
946	44	3	it's like a
947	44	3	know how you
948	44	3	know that i
949	44	3	on my way
950	44	3	one of them
951	44	5	seem to be
952	44	3	she has to
953	44	3	sure you don't
954	44	3	think i'm gonna
955	44	5	to show you

956	44	4	to worry about
957	44	3	want to say
958	44	3	what if we
959	44	3	you didn't have
960	44	4	you just have
961	43	4	and i would
962	43	4	can get a
963	43	3	can you just
964	43	4	can't do this
965	43	3	get in the
966	43	3	give you the
967	43	5	have the same
968	43	3	her to the
969	43	3	i didn't get
970	43	3	i thought you'd
971	43	4	i'll tell you
972	43	4	i'm just saying
973	43	3	i'm not talking
974	43	3	if i was
975	43	3	is why you
976	43	4	need to find
977	43	4	no matter what
978	43	4	only thing that
979	43	4	or you can
980	43	3	out how to
981	43	5	that we have
982	43	3	the sort of
983	43	3	the time you
984	43	3	there are only
985	43	4	to get it
986	43	5	to go into
987	43	4	wish i could
988	43	3	you might be
989	42	4	and in the
990	42	3	and you get
991	42	3	as good as
992	42	4	because i was
993	42	3	does that mean
994	42	4	figure out what
995	42	4	i gave you
996	42	3	i really don't
997	42	4	i was on
998	42	4	is the one
999	42	3	it's a big

1000	42	3	know what the
1001	42	3	like to see
1002	42	5	this is where
1003	42	3	to let you
1004	42	3	to work with
1005	42	4	we have the
1006	42	4	where are we
1007	42	3	would you do
1008	42	4	you could do
1009	41	5	able to do
1010	41	3	and then we'll
1011	41	4	and we can
1012	41	3	and you can't
1013	41	3	but i want
1014	41	4	but that's not
1015	41	4	doesn't have to
1016	41	3	front of the
1017	41	4	going to tell
1018	41	3	hit by a
1019	41	3	i just can't
1020	41	3	i will not
1021	41	4	it doesn't mean
1022	41	3	it on the
1023	41	5	it turns out
1024	41	4	it will be
1025	41	4	know if you
1026	41	3	know what that
1027	41	3	let me see
1028	41	5	no way to
1029	41	5	of us are
1030	41	4	out to be
1031	41	4	seems to be
1032	41	5	should be able
1033	41	5	tell you about
1034	41	3	that you are
1035	41	3	the way i
1036	41	3	to get some
1037	41	3	to get your
1038	41	4	to make this
1039	41	4	to pick up
1040	41	3	to take you
1041	41	4	we'll have to
1042	41	4	what you can
1043	41	3	why is it

1044	41	4	you know about
1045	40	3	a good thing.
1046	40	3	a really good
1047	40	4	and the other
1048	40	5	and then they
1049	40	3	are you a
1050	40	4	been trying to
1051	40	3	can i ask
1052	40	5	could have a
1053	40	3	do it in
1054	40	3	fact that you
1055	40	3	find out what
1056	40	4	for us to
1057	40	4	have a problem
1058	40	4	have been a
1059	40	3	i thought i'd
1060	40	5	if we could
1061	40	3	in a lot
1062	40	4	it is the
1063	40	4	know anything about
1064	40	3	let you know
1065	40	4	like to do
1066	40	3	me to tell
1067	40	5	more than a
1068	40	4	not just a
1069	40	3	of your life.
1070	40	5	on to the
1071	40	3	she has no
1072	40	3	so that you
1073	40	3	take you to
1074	40	5	the ability to
1075	40	4	the last thing
1076	40	5	to go out
1077	40	3	to stop the
1078	40	3	were in the
1079	40	5	what happens when
1080	40	4	when we get
1081	40	5	when we were
1082	40	3	you can make
1083	40	3	you got the
1084	40	3	you to know
1085	39	4	and i were
1086	39	4	and see if
1087	39	3	and then he

1088	39	4	and then she
1089	39	4	and you need
1090	39	3	and you're not
1091	39	4	any of the
1092	39	3	because you don't
1093	39	3	but i do
1094	39	3	but it was
1095	39	4	by the end
1096	39	4	can see the
1097	39	4	didn't have to
1098	39	3	do what i
1099	39	3	do you get
1100	39	4	have to ask
1101	39	3	i had the
1102	39	3	i just thought
1103	39	4	i might be
1104	39	4	i was wondering
1105	39	3	if you like
1106	39	4	it's in the
1107	39	4	like to be
1108	39	4	put it in
1109	39	4	so i think
1110	39	3	so i'm gonna
1111	39	3	thanks for the
1112	39	5	that there are
1113	39	3	that's what i'm
1114	39	3	then why are
1115	39	5	to be an
1116	39	4	to find the
1117	39	4	to go through
1118	39	3	to know how
1119	39	3	to listen to
1120	39	5	told me that
1121	39	5	trying to find
1122	39	4	was on the
1123	39	3	we are going
1124	39	4	we were able
1125	39	3	what i did
1126	39	4	what to do
1127	39	3	what you want
1128	39	5	when it comes
1129	39	5	you know if
1130	39	4	you see the
1131	39	3	you're in the

1132	38	3	a patient with
1133	38	4	all sorts of
1134	38	3	and in fact
1135	38	3	but it doesn't
1136	38	3	could be the
1137	38	3	do it for
1138	38	3	do you mean
1139	38	3	don't you just
1140	38	3	get a little
1141	38	3	have something to
1142	38	4	have to find
1143	38	3	he wanted to
1144	38	3	her in the
1145	38	3	him on the
1146	38	3	how could you
1147	38	5	i don't really
1148	38	3	i know i'm
1149	38	3	i take it
1150	38	4	it may be
1151	38	4	it should be
1152	38	4	it would have
1153	38	3	just to be
1154	38	4	just try to
1155	38	3	know what it's
1156	38	3	might not be
1157	38	4	more likely to
1158	38	3	not trying to
1159	38	4	of course it
1160	38	3	pick up the
1161	38	4	so i just
1162	38	4	so we have
1163	38	3	tell me how
1164	38	3	that what you
1165	38	4	the amount of
1166	38	4	think of it
1167	38	3	this is going
1168	38	4	to say to
1169	38	3	up with the
1170	38	4	was in a
1171	38	3	was the last
1172	38	4	when i say
1173	38	3	when you have
1174	38	3	you know what's
1175	38	4	you take a

1176	38	3	you'd like to
1177	37	3	a good thing
1178	37	4	and all the
1179	37	3	and tell me
1180	37	4	and there's a
1181	37	3	and there's no
1182	37	3	and you were
1183	37	3	be on the
1184	37	4	because this is
1185	37	5	bit of a
1186	37	4	can do the
1187	37	3	do this to
1188	37	3	do you like
1189	37	3	do you mind
1190	37	4	don't know about
1191	37	3	due to the
1192	37	3	give it a
1193	37	3	i ask you
1194	37	3	i can't find
1195	37	3	i needed to
1196	37	4	if you are
1197	37	4	is it a
1198	37	3	is what you
1199	37	3	it can't be
1200	37	3	just need a
1201	37	4	lot of people
1202	37	3	she told me
1203	37	3	so what are
1204	37	3	so you have
1205	37	3	talk about it
1206	37	3	that doesn't make
1207	37	3	that i would
1208	37	3	that's all i
1209	37	5	there are two
1210	37	4	they need to
1211	37	3	think it was
1212	37	3	to have you
1213	37	3	to make you
1214	37	3	want to spend
1215	37	3	we should have
1216	37	3	were supposed to
1217	37	4	which means that
1218	37	3	you can say
1219	37	3	you look like

1220	37	4	you should get
1221	37	4	you'll have to
1222	36	3	a hole in
1223	36	4	and it's a
1224	36	4	and she was
1225	36	4	and we have
1226	36	3	and we're gonna
1227	36	5	are in the
1228	36	3	but if i
1229	36	5	do with the
1230	36	3	don't care if
1231	36	5	for the last
1232	36	5	go into the
1233	36	3	going back to
1234	36	3	i can make
1235	36	3	i did it
1236	36	3	if you go
1237	36	3	in here and
1238	36	4	in the blood
1239	36	5	in the right
1240	36	4	is a little
1241	36	3	it doesn't make
1242	36	3	it was like
1243	36	3	it's one of
1244	36	3	know that you're
1245	36	3	me to get
1246	36	3	me to take
1247	36	5	on the way
1248	36	3	sounds like a
1249	36	4	tell you what
1250	36	5	that there is
1251	36	3	that you're not
1252	36	3	the head of
1253	36	3	the last two
1254	36	5	the same thing
1255	36	5	the things that
1256	36	3	the way it
1257	36	4	think that you
1258	36	4	this is one
1259	36	3	this is so
1260	36	5	to the next
1261	36	3	to worry about.
1262	36	3	want to keep
1263	36	3	what i can

1264	36	3	which one of
1265	36	3	who are you
1266	36	3	who wants to
1267	35	3	a list of
1268	35	3	and there is
1269	35	4	and there was
1270	35	3	and we are
1271	35	3	and you should
1272	35	5	been able to
1273	35	4	but it's a
1274	35	3	but there is
1275	35	4	can i just
1276	35	3	don't think that
1277	35	3	even know what
1278	35	3	feel like a
1279	35	4	for the past
1280	35	4	got a lot
1281	35	4	got to do
1282	35	3	have to work
1283	35	3	he was just
1284	35	4	here in the
1285	35	4	i can tell
1286	35	3	i could use
1287	35	3	if i'm gonna
1288	35	4	if it is
1289	35	5	it comes to
1290	35	4	it is to
1291	35	4	job is to
1292	35	4	look at your
1293	35	4	more of a
1294	35	5	not to be
1295	35	4	or are you
1296	35	3	or do you
1297	35	3	put it on
1298	35	3	put on a
1299	35	3	rest of his
1300	35	5	that in the
1301	35	4	that you would
1302	35	4	that's why we
1303	35	5	the number of
1304	35	3	the one who's
1305	35	4	the risk of
1306	35	3	there's only one
1307	35	4	think we can

1308	35	3	to do what
1309	35	3	to have an
1310	35	5	was able to
1311	35	5	we could do
1312	35	4	we have an
1313	35	3	what did she
1314	35	3	you just need
1315	35	4	you know when
1316	35	3	you're about to
1317	35	3	you, but i
1318	34	4	a matter of
1319	34	4	a series of
1320	34	4	and i'm going
1321	34	3	and talk to
1322	34	3	and the only
1323	34	3	and you want
1324	34	4	at this point,
1325	34	3	could have been
1326	34	5	have to wait
1327	34	3	i did a
1328	34	4	i was doing
1329	34	4	if there was
1330	34	3	is a big
1331	34	5	is just a
1332	34	3	is that why
1333	34	3	is this a
1334	34	5	it in a
1335	34	3	it is. i
1336	34	5	it won't be
1337	34	4	no one is
1338	34	3	not one of
1339	34	5	on the right
1340	34	3	out of her
1341	34	4	out on the
1342	34	3	she wanted to
1343	34	3	that you need
1344	34	5	the answer is
1345	34	4	the case of
1346	34	5	the difference between
1347	34	5	the first one
1348	34	3	there's no reason
1349	34	5	this was a
1350	34	3	to come to
1351	34	3	to get this

1352	34	3	to keep the
1353	34	5	to know about
1354	34	3	was a bad
1355	34	5	we talk about
1356	34	5	we talked about
1357	34	4	we were in
1358	34	3	were going to
1359	34	4	would be the
1360	34	3	yeah, i think
1361	34	3	you do a
1362	34	3	you do to
1363	34	4	you must have
1364	34	4	you take the
1365	34	3	you to the
1366	34	3	you, and i
1367	33	5	a picture of
1368	33	3	a waste of
1369	33	4	and make sure
1370	33	4	and that's the
1371	33	3	and then i'm
1372	33	3	because i have
1373	33	4	because you have
1374	33	4	damage to the
1375	33	4	didn't have a
1376	33	3	do you remember
1377	33	3	don't want me
1378	33	4	get used to
1379	33	3	give me some
1380	33	5	going on in
1381	33	3	he is a
1382	33	4	i believe that
1383	33	3	i don't understand
1384	33	3	i will do
1385	33	3	i wouldn't be
1386	33	3	i'd like you
1387	33	4	i'm looking at
1388	33	3	if she doesn't
1389	33	5	if you look
1390	33	4	in the way
1391	33	4	is the first
1392	33	4	is the last
1393	33	3	is this the
1394	33	3	it seems like
1395	33	4	long as you

1396	33	3	middle of a
1397	33	4	none of the
1398	33	3	now this is
1399	33	4	of course the
1400	33	5	of the best
1401	33	4	of you will
1402	33	3	pain in the
1403	33	3	see if you
1404	33	5	so there are
1405	33	3	soon as we
1406	33	3	that doesn't mean
1407	33	3	the one with
1408	33	3	the right thing
1409	33	5	they don't have
1410	33	4	to be here
1411	33	3	to get into
1412	33	5	to look for
1413	33	3	wanted to say
1414	33	5	way to get
1415	33	3	what it's like
1416	33	4	what was the
1417	33	3	when he was
1418	33	3	who do you
1419	33	3	you could be
1420	33	3	you if you
1421	33	3	you say to
1422	33	3	you should know
1423	33	3	you were going
1424	33	3	you were just
1425	33	3	you will not
1426	32	4	and get a
1427	32	3	and now he's
1428	32	4	and when you
1429	32	4	because he was
1430	32	4	can do for
1431	32	3	gonna have a
1432	32	3	have a little
1433	32	3	how would you
1434	32	3	i do know
1435	32	3	i go to
1436	32	3	i just have
1437	32	3	i never thought
1438	32	3	i would love
1439	32	3	i'm glad you

1440	32	4	if i were
1441	32	5	is not an
1442	32	3	it must be
1443	32	3	it's supposed to
1444	32	3	know it's not
1445	32	3	let's take a
1446	32	3	look at his
1447	32	4	may be a
1448	32	3	me to go
1449	32	3	no signs of
1450	32	3	not gonna get
1451	32	4	of you to
1452	32	3	ready for the
1453	32	4	said that you
1454	32	4	so if we
1455	32	3	soon as you
1456	32	3	that i know
1457	32	4	the face of
1458	32	5	the one thing
1459	32	4	the other side
1460	32	3	the side of
1461	32	3	they don't know
1462	32	4	to do to
1463	32	3	to do your
1464	32	3	to see me
1465	32	4	we just have
1466	32	4	we went to
1467	32	4	were trying to
1468	32	3	what you're talking
1469	32	3	which means you
1470	32	3	with you and
1471	32	4	you can give
1472	32	3	you did the
1473	32	3	you shouldn't be
1474	32	3	you think that's
1475	32	3	you think we
1476	32	3	you're not the
1477	31	3	a bit more
1478	31	3	a man who
1479	31	5	a patient in
1480	31	3	a reason to
1481	31	3	all i can
1482	31	5	all of these
1483	31	4	all the other

1484	31	4	are a lot
1485	31	3	at least a
1486	31	3	be in a
1487	31	5	being able to
1488	31	3	but if you're
1489	31	4	came up with
1490	31	3	can't do it
1491	31	3	come out of
1492	31	3	could be an
1493	31	3	down in the
1494	31	4	first of all,
1495	31	3	have to stop
1496	31	3	he was on
1497	31	3	he's in the
1498	31	3	i don't remember
1499	31	3	i look at
1500	31	4	i really need
1501	31	3	i was talking
1502	31	3	is a lot
1503	31	4	is going on
1504	31	3	it's all about
1505	31	3	last time i
1506	31	3	me if you
1507	31	4	might be able
1508	31	5	not the only
1509	31	4	of the things
1510	31	4	part of your
1511	31	3	people who are
1512	31	3	since i was
1513	31	3	so i have
1514	31	4	so much for
1515	31	3	tell me the
1516	31	4	that i'm not
1517	31	3	that you know
1518	31	3	the first time.
1519	31	3	the patient was
1520	31	5	the presence of
1521	31	3	there's a chance
1522	31	4	thing in the
1523	31	4	those of you
1524	31	3	to go on
1525	31	4	to keep it
1526	31	4	to run a
1527	31	3	to save the

1528	31	4	try to get
1529	31	3	wait for the
1530	31	4	what i'm talking
1531	31	5	what it is
1532	31	5	would have to
1533	31	4	you do it
1534	31	3	you don't do
1535	31	4	you have your
1536	31	3	you know it's
1537	31	3	you on the
1538	31	3	you think of
1539	30	3	a hundred and
1540	30	4	able to get
1541	30	3	and you think
1542	30	3	any of you
1543	30	4	at the top
1544	30	4	be a lot
1545	30	3	be happy to
1546	30	4	but we have
1547	30	3	but what if
1548	30	3	deal with the
1549	30	4	didn't know what
1550	30	3	didn't you tell
1551	30	3	don't see any
1552	30	4	find a way
1553	30	3	for you and
1554	30	4	get you to
1555	30	4	going to make
1556	30	3	got out of
1557	30	3	got to go
1558	30	3	had to do
1559	30	3	he could have
1560	30	4	i try to
1561	30	3	i was looking
1562	30	3	i won't be
1563	30	3	i'll let you
1564	30	4	i'm not an
1565	30	3	i'm not in
1566	30	4	if it were
1567	30	3	if we have
1568	30	3	if you feel
1569	30	3	in on the
1570	30	5	it as a
1571	30	4	it's important to

1572	30	4	know if i
1573	30	5	last time you
1574	30	3	like i was
1575	30	3	like to have
1576	30	3	me about the
1577	30	5	move on to
1578	30	3	not on the
1579	30	3	not that i
1580	30	4	one of our
1581	30	4	one or two
1582	30	3	so that i
1583	30	4	so you could
1584	30	3	that can't be
1585	30	3	that i didn't
1586	30	3	that is what
1587	30	3	that's a lot
1588	30	3	there was nothing
1589	30	3	to go with
1590	30	3	to take her
1591	30	5	to think of
1592	30	3	trying to kill
1593	30	3	want to come
1594	30	3	we do it
1595	30	4	what if you
1596	30	3	what makes you
1597	30	3	what you do
1598	30	3	where have you
1599	30	3	why am i
1600	30	4	you can't see
1601	30	4	you might have
1602	30	5	you see a
1603	30	3	you won't be
1604	30	3	you're talking about
1605	29	4	a bone marrow
1606	29	3	a history of
1607	29	3	all about the
1608	29	4	and i do
1609	29	3	and so on
1610	29	3	and that's what
1611	29	3	and then i'll
1612	29	3	are gonna be
1613	29	3	be the best
1614	29	3	but i will
1615	29	4	can't do anything

1616	29	5	exactly the same
1617	29	3	give him the
1618	29	3	go home and
1619	29	3	have to keep
1620	29	3	have to see
1621	29	4	have to worry
1622	29	5	how can we
1623	29	3	how many times
1624	29	3	how much i
1625	29	3	i had an
1626	29	3	i knew that
1627	29	3	i would never
1628	29	4	i would say
1629	29	3	if i'm not
1630	29	3	in the hospital
1631	29	3	in the waiting
1632	29	4	in this case
1633	29	4	is kind of
1634	29	4	is supposed to
1635	29	5	is that we
1636	29	3	it's not going
1637	29	4	know what this
1638	29	5	lot of the
1639	29	3	me in a
1640	29	3	me to the
1641	29	4	of them are
1642	29	3	one with the
1643	29	3	part of my
1644	29	3	right thing to
1645	29	3	she was just
1646	29	3	so i'm just
1647	29	4	talk about this
1648	29	4	that he was
1649	29	3	that i can't
1650	29	3	that she was
1651	29	5	that sort of
1652	29	4	that you could
1653	29	3	that's not true.
1654	29	4	the next few
1655	29	4	the thing that
1656	29	5	there are many
1657	29	3	think i was
1658	29	4	to help you
1659	29	4	to see her

1660	29	4	trying to do
1661	29	4	wanted to do
1662	29	3	wanted to make
1663	29	3	we get a
1664	29	3	why do i
1665	29	3	you can't have
1666	29	4	you have been
1667	29	3	you tried to
1668	29	5	you're looking at
1669	28	3	a hell of
1670	28	5	all kinds of
1671	28	4	all of you
1672	28	4	and a half
1673	28	3	and if we
1674	28	3	and look at
1675	28	4	and she has
1676	28	4	and that's why
1677	28	3	and these are
1678	28	4	back into the
1679	28	5	based on the
1680	28	3	but there's no
1681	28	3	can you be
1682	28	3	can you give
1683	28	4	every time you
1684	28	3	get to do
1685	28	4	going to say
1686	28	5	have to know
1687	28	4	have to move
1688	28	3	have to put
1689	28	3	i am in
1690	28	3	i guess you
1691	28	3	i'm thinking about
1692	28	3	i've never done
1693	28	3	if he was
1694	28	3	if i do
1695	28	3	if there is
1696	28	3	if there's any
1697	28	4	if you know
1698	28	3	if you really
1699	28	3	it's good to
1700	28	3	it's not his
1701	28	3	it's not what
1702	28	4	know what's going
1703	28	3	know who i

1704	28	4	little bit more
1705	28	3	need to figure
1706	28	3	need to start
1707	28	3	need to stop
1708	28	3	or is it
1709	28	3	or we could
1710	28	3	out of that
1711	28	4	people in the
1712	28	4	rid of the
1713	28	3	right in front
1714	28	3	she was on
1715	28	4	so how do
1716	28	3	so i'm not
1717	28	3	so what's the
1718	28	3	so you know
1719	28	4	so you need
1720	28	4	something to do
1721	28	3	take it out
1722	28	4	that i need
1723	28	4	the definition of
1724	28	4	the other thing
1725	28	3	the point is
1726	28	4	the point of
1727	28	3	the right to
1728	28	3	the time to
1729	28	4	the way we
1730	28	4	then we can
1731	28	3	they don't want
1732	28	4	this is how
1733	28	5	those are the
1734	28	4	to be at
1735	28	4	to get rid
1736	28	3	to see how
1737	28	3	to use the
1738	28	3	to work on
1739	28	3	was a little
1740	28	4	we all have
1741	28	4	we still have
1742	28	3	why do we
1743	28	3	would you say
1744	28	3	you can't get
1745	28	4	you had the
1746	28	3	you know why
1747	28	4	you may have

1748	28	3	you really don't
1749	28	3	you will have
1750	27	3	a few months
1751	27	3	a good idea
1752	27	4	a variety of
1753	27	5	and all of
1754	27	4	and one of
1755	27	4	and they are
1756	27	3	and what you
1757	27	3	and you've got
1758	27	3	but i didn't
1759	27	3	but i just
1760	27	3	but in the
1761	27	5	can do a
1762	27	3	can do it
1763	27	4	can we do
1764	27	3	can we just
1765	27	3	can you take
1766	27	5	cells in the
1767	27	5	do not have
1768	27	3	for her to
1769	27	3	for them to
1770	27	5	give you an
1771	27	4	had to go
1772	27	3	have to start
1773	27	3	how much you
1774	27	3	i can't remember
1775	27	3	i could tell
1776	27	3	i have this
1777	27	4	i mean i
1778	27	3	i think if
1779	27	5	in for a
1780	27	3	in the back
1781	27	5	is like a
1782	27	3	is why we
1783	27	4	it all the
1784	27	3	it in my
1785	27	3	it with a
1786	27	3	it's a bad
1787	27	4	kind of thing
1788	27	3	know what else
1789	27	3	me on the
1790	27	3	me when i
1791	27	5	not just the

1792	27	4	now i know
1793	27	4	now if you
1794	27	3	now you can
1795	27	4	of the body
1796	27	4	on the same
1797	27	3	once in a
1798	27	3	or it could
1799	27	4	running out of
1800	27	3	see what i
1801	27	3	she had to
1802	27	3	should have been
1803	27	4	that we could
1804	27	3	that you want
1805	27	3	that's what we
1806	27	5	the age of
1807	27	3	the tumor is
1808	27	4	the whole thing
1809	27	3	there was an
1810	27	3	there we go.
1811	27	3	think i know
1812	27	4	this could be
1813	27	3	thought it would
1814	27	4	to be so
1815	27	4	to give her
1816	27	3	to tell you.
1817	27	3	to wait for
1818	27	3	to you and
1819	27	3	up to a
1820	27	3	want to look
1821	27	3	wanted to go
1822	27	5	was the first
1823	27	4	way to the
1824	27	4	we could have
1825	27	3	we get the
1826	27	5	we know that
1827	27	4	we wanted to
1828	27	4	we've talked about
1829	27	3	well, you know
1830	27	4	what i said
1831	27	4	what i think
1832	27	3	what this is
1833	27	3	what you have
1834	27	3	where did you
1835	27	3	which is what

1836	27	3	will be in
1837	27	3	will not be
1838	27	3	you do that
1839	27	4	you get that
1840	27	3	you to keep
1841	27	4	you're looking for
1842	26	5	a few days
1843	26	3	a little bit.
1844	26	3	and if it's
1845	26	3	and that was
1846	26	3	and the next
1847	26	4	and then it
1848	26	3	because it was
1849	26	3	but if we
1850	26	3	but there's a
1851	26	4	can do is
1852	26	5	can give you
1853	26	3	can tell me
1854	26	3	does that make
1855	26	3	doesn't mean you
1856	26	4	doesn't need to
1857	26	3	don't like to
1858	26	3	don't look at
1859	26	4	don't think that's
1860	26	4	don't think we
1861	26	4	every one of
1862	26	3	first of all
1863	26	4	for a couple
1864	26	3	for a while,
1865	26	4	go to a
1866	26	4	going to give
1867	26	3	have a better
1868	26	3	have nothing to
1869	26	4	have you heard
1870	26	3	he has no
1871	26	4	i can give
1872	26	4	i do have
1873	26	4	i would do
1874	26	3	if she has
1875	26	3	if they don't
1876	26	3	if you didn't
1877	26	5	if you take
1878	26	3	in the past
1879	26	3	is that your

1880	26	3	is there something
1881	26	4	it through the
1882	26	3	know how i
1883	26	4	let me tell
1884	26	3	like to think
1885	26	3	maybe it was
1886	26	4	might have a
1887	26	3	my way to
1888	26	3	not a bad
1889	26	3	of the day
1890	26	4	of the other
1891	26	3	on the left
1892	26	3	on your own
1893	26	4	so if i
1894	26	3	so you think
1895	26	3	so you've got
1896	26	3	something wrong with
1897	26	3	take it from
1898	26	4	tell you how
1899	26	4	that can be
1900	26	3	that in a
1901	26	3	that is so
1902	26	3	the best thing
1903	26	4	the cancer cells
1904	26	3	the chance to
1905	26	4	the majority of
1906	26	3	the next time
1907	26	3	the worst thing
1908	26	4	there in the
1909	26	4	there's got to
1910	26	4	thing that you
1911	26	4	think i need
1912	26	4	time in the
1913	26	4	to come back
1914	26	3	to come out
1915	26	4	to focus on
1916	26	3	to get that
1917	26	4	to get through
1918	26	3	to operate on
1919	26	4	was the only
1920	26	4	we can be
1921	26	4	we can go
1922	26	4	we can see
1923	26	3	we need more

1924	26	3	we used to
1925	26	4	what we do
1926	26	3	what we do.
1927	26	4	will be a
1928	26	3	you for your
1929	26	3	you have something
1930	26	4	you still have
1931	26	3	you very much
1932	26	3	you want it
1933	26	3	you went to
1934	26	3	you'd have to
1935	26	3	you're in a
1936	25	3	about the fact
1937	25	4	all you have
1938	25	3	and he's a
1939	25	4	and so the
1940	25	3	and some of
1941	25	3	and then there's
1942	25	3	are you the
1943	25	3	as you know
1944	25	3	at the back
1945	25	4	be a better
1946	25	3	be nice to
1947	25	3	because i can't
1948	25	3	but i would
1949	25	3	but we need
1950	25	3	but you need
1951	25	4	but you should
1952	25	3	but you're gonna
1953	25	4	can tell you
1954	25	3	can you see
1955	25	3	can't tell you
1956	25	3	do not want
1957	25	4	end up with
1958	25	4	go in and
1959	25	4	going to die
1960	25	3	got a good
1961	25	3	have a patient
1962	25	4	he would have
1963	25	3	him on a
1964	25	3	how long do
1965	25	3	i can hear
1966	25	3	i can say
1967	25	3	i could take

1968	25	5	i have one
1969	25	4	i just didn't
1970	25	3	i say we
1971	25	3	i still think
1972	25	4	i think they
1973	25	3	i'm just going
1974	25	3	if i just
1975	25	4	if you're a
1976	25	3	in the body
1977	25	3	it if you
1978	25	3	it sounds like
1979	25	3	it was all
1980	25	3	it's a great
1981	25	3	it's got to
1982	25	3	it's the right
1983	25	4	it's the same
1984	25	3	just like you
1985	25	4	just look at
1986	25	5	kind of like
1987	25	3	know it's a
1988	25	3	know what they
1989	25	3	need to have
1990	25	4	not the same
1991	25	4	now you have
1992	25	4	or at least
1993	25	4	other side of
1994	25	3	she didn't want
1995	25	4	should be a
1996	25	3	so it's a
1997	25	3	so that we
1998	25	3	so there's a
1999	25	4	so we need
2000	25	3	so you want
2001	25	4	still in the
2002	25	4	talking about a
2003	25	4	that you should
2004	25	3	the bone marrow
2005	25	4	the people who
2006	25	3	the source of
2007	25	4	the surface of
2008	25	4	the way they
2009	25	4	there must be
2010	25	4	there's been a
2011	25	4	think that's a

2012	25	4	to be more
2013	25	3	to get on
2014	25	3	to hear about
2015	25	5	to see what
2016	25	3	to stay in
2017	25	4	two or three
2018	25	3	up on a
2019	25	4	was diagnosed with
2020	25	4	we could get
2021	25	3	we don't even
2022	25	3	we don't get
2023	25	5	we look at
2024	25	4	we're looking at
2025	25	3	what about a
2026	25	3	what does this
2027	25	3	what have you
2028	25	4	what i have
2029	25	3	what was i
2030	25	3	what you were
2031	25	3	when do you
2032	25	3	who i am
2033	25	4	will be the
2034	25	4	you do this
2035	25	4	you don't feel
2036	25	4	you know it
2037	25	4	you think is
2038	25	3	you to come
2039	25	3	you were supposed
2040	25	3	you've got the
2041	24	3	a few hours
2042	24	3	about to go
2043	24	3	and i really
2044	24	3	and i'm sure
2045	24	4	and talk about
2046	24	3	and we don't
2047	24	3	and we will
2048	24	4	at least one
2049	24	3	because of a
2050	24	4	can do this
2051	24	3	can't believe you
2052	24	4	could do a
2053	24	5	do a lot
2054	24	5	do i get
2055	24	4	do we know

2056	24	3	expect you to
2057	24	3	fact that you're
2058	24	4	for a while
2059	24	5	for it to
2060	24	3	get you some
2061	24	5	has to do
2062	24	3	he told me
2063	24	4	hole in the
2064	24	3	how the hell
2065	24	3	i know we
2066	24	4	i should say
2067	24	3	if the patient
2068	24	4	in a little
2069	24	3	in the er
2070	24	5	in the united
2071	24	3	it is that
2072	24	4	it takes a
2073	24	3	it was really
2074	24	3	it's just the
2075	24	3	it's like to
2076	24	3	it's the first
2077	24	3	just because you
2078	24	3	know it was
2079	24	4	less than a
2080	24	3	look at that
2081	24	5	looking for the
2082	24	3	lot of time
2083	24	3	might be the
2084	24	5	might have been
2085	24	3	not sure i
2086	24	3	nothing we can
2087	24	3	of you are
2088	24	3	right to be
2089	24	3	see if we
2090	24	4	set up for
2091	24	3	should be in
2092	24	3	so if you're
2093	24	4	some of us
2094	24	4	something like that
2095	24	3	talking to you
2096	24	3	tell me where
2097	24	3	terms of the
2098	24	3	that i want
2099	24	4	that i'm a

2100	24	4	that might be
2101	24	3	that's what they
2102	24	3	the bottom of
2103	24	5	the level of
2104	24	3	the one you
2105	24	3	the patient has
2106	24	4	the reason i
2107	24	4	the same way
2108	24	4	them in the
2109	24	4	they're trying to
2110	24	4	this sort of
2111	24	4	this thing is
2112	24	3	to get an
2113	24	3	to give up
2114	24	4	to see a
2115	24	3	to speak to
2116	24	4	to tell the
2117	24	3	to tell them
2118	24	3	uh, this is
2119	24	3	want them to
2120	24	3	want to ask
2121	24	3	wanted to talk
2122	24	3	was out of
2123	24	4	was wondering if
2124	24	4	we can find
2125	24	3	we got the
2126	24	3	we like to
2127	24	4	we would have
2128	24	4	we'd like to
2129	24	5	we're in the
2130	24	3	well, it's not
2131	24	3	went to a
2132	24	3	what are they
2133	24	3	what i do
2134	24	4	what you need
2135	24	3	whatever it is
2136	24	3	will have to
2137	24	3	wrong with the
2138	24	3	you and the
2139	24	3	you didn't know
2140	24	4	you don't see
2141	24	3	you get your
2142	24	4	you say that
2143	24	4	you so much

2144	23	4	a kind of
2145	23	4	a sense of
2146	23	4	all of our
2147	23	5	all of your
2148	23	3	am not a
2149	23	3	and get the
2150	23	4	and i said
2151	23	3	and then there
2152	23	3	and we need
2153	23	5	appears to be
2154	23	3	at least i
2155	23	3	be a doctor
2156	23	3	but we're not
2157	23	5	can do that
2158	23	3	can have a
2159	23	3	care about the
2160	23	3	coming out of
2161	23	4	do anything about
2162	23	4	do we get
2163	23	3	does he have
2164	23	4	doesn't mean that
2165	23	3	doesn't seem to
2166	23	3	don't know yet.
2167	23	3	for all of
2168	23	3	give up on
2169	23	4	go back in
2170	23	4	going to happen
2171	23	4	had no idea
2172	23	3	have a problem.
2173	23	3	have to come
2174	23	5	have to look
2175	23	3	haven't had a
2176	23	3	he's trying to
2177	23	3	i came to
2178	23	3	i can only
2179	23	3	i just think
2180	23	4	i was at
2181	23	3	i will take
2182	23	3	i've been in
2183	23	4	if i want
2184	23	4	if you weren't
2185	23	3	if you would
2186	23	3	in a hospital
2187	23	3	in a moment

2188	23	4	in one of
2189	23	3	in with a
2190	23	5	is it that
2191	23	4	know what we
2192	23	4	know you have
2193	23	3	like to go
2194	23	3	look at you
2195	23	5	looking at a
2196	23	4	may have to
2197	23	3	never have to
2198	23	4	next time you
2199	23	4	no evidence of
2200	23	4	of the brain
2201	23	5	of you who
2202	23	3	on the phone
2203	23	3	on the side
2204	23	4	on the table
2205	23	3	out for a
2206	23	3	problem with the
2207	23	4	that i think
2208	23	4	that there was
2209	23	4	that you had
2210	23	5	that's the only
2211	23	4	the name of
2212	23	5	the right side
2213	23	4	the same time
2214	23	5	the united states
2215	23	4	there is an
2216	23	4	there was some
2217	23	4	think it's the
2218	23	5	this would be
2219	23	3	time to get
2220	23	3	to come in
2221	23	3	to do that,
2222	23	3	to give it
2223	23	3	to go home
2224	23	3	to have some
2225	23	3	to me that
2226	23	3	to put in
2227	23	3	to rule out
2228	23	3	to take that
2229	23	5	turns out that
2230	23	3	waiting for the
2231	23	3	want to help

2232	23	3	want to work
2233	23	3	was kind of
2234	23	3	we don't do
2235	23	3	we know what
2236	23	3	we try to
2237	23	3	we're not doing
2238	23	4	what happens if
2239	23	3	what i need
2240	23	3	what you're gonna
2241	23	3	what's happening to
2242	23	3	with all the
2243	23	3	you can come
2244	23	3	you can just
2245	23	3	you can think
2246	23	4	you know i'm
2247	23	3	you know this
2248	23	3	you make a
2249	23	3	you might as
2250	23	3	you see that
2251	23	3	you wouldn't be
2252	23	3	you're not getting
2253	22	4	a cup of
2254	22	3	a pain in
2255	22	4	a pretty good
2256	22	3	all of this
2257	22	3	and a lot
2258	22	3	and i'll be
2259	22	3	and then we're
2260	22	4	and then when
2261	22	5	and try to
2262	22	4	are supposed to
2263	22	4	at the beginning
2264	22	3	back of the
2265	22	3	but it could
2266	22	3	but we can
2267	22	3	can make it
2268	22	3	can't believe i
2269	22	3	close to the
2270	22	3	did you ever
2271	22	4	don't have an
2272	22	3	don't you want
2273	22	3	down on the
2274	22	3	even if we
2275	22	3	first time in

2276	22	3	front of a
2277	22	3	gave me a
2278	22	3	get on the
2279	22	4	go back and
2280	22	3	go for a
2281	22	4	go in there
2282	22	3	gonna do it.
2283	22	3	got a better
2284	22	3	had nothing to
2285	22	4	happens to be
2286	22	3	have any more
2287	22	3	he went to
2288	22	4	how many people
2289	22	3	i believe in
2290	22	3	i give you
2291	22	3	i just wanna
2292	22	3	i must have
2293	22	3	i spoke to
2294	22	5	i think about
2295	22	3	i thought the
2296	22	3	i'll get the
2297	22	3	i'm doing a
2298	22	3	i'm the only
2299	22	3	i've got the
2300	22	5	if you ask
2301	22	3	if you ever
2302	22	3	if you give
2303	22	4	in this room
2304	22	4	it doesn't have
2305	22	3	it had to
2306	22	4	it kind of
2307	22	4	it might not
2308	22	3	it takes to
2309	22	4	it was so
2310	22	3	it's a long
2311	22	3	know that we
2312	22	3	know what he
2313	22	5	let's talk about
2314	22	4	likely to be
2315	22	4	little bit about
2316	22	3	make sure she
2317	22	3	maybe we can
2318	22	3	me as a
2319	22	3	me to come

2320	22	3	more than one
2321	22	3	nice to see
2322	22	4	not be able
2323	22	3	on one of
2324	22	5	part of a
2325	22	4	set up a
2326	22	3	so we don't
2327	22	4	some of them
2328	22	5	some of these
2329	22	4	still have a
2330	22	3	still have to
2331	22	3	take him to
2332	22	3	talk to my
2333	22	4	that could be
2334	22	3	that means that
2335	22	5	that there's a
2336	22	4	that you can't
2337	22	3	the one to
2338	22	4	the process of
2339	22	4	the way up
2340	22	3	the world is
2341	22	3	there is one
2342	22	4	there may be
2343	22	3	there's a reason
2344	22	4	this is really
2345	22	4	to be some
2346	22	3	to give him
2347	22	3	to leave the
2348	22	3	to move in
2349	22	3	to put a
2350	22	3	to see your
2351	22	4	to the other
2352	22	4	to the right
2353	22	3	turned out to
2354	22	3	we all know
2355	22	4	we can take
2356	22	4	we don't want
2357	22	3	what i'm going
2358	22	3	what was that
2359	22	4	what we can
2360	22	3	when you think
2361	22	4	who's going to
2362	22	3	won't have to
2363	22	3	would i do

2364	22	5	you about the
2365	22	4	you could get
2366	22	3	you don't just
2367	22	3	you know that,
2368	22	3	you made it
2369	22	3	you to talk
2370	22	3	you took a
2371	22	3	you were right
2372	22	3	you're okay with
2373	21	3	a call from
2374	21	3	a few minutes
2375	21	3	a symptom of
2376	21	4	according to the
2377	21	4	all of a
2378	21	5	an example of
2379	21	4	and for the
2380	21	3	and it doesn't
2381	21	4	and most of
2382	21	3	and out of
2383	21	4	and see what
2384	21	3	and you just
2385	21	4	are on the
2386	21	3	ask you a
2387	21	4	been thinking about
2388	21	3	but i got
2389	21	4	but there was
2390	21	3	but we don't
2391	21	3	but when i
2392	21	3	can't talk about
2393	21	3	come in and
2394	21	3	didn't have any
2395	21	3	didn't know that
2396	21	3	do all the
2397	21	4	do have a
2398	21	4	do something about
2399	21	4	do the same
2400	21	3	even if it
2401	21	3	every now and
2402	21	4	fluid in the
2403	21	3	focus on the
2404	21	3	for him to
2405	21	3	get a good
2406	21	3	get into the
2407	21	3	give you some

2408	21	4	going to see
2409	21	3	he won't be
2410	21	3	head of the
2411	21	3	i can think
2412	21	3	i take a
2413	21	4	i understand that
2414	21	4	i will never
2415	21	3	i'd have to
2416	21	4	if you see
2417	21	4	if you're going
2418	21	3	in case you
2419	21	3	in in the
2420	21	3	in the eye
2421	21	3	in the room
2422	21	4	is much more
2423	21	4	is the most
2424	21	4	is what we
2425	21	4	it has a
2426	21	3	it was in
2427	21	4	it's about the
2428	21	3	it's only been
2429	21	3	just for a
2430	21	3	know about the
2431	21	3	let's have a
2432	21	3	like me to
2433	21	3	like to get
2434	21	5	look at it
2435	21	3	looks like you
2436	21	3	me feel like
2437	21	4	next to the
2438	21	3	no time for
2439	21	3	on the floor
2440	21	4	out on a
2441	21	3	patient with a
2442	21	3	say that i
2443	21	4	so i'm going
2444	21	5	so that's what
2445	21	3	stop looking at
2446	21	4	such a good
2447	21	3	take you up
2448	21	3	that i had
2449	21	3	that she has
2450	21	4	that they are
2451	21	3	that's exactly what

2452	21	4	that's what the
2453	21	3	that's why he
2454	21	3	the brain is
2455	21	4	the cause of
2456	21	3	the inside of
2457	21	5	the most common
2458	21	5	the problem is
2459	21	4	there might be
2460	21	3	think about what
2461	21	3	think it would
2462	21	3	thought you might
2463	21	3	three or four
2464	21	3	to be there
2465	21	3	to do about
2466	21	3	to each other.
2467	21	3	to find out.
2468	21	3	to go and
2469	21	4	to kill the
2470	21	4	to take this
2471	21	3	to the hospital
2472	21	4	to use a
2473	21	3	to want to
2474	21	3	took care of
2475	21	3	we can talk
2476	21	3	we go to
2477	21	3	went back to
2478	21	4	what else is
2479	21	3	what i mean
2480	21	3	what you said
2481	21	3	where are the
2482	21	3	why i didn't
2483	21	4	would love to
2484	21	3	would you have
2485	21	3	you can look
2486	21	3	you can stop
2487	21	4	you could take
2488	21	3	you do know
2489	21	3	you have some
2490	21	3	you know you're
2491	21	5	you try to
2492	21	3	you very much.
2493	21	3	you want your
2494	21	3	you'll be able
2495	21	3	you're on the

2496	20	3	a little too
2497	20	3	a load of
2498	20	3	a patient who
2499	20	3	a person who
2500	20	3	all we have
2501	20	3	and get her
2502	20	3	and i like
2503	20	4	and they don't
2504	20	4	and this was
2505	20	3	any of your
2506	20	3	are plenty of
2507	20	3	ask me to
2508	20	3	at least you
2509	20	4	be one of
2510	20	4	be talking about
2511	20	4	can do about
2512	20	4	can do to
2513	20	4	can look at
2514	20	4	can think of
2515	20	4	caused by the
2516	20	3	do anything to
2517	20	4	do i know
2518	20	3	do we need
2519	20	3	doesn't have any
2520	20	3	don't know anything
2521	20	3	don't think she
2522	20	3	exactly what i
2523	20	4	for more than
2524	20	4	going to try
2525	20	3	got a big
2526	20	3	got a new
2527	20	4	have a new
2528	20	4	have had a
2529	20	3	have to deal
2530	20	3	he could be
2531	20	3	he didn't have
2532	20	3	he had to
2533	20	3	he was the
2534	20	3	he's not going
2535	20	3	how to get
2536	20	3	i bet you
2537	20	3	i know about
2538	20	4	i said to
2539	20	5	i think is

2540	20	3	i will have
2541	20	3	i'll have to
2542	20	3	i'm done with
2543	20	3	i'm telling you
2544	20	4	if it doesn't
2545	20	3	if we do
2546	20	3	if you find
2547	20	4	if you let
2548	20	3	if you wanna
2549	20	3	in less than
2550	20	5	is in a
2551	20	3	is it the
2552	20	3	is the right
2553	20	4	is the same
2554	20	4	is the way
2555	20	3	is trying to
2556	20	4	it for the
2557	20	3	it means that
2558	20	3	it out on
2559	20	3	it's just not
2560	20	3	it's not as
2561	20	3	just saying that
2562	20	5	lot of things
2563	20	3	maybe you can
2564	20	4	me tell you
2565	20	4	more and more
2566	20	3	need to ask
2567	20	3	needs to go
2568	20	3	no history of
2569	20	3	not interested in
2570	20	3	nothing you can
2571	20	4	of it as
2572	20	5	of you have
2573	20	4	parts of the
2574	20	3	people who have
2575	20	5	see that the
2576	20	3	she's a little
2577	20	4	should not be
2578	20	4	size of a
2579	20	4	sort of a
2580	20	3	tell us what
2581	20	4	tell you the
2582	20	4	that it is
2583	20	4	that it's not

2584	20	4	that they were
2585	20	4	that we are
2586	20	3	the blood supply
2587	20	5	the blood vessels
2588	20	4	the course of
2589	20	4	the last few
2590	20	3	the lymph node
2591	20	3	the one i
2592	20	4	the ones that
2593	20	3	the same as
2594	20	3	the time i
2595	20	3	there has to
2596	20	5	there is some
2597	20	3	they had to
2598	20	3	think it's gonna
2599	20	4	time to go
2600	20	4	to be something
2601	20	3	to be very
2602	20	3	to do everything
2603	20	4	to do some
2604	20	3	to get over
2605	20	3	to get up
2606	20	4	to give them
2607	20	4	to have that
2608	20	3	to have this
2609	20	3	to keep him
2610	20	3	to put on
2611	20	3	to think that
2612	20	3	very difficult to
2613	20	4	we can have
2614	20	4	we could be
2615	20	3	we're not going
2616	20	3	well, you know,
2617	20	3	what did we
2618	20	4	what is it
2619	20	3	what is that
2620	20	4	what is this
2621	20	4	when you do
2622	20	3	which means the
2623	20	3	will give you
2624	20	4	you can start
2625	20	4	you come to
2626	20	3	you do what
2627	20	3	you don't give

2628	20	3	you don't really
2629	20	3	you may not
2630	20	5	you might not
2631	20	4	you put the
2632	20	4	you really need
2633	20	3	you that you
2634	20	3	you think they
2635	20	3	you would be
2636	20	4	you're not in
2637	19	3	a few things
2638	19	5	a few weeks
2639	19	5	a few years
2640	19	5	a group of
2641	19	3	a very small
2642	19	3	a whole lot
2643	19	3	all i have
2644	19	4	and it can
2645	19	4	and see how
2646	19	4	and so i
2647	19	3	and they have
2648	19	3	and what do
2649	19	3	any of this
2650	19	4	are trying to
2651	19	4	as i said
2652	19	3	at this point
2653	19	4	because i think
2654	19	3	because it's not
2655	19	3	because of my
2656	19	4	because that's what
2657	19	4	because we have
2658	19	3	because you know
2659	19	3	because you want
2660	19	3	because you're not
2661	19	4	between the two
2662	19	3	but i thought
2663	19	3	but the fact
2664	19	3	came in with
2665	19	5	came out of
2666	19	4	can go to
2667	19	3	can take a
2668	19	3	can't be a
2669	19	3	can't have a
2670	19	3	caused by a
2671	19	3	couple of weeks

2672	19	4	do it on
2673	19	3	don't know that
2674	19	3	don't need any
2675	19	3	feel free to
2676	19	4	find out about
2677	19	4	find out if
2678	19	3	find out what's
2679	19	3	first time i
2680	19	5	for the most
2681	19	4	gave you a
2682	19	3	go on to
2683	19	4	go up to
2684	19	3	going to come
2685	19	4	going to find
2686	19	3	half an hour
2687	19	3	have a heart
2688	19	3	have a look
2689	19	3	have anything to
2690	19	4	have one of
2691	19	3	here. this is
2692	19	4	him to be
2693	19	3	i can still
2694	19	3	i guess it's
2695	19	3	i hope you're
2696	19	3	i may be
2697	19	4	i said you
2698	19	3	i showed you
2699	19	3	i suggest you
2700	19	3	i think there's
2701	19	3	i'll do it
2702	19	3	i'm kind of
2703	19	3	i'm on the
2704	19	3	i'm sorry to
2705	19	3	i'm starting to
2706	19	4	i'm sure that
2707	19	4	immune system is
2708	19	3	is full of
2709	19	4	is part of
2710	19	3	is such a
2711	19	4	is that it
2712	19	3	it at the
2713	19	4	it may not
2714	19	3	just don't know
2715	19	3	just make sure

2716	19	4	know if it's
2717	19	3	know why you
2718	19	3	let me go
2719	19	4	look at a
2720	19	3	looks like the
2721	19	3	much as i
2722	19	4	none of them
2723	19	3	not having an
2724	19	3	of the world
2725	19	5	on in the
2726	19	3	on the surface
2727	19	4	out for the
2728	19	3	patient has a
2729	19	3	saying that you
2730	19	4	see if it
2731	19	3	see if the
2732	19	3	see what we
2733	19	3	see what you
2734	19	3	she has an
2735	19	3	should have a
2736	19	3	so do you
2737	19	3	so you're not
2738	19	3	something that i
2739	19	3	source of the
2740	19	3	take a few
2741	19	3	that we were
2742	19	5	that when you
2743	19	4	that's going to
2744	19	4	that's kind of
2745	19	4	the best way
2746	19	4	the idea of
2747	19	3	the only person
2748	19	3	the patient to
2749	19	3	the person who
2750	19	3	the results of
2751	19	5	the use of
2752	19	3	them out of
2753	19	4	then you have
2754	19	3	there are plenty
2755	19	3	there could be
2756	19	3	there's no such
2757	19	5	they can be
2758	19	4	they wanted to
2759	19	3	they're gonna be

2760	19	3	things that i
2761	19	3	think about the
2762	19	4	this is exactly
2763	19	3	this part of
2764	19	4	this was the
2765	19	3	to do it,
2766	19	4	to get his
2767	19	3	to put him
2768	19	4	told you that
2769	19	3	until you get
2770	19	3	up at the
2771	19	4	very hard to
2772	19	4	we did a
2773	19	3	we should just
2774	19	4	we're not talking
2775	19	4	what can you
2776	19	3	what else could
2777	19	5	what i would
2778	19	3	what i'm doing
2779	19	3	what we're doing
2780	19	3	what you're doing
2781	19	4	which one is
2782	19	3	while i was
2783	19	3	why is this
2784	19	3	work on the
2785	19	3	would be an
2786	19	4	would it be
2787	19	3	would you be
2788	19	5	you can find
2789	19	4	you can still
2790	19	3	you give them
2791	19	4	you got it
2792	19	3	you may be
2793	19	3	you need an
2794	19	4	you on a
2795	19	3	you put a
2796	19	3	you when you
2797	18	3	a heart attack
2798	18	5	a lot about
2799	18	3	a period of
2800	18	4	a result of
2801	18	4	a way of
2802	18	3	about it and
2803	18	4	about to be

2804	18	3	about what i
2805	18	4	all of those
2806	18	3	always wanted to
2807	18	4	and by the
2808	18	3	and do a
2809	18	3	and even if
2810	18	3	and he doesn't
2811	18	4	and he has
2812	18	4	and i'm just
2813	18	3	and if you're
2814	18	3	and it may
2815	18	4	and lots of
2816	18	3	and none of
2817	18	4	and on the
2818	18	3	and the rest
2819	18	3	and wait for
2820	18	4	and what i
2821	18	4	any kind of
2822	18	5	any one of
2823	18	5	are in a
2824	18	5	are there any
2825	18	3	be more than
2826	18	3	because she was
2827	18	3	because you are
2828	18	3	but at least
2829	18	3	can get the
2830	18	3	can make a
2831	18	3	can make you
2832	18	4	care of the
2833	18	4	could take a
2834	18	3	did a good
2835	18	3	didn't do it
2836	18	3	do you not
2837	18	3	does it look
2838	18	3	got a problem
2839	18	4	have a chance
2840	18	3	have to learn
2841	18	3	have you done
2842	18	4	how does that
2843	18	3	i agree with
2844	18	3	i can find
2845	18	3	i can't say
2846	18	3	i did what
2847	18	4	i forgot to

2848	18	3	i have two
2849	18	3	i just said
2850	18	3	i made it
2851	18	3	i should just
2852	18	4	i think we're
2853	18	4	i think you'll
2854	18	3	i will give
2855	18	3	i'm sure it's
2856	18	3	i've got an
2857	18	3	if it makes
2858	18	3	if you just
2859	18	5	in a very
2860	18	5	in the brain
2861	18	4	is no longer
2862	18	4	is not going
2863	18	4	is that i
2864	18	3	it look like
2865	18	3	it on my
2866	18	3	it wouldn't be
2867	18	3	it's the most
2868	18	3	just to make
2869	18	5	know that it
2870	18	4	know that it's
2871	18	3	like to hear
2872	18	3	like to know
2873	18	3	look out for
2874	18	3	made me feel
2875	18	4	make a big
2876	18	4	me a little
2877	18	3	me because i
2878	18	3	me that i
2879	18	3	me to talk
2880	18	5	much of the
2881	18	3	no idea how
2882	18	3	none of this
2883	18	4	not enough to
2884	18	3	not sure what
2885	18	3	of all of
2886	18	3	of people with
2887	18	4	of the way
2888	18	5	of this is
2889	18	3	of us to
2890	18	4	one in the
2891	18	4	or something like

2892	18	3	out of bed
2893	18	3	put your hand
2894	18	3	right in the
2895	18	5	say that you
2896	18	3	she's in a
2897	18	3	show up on
2898	18	3	so i thought
2899	18	4	some of you
2900	18	3	spend time with
2901	18	3	spent the last
2902	18	3	tell me if
2903	18	4	that you might
2904	18	3	that's all i'm
2905	18	3	that's what it
2906	18	5	the beginning of
2907	18	4	the best of
2908	18	5	the history of
2909	18	3	the key to
2910	18	4	the last three
2911	18	3	the other one
2912	18	5	the people that
2913	18	4	the way the
2914	18	4	then there was
2915	18	4	then you get
2916	18	3	there won't be
2917	18	4	there's no sign
2918	18	3	there's nothing we
2919	18	3	think about it,
2920	18	3	think he was
2921	18	3	think i got
2922	18	4	this might be
2923	18	4	this will be
2924	18	3	to call the
2925	18	3	to cut off
2926	18	3	to keep your
2927	18	3	to need to
2928	18	3	to repair the
2929	18	3	to see that
2930	18	3	to the left
2931	18	3	to work for
2932	18	3	to you that
2933	18	3	try to make
2934	18	3	um, i have
2935	18	4	want to find

2936	18	3	want to lose
2937	18	4	want to use
2938	18	3	wanted to have
2939	18	5	wanted to know
2940	18	4	was one of
2941	18	3	was talking to
2942	18	4	we can make
2943	18	4	we can put
2944	18	3	we can try
2945	18	3	we can't get
2946	18	4	we do the
2947	18	4	we do this
2948	18	4	we looked at
2949	18	4	we might be
2950	18	3	were looking for
2951	18	3	what about your
2952	18	4	what we have
2953	18	3	when you say
2954	18	4	which is not
2955	18	3	why is that
2956	18	3	would you want
2957	18	3	you a few
2958	18	3	you believe in
2959	18	3	you do with
2960	18	3	you got me
2961	18	3	you gotta do
2962	18	3	you have two
2963	18	3	you just had
2964	18	3	you see what
2965	18	3	you talk about
2966	18	3	you to make
2967	18	3	you were trying
2968	18	3	you when i
2969	18	3	you won't have
2970	18	3	you're doing a
2971	17	3	a lot better
2972	17	3	a very nice
2973	17	3	able to tell
2974	17	3	all you can
2975	17	3	am i going
2976	17	3	and have a
2977	17	3	and he had
2978	17	3	and he's got
2979	17	4	and i won't

2980	17	4	and into the
2981	17	3	and tell them
2982	17	3	and that's a
2983	17	4	and the reason
2984	17	3	and you could
2985	17	4	any of us
2986	17	4	anything to do
2987	17	5	are not the
2988	17	3	at least we
2989	17	3	be a part
2990	17	3	be on your
2991	17	3	because you can't
2992	17	4	best way to
2993	17	3	better than i
2994	17	4	but he was
2995	17	3	but that doesn't
2996	17	3	but then you
2997	17	4	but they don't
2998	17	3	came to see
2999	17	3	can get to
3000	17	4	can lead to
3001	17	3	can you make
3002	17	3	comes down to
3003	17	3	did you want
3004	17	3	don't know the
3005	17	3	don't see a
3006	17	5	fact that the
3007	17	5	far as i
3008	17	3	figure out why
3009	17	4	get them to
3010	17	4	go through the
3011	17	3	go with the
3012	17	4	going on here?
3013	17	4	going to ask
3014	17	3	going to look
3015	17	4	good at it.
3016	17	3	good luck with
3017	17	3	got a patient
3018	17	3	have a couple
3019	17	3	have you got
3020	17	3	how many of
3021	17	3	i did an
3022	17	3	i had my
3023	17	4	i just couldn't

3024	17	3	i made the
3025	17	3	i really want
3026	17	3	i won't have
3027	17	3	i'm saying that
3028	17	4	i've been doing
3029	17	3	if it's the
3030	17	4	if they have
3031	17	3	if we had
3032	17	3	if we were
3033	17	3	if you keep
3034	17	3	in the laboratory
3035	17	3	in the lungs
3036	17	3	in the lungs.
3037	17	3	is exactly what
3038	17	4	is if you
3039	17	3	is it just
3040	17	4	is to be
3041	17	4	is what i'm
3042	17	3	it means you
3043	17	3	it seems to
3044	17	4	it to a
3045	17	3	it took me
3046	17	3	it used to
3047	17	3	it's not too
3048	17	3	just go back
3049	17	4	just to get
3050	17	3	know when you
3051	17	3	know you're not
3052	17	5	let me just
3053	17	4	like a good
3054	17	3	make sure i
3055	17	3	meant to be
3056	17	4	more or less
3057	17	3	more than you
3058	17	3	must be a
3059	17	3	need to come
3060	17	3	need to think
3061	17	3	no way of
3062	17	4	not a very
3063	17	3	not that i'm
3064	17	3	now i'm gonna
3065	17	4	of the heart
3066	17	3	of the virus
3067	17	3	once you get

3068	17	4	other than the
3069	17	3	over and over
3070	17	5	put them in
3071	17	3	quite a lot
3072	17	5	really have to
3073	17	3	see it in
3074	17	3	she asked me
3075	17	3	she didn't have
3076	17	3	she is a
3077	17	3	she was the
3078	17	3	she would be
3079	17	4	so i will
3080	17	3	so that's a
3081	17	3	so there is
3082	17	3	so what does
3083	17	3	so what is
3084	17	3	so why are
3085	17	3	stay in the
3086	17	3	still trying to
3087	17	3	tell me it's
3088	17	4	that as a
3089	17	3	that should be
3090	17	3	that the patient
3091	17	3	that's the way
3092	17	5	the answer to
3093	17	4	the heart and
3094	17	3	the most likely
3095	17	5	the part of
3096	17	4	the point where
3097	17	3	the thing about
3098	17	4	the type of
3099	17	3	the virus is
3100	17	3	the wall of
3101	17	3	then i have
3102	17	4	then i was
3103	17	3	then i will
3104	17	3	there's lots of
3105	17	3	they are not
3106	17	3	they tend to
3107	17	4	think i'm a
3108	17	4	think of a
3109	17	3	this isn't the
3110	17	3	thought i had
3111	17	3	three times a

3112	17	4	to all the
3113	17	4	to do anything
3114	17	4	to do in
3115	17	4	to go for
3116	17	3	to know why
3117	17	3	to open up
3118	17	3	to remind you
3119	17	4	to say about
3120	17	3	to see it
3121	17	3	to stand up
3122	17	3	to the patient
3123	17	4	to you in
3124	17	3	trying to take
3125	17	4	us to do
3126	17	3	wake up and
3127	17	3	want to stop
3128	17	3	want to try
3129	17	3	wanted to give
3130	17	3	was at the
3131	17	3	was the one
3132	17	4	we do a
3133	17	4	what can we
3134	17	3	what happens to
3135	17	4	why i don't
3136	17	4	why would we
3137	17	3	wondering if you
3138	17	3	years ago, i
3139	17	4	you can use
3140	17	3	you can't make
3141	17	3	you could use
3142	17	3	you didn't say
3143	17	3	you got any
3144	17	3	you have my
3145	17	4	you kind of
3146	17	4	you know a
3147	17	3	you like the
3148	17	4	you put it
3149	17	4	you start to
3150	17	3	you were my
3151	17	3	you're not supposed
3152	17	3	you've got an
3153	16	3	able to take
3154	16	3	allow you to
3155	16	4	and all that

3156	16	4	and in a
3157	16	3	and make it
3158	16	3	and she doesn't
3159	16	3	and there's nothing
3160	16	3	and think about
3161	16	3	and we were
3162	16	3	and what is
3163	16	3	and you got
3164	16	3	are not going
3165	16	4	at some point
3166	16	4	be a very
3167	16	3	be aware of
3168	16	3	be the last
3169	16	3	because i had
3170	16	3	been in the
3171	16	3	blood flow to
3172	16	3	but if it's
3173	16	3	but that's the
3174	16	3	but what i
3175	16	3	can you say
3176	16	3	can't do that
3177	16	3	come back and
3178	16	3	does that tell
3179	16	3	don't be a
3180	16	3	don't do it
3181	16	3	don't have anything
3182	16	3	done with the
3183	16	5	even if they
3184	16	3	even know how
3185	16	3	figure it out
3186	16	3	gave him a
3187	16	3	get a lot
3188	16	4	get all the
3189	16	3	get him back
3190	16	4	giving you a
3191	16	3	going to a
3192	16	3	going to show
3193	16	3	good idea to
3194	16	3	got a little
3195	16	3	got to take
3196	16	3	had a lot
3197	16	3	happens when you
3198	16	3	have to check
3199	16	4	have to use

3200	16	3	he said that
3201	16	3	he was an
3202	16	3	he's just a
3203	16	3	hope for the
3204	16	5	how does it
3205	16	4	how you do
3206	16	3	i decided to
3207	16	3	i have had
3208	16	4	i hope that
3209	16	3	i said it
3210	16	3	i saw her
3211	16	3	i spent the
3212	16	3	i was really
3213	16	3	i'll get a
3214	16	5	i'll show you
3215	16	3	if there are
3216	16	3	if this was
3217	16	5	in addition to
3218	16	3	in the left
3219	16	4	in the whole
3220	16	4	is not just
3221	16	3	is so much
3222	16	3	is still in
3223	16	3	it is for
3224	16	3	it makes it
3225	16	4	it seems that
3226	16	3	it to you
3227	16	3	it was not
3228	16	3	it was very
3229	16	3	it's important that
3230	16	3	it's not in
3231	16	3	it's up to
3232	16	3	just one more
3233	16	3	know if you're
3234	16	3	know that this
3235	16	3	know what she
3236	16	3	know what we're
3237	16	5	let's go back
3238	16	3	live in the
3239	16	3	long do you
3240	16	3	long enough to
3241	16	3	look at me,
3242	16	3	make a good
3243	16	3	make sure the

3244	16	4	may be the
3245	16	3	maybe this is
3246	16	5	most of you
3247	16	4	much of a
3248	16	3	no. i think
3249	16	3	not a lot
3250	16	5	not be a
3251	16	3	of being a
3252	16	5	of the patients
3253	16	4	of the united
3254	16	3	of you is
3255	16	4	of your life
3256	16	3	on it. i
3257	16	3	on the list
3258	16	4	right side of
3259	16	3	see if they
3260	16	3	set up the
3261	16	3	sign of a
3262	16	4	so it is
3263	16	3	so she can
3264	16	3	stand up for
3265	16	3	tell that to
3266	16	4	tend to be
3267	16	4	that is that
3268	16	3	that to the
3269	16	3	that you will
3270	16	3	that's what he
3271	16	3	the need to
3272	16	4	the risks of
3273	16	3	the world to
3274	16	3	there are people
3275	16	5	there are three
3276	16	4	there for a
3277	16	3	they don't get
3278	16	3	they have no
3279	16	4	things that you
3280	16	4	think we need
3281	16	3	this is no
3282	16	4	this is very
3283	16	3	time to do
3284	16	3	to ask for
3285	16	3	to be doing
3286	16	4	to be sure
3287	16	4	to make that

3288	16	3	to put her
3289	16	3	to see him
3290	16	3	too much of
3291	16	3	trying to tell
3292	16	3	was about to
3293	16	3	was on my
3294	16	3	way that i
3295	16	3	we are in
3296	16	4	we can use
3297	16	4	we have been
3298	16	3	we just don't
3299	16	3	we'd have to
3300	16	3	what do they
3301	16	3	what i don't
3302	16	4	what i've been
3303	16	3	when you can
3304	16	3	who needs a
3305	16	3	worried about the
3306	16	3	would you mind
3307	16	3	wouldn't have been
3308	16	5	you a little
3309	16	3	you can see,
3310	16	3	you can stay
3311	16	3	you find a
3312	16	3	you got some
3313	16	3	you gotta be
3314	16	4	you like a
3315	16	3	you might get
3316	16	3	you need more
3317	16	4	you read the
3318	16	4	you think there's
3319	16	3	you will get
3320	16	3	you with a
3321	16	3	you'll need to
3322	16	3	you're a little
3323	16	3	your first day
3324	15	3	a hard time
3325	15	3	a needle in
3326	15	3	able to see
3327	15	3	all of that
3328	15	3	an hour and
3329	15	3	and as a
3330	15	3	and he just
3331	15	3	and i say

3332	15	3	and i went
3333	15	3	and i'm trying
3334	15	4	and if they
3335	15	4	and it looks
3336	15	3	and it's the
3337	15	3	and then a
3338	15	3	and you had
3339	15	3	are gonna have
3340	15	4	are the most
3341	15	4	are the ones
3342	15	4	are we going
3343	15	4	as i can
3344	15	4	at the time
3345	15	3	back to that
3346	15	3	be better than
3347	15	3	be caused by
3348	15	4	be the only
3349	15	5	because it's a
3350	15	4	been a lot
3351	15	3	bleeding in the
3352	15	3	but i won't
3353	15	3	but it's the
3354	15	4	but that is
3355	15	3	but you just
3356	15	3	can see here
3357	15	3	can talk about
3358	15	4	can think about
3359	15	3	can't deal with
3360	15	3	can't think of
3361	15	3	come back in
3362	15	3	come to the
3363	15	3	connected to the
3364	15	3	could also be
3365	15	3	could end up
3366	15	4	could get a
3367	15	3	didn't do anything
3368	15	3	do any of
3369	15	3	do with it.
3370	15	4	don't make it
3371	15	3	don't think it
3372	15	4	end up in
3373	15	4	even if it's
3374	15	3	fell in love
3375	15	3	for a new

3376	15	4	for a second
3377	15	4	for some reason
3378	15	3	from one of
3379	15	3	get on with
3380	15	4	give them a
3381	15	3	give you more
3382	15	4	goes back to
3383	15	5	goes to the
3384	15	3	good news for
3385	15	3	got to tell
3386	15	3	had a big
3387	15	3	had a little
3388	15	3	had to have
3389	15	4	has never been
3390	15	3	has to go
3391	15	3	have to live
3392	15	3	he had an
3393	15	3	he used to
3394	15	3	he's had a
3395	15	4	heart rate is
3396	15	3	here if you
3397	15	3	hope you don't
3398	15	3	how to use
3399	15	3	i am very
3400	15	3	i came up
3401	15	3	i can't stand
3402	15	3	i gave up
3403	15	3	i going to
3404	15	3	i got this
3405	15	4	i guess the
3406	15	3	i help you
3407	15	3	i just told
3408	15	3	i mean you
3409	15	4	i really think
3410	15	3	i said i'm
3411	15	4	i think they're
3412	15	3	i thought this
3413	15	3	i wonder if
3414	15	3	i'm one of
3415	15	3	i've been trying
3416	15	3	if i told
3417	15	4	if one of
3418	15	3	if you'd like
3419	15	3	in a way

3420	15	3	in the back.
3421	15	3	in the front
3422	15	4	is a really
3423	15	3	is about to
3424	15	4	is at the
3425	15	3	is in fact
3426	15	3	is no way
3427	15	3	is out of
3428	15	3	is starting to
3429	15	4	is that in
3430	15	5	is that there
3431	15	3	is this gonna
3432	15	4	is to get
3433	15	4	is to make
3434	15	3	is where you
3435	15	4	it in your
3436	15	3	it under control.
3437	15	3	it wasn't the
3438	15	3	it's only a
3439	15	4	it's part of
3440	15	5	just a couple
3441	15	3	just take it
3442	15	4	know is that
3443	15	3	know that the
3444	15	3	know you can
3445	15	4	likely to get
3446	15	3	live in a
3447	15	3	long time to
3448	15	4	looked at the
3449	15	3	looking out for
3450	15	3	make a decision
3451	15	4	many of the
3452	15	5	many of you
3453	15	3	means you have
3454	15	3	member of the
3455	15	3	more important than
3456	15	3	needs a new
3457	15	3	no one has
3458	15	3	no. i'm just
3459	15	3	not because of
3460	15	3	not gonna go
3461	15	4	not have a
3462	15	3	not what we
3463	15	4	now i don't

3464	15	3	of the people
3465	15	5	of you in
3466	15	3	period of time
3467	15	3	put in the
3468	15	3	results of the
3469	15	3	right now. and
3470	15	3	see how it
3471	15	3	she has the
3472	15	3	she tried to
3473	15	3	sit down and
3474	15	4	so they can
3475	15	3	so when you
3476	15	3	so you do
3477	15	4	some of my
3478	15	4	something that you
3479	15	4	take a little
3480	15	3	talked about the
3481	15	3	talking about your
3482	15	4	that has a
3483	15	3	that i love
3484	15	5	that if we
3485	15	5	that you think
3486	15	3	that's one of
3487	15	3	that's the best
3488	15	3	the chances of
3489	15	3	the day i
3490	15	5	the extent of
3491	15	3	the hospital is
3492	15	3	the length of
3493	15	3	the other way
3494	15	3	the pain is
3495	15	4	the rate of
3496	15	3	the things you
3497	15	4	the time we
3498	15	5	them to be
3499	15	3	there have been
3500	15	4	there is something
3501	15	3	there was something
3502	15	4	thing you can
3503	15	5	think there's a
3504	15	5	think we have
3505	15	3	think you could
3506	15	3	this for a
3507	15	3	this is for

3508	15	4	this one is
3509	15	4	thought that was
3510	15	3	thought you could
3511	15	3	times a day.
3512	15	3	to become a
3513	15	3	to do now
3514	15	3	to do this,
3515	15	3	to know where
3516	15	5	to reduce the
3517	15	4	to understand that
3518	15	3	to wait until
3519	15	3	told you about
3520	15	3	try and get
3521	15	4	two of the
3522	15	3	two units of
3523	15	3	us to be
3524	15	3	waiting for me
3525	15	3	wall of the
3526	15	4	want to put
3527	15	3	want to wait
3528	15	4	was a lot
3529	15	3	was thinking about
3530	15	3	we can all
3531	15	4	we can give
3532	15	3	we could go
3533	15	3	we go back
3534	15	3	we just got
3535	15	3	we still don't
3536	15	3	we won't be
3537	15	4	we'll talk about
3538	15	3	we're out of
3539	15	3	what did the
3540	15	3	what does the
3541	15	4	what happens in
3542	15	3	what i'm saying
3543	15	3	what would be
3544	15	3	what's going on,
3545	15	3	where i am.
3546	15	3	which would be
3547	15	3	white cell count
3548	15	3	would you tell
3549	15	4	you can put
3550	15	3	you get an
3551	15	3	you get it

3552	15	3	you know she
3553	15	3	you know that's
3554	15	3	you really are
3555	15	3	you talking about
3556	15	3	you were talking
3557	15	4	you will find
3558	14	3	a friend of
3559	14	3	a set of
3560	14	3	about what happened
3561	14	3	an awful lot
3562	14	3	and hope for
3563	14	3	and i wasn't
3564	14	3	and if the
3565	14	3	and the fact
3566	14	3	and the last
3567	14	3	and we can't
3568	14	3	and we just
3569	14	3	and what about
3570	14	3	and you won't
3571	14	3	any chance you
3572	14	3	are a little
3573	14	3	are you taking
3574	14	3	as a result
3575	14	5	at least not
3576	14	4	at the bottom
3577	14	3	at this point.
3578	14	3	away from my
3579	14	3	back from the
3580	14	3	back of your
3581	14	3	be a long
3582	14	3	be looking at
3583	14	4	because of this
3584	14	3	because you're a
3585	14	3	blood to the
3586	14	3	but she has
3587	14	4	but that was
3588	14	3	but we are
3589	14	3	but you do
3590	14	3	but you won't
3591	14	3	call it a
3592	14	3	can have the
3593	14	3	come back from
3594	14	3	did i say
3595	14	3	do now is

3596	14	3	does it hurt
3597	14	3	doesn't matter what
3598	14	3	don't feel like
3599	14	3	don't get a
3600	14	3	don't see why
3601	14	3	don't think the
3602	14	3	down here and
3603	14	4	even in the
3604	14	3	first day of
3605	14	3	for me and
3606	14	4	for over a
3607	14	3	from all the
3608	14	3	get her back
3609	14	3	get them out
3610	14	4	get through the
3611	14	4	going to help
3612	14	4	going to let
3613	14	3	going to stop
3614	14	3	got off the
3615	14	3	got on the
3616	14	3	has got to
3617	14	3	have a job
3618	14	3	have done it
3619	14	3	have to think
3620	14	3	he was going
3621	14	3	he's on the
3622	14	5	here is the
3623	14	3	how much do
3624	14	3	i can just
3625	14	3	i can't think
3626	14	3	i do it
3627	14	3	i find it
3628	14	3	i had this
3629	14	3	i happen to
3630	14	3	i just heard
3631	14	3	i know your
3632	14	4	i mean by
3633	14	4	i think there
3634	14	3	i understand you
3635	14	3	i was taking
3636	14	3	i work with
3637	14	4	i'm interested in
3638	14	3	i'm sure he
3639	14	3	i've got it

3640	14	4	i've tried to
3641	14	3	if you make
3642	14	3	if you say
3643	14	3	in and out
3644	14	4	in just a
3645	14	3	in the car
3646	14	3	in the dark.
3647	14	3	in this situation
3648	14	4	is a problem
3649	14	4	is for the
3650	14	3	is that all
3651	14	3	is what it
3652	14	4	it a little
3653	14	3	it and i
3654	14	3	it because i
3655	14	3	it is an
3656	14	3	it when i
3657	14	3	it's a beautiful
3658	14	3	it's better to
3659	14	5	it's easy to
3660	14	3	it. so i
3661	14	3	just focus on
3662	14	3	know what a
3663	14	3	know where to
3664	14	3	know why i'm
3665	14	3	let's start with
3666	14	3	like that in
3667	14	4	lots and lots
3668	14	4	make sure it's
3669	14	4	may have been
3670	14	3	me in my
3671	14	3	mean i don't
3672	14	3	meet you in
3673	14	3	might need to
3674	14	3	most important thing
3675	14	5	most of them
3676	14	3	needed to be
3677	14	3	next thing i
3678	14	3	none of you
3679	14	3	not sure that
3680	14	3	not to get
3681	14	3	not used to
3682	14	3	now i can't
3683	14	4	now we can

3684	14	4	of the blood
3685	14	4	of the cancer
3686	14	3	of them is
3687	14	3	one in a
3688	14	3	one wants to
3689	14	3	or if you
3690	14	5	order to get
3691	14	4	over the last
3692	14	3	over the past
3693	14	3	over the place.
3694	14	3	over to the
3695	14	3	people who don't
3696	14	3	right now and
3697	14	3	seems like a
3698	14	3	seems to have
3699	14	3	she had no
3700	14	3	she was trying
3701	14	3	she's allergic to
3702	14	3	so he was
3703	14	3	so it doesn't
3704	14	3	so that the
3705	14	4	so what i'm
3706	14	3	so you can't
3707	14	3	so you just
3708	14	3	some of that
3709	14	3	something that we
3710	14	3	sort of thing
3711	14	4	stuck in the
3712	14	4	sure that the
3713	14	5	tell you a
3714	14	3	that if i
3715	14	5	that it would
3716	14	3	that it's a
3717	14	3	that she had
3718	14	3	that we would
3719	14	3	that's all that
3720	14	4	that's quite a
3721	14	5	that's what we're
3722	14	3	that's why it's
3723	14	4	the absence of
3724	14	3	the best we
3725	14	3	the blood pressure
3726	14	3	the guy with
3727	14	3	the heart rate

3728	14	3	the in the
3729	14	3	the last six
3730	14	4	the loss of
3731	14	3	the nature of
3732	14	3	the person you
3733	14	4	the problem with
3734	14	3	the surgery is
3735	14	3	the time of
3736	14	4	them to do
3737	14	3	there anything else
3738	14	4	there is another
3739	14	3	there were no
3740	14	3	there's a good
3741	14	3	there's always a
3742	14	3	there's no time
3743	14	3	there's nothing you
3744	14	3	there's something wrong
3745	14	3	there's still a
3746	14	4	they have the
3747	14	3	thing to do
3748	14	3	thing to say
3749	14	3	think i might
3750	14	3	think it is
3751	14	3	think of the
3752	14	3	think you know
3753	14	3	thought i would
3754	14	3	thought it might
3755	14	3	to be done
3756	14	4	to be that
3757	14	4	to come down
3758	14	3	to do and
3759	14	4	to each other
3760	14	3	to fix the
3761	14	3	to give a
3762	14	3	to live in
3763	14	3	to me and
3764	14	3	to put the
3765	14	3	to take them
3766	14	3	to the point
3767	14	3	told him that
3768	14	3	told me about
3769	14	3	too much for
3770	14	4	unless you have
3771	14	4	very good at

3772	14	3	wants to get
3773	14	3	wants us to
3774	14	4	was on a
3775	14	3	was talking about
3776	14	3	way to do
3777	14	3	way up to
3778	14	4	way you can
3779	14	3	we can still
3780	14	3	we don't really
3781	14	3	we have some
3782	14	4	we know it's
3783	14	4	we took a
3784	14	3	we'll let you
3785	14	4	were on the
3786	14	4	what is a
3787	14	4	what it looks
3788	14	4	when you come
3789	14	3	when you wake
3790	14	4	where you can
3791	14	3	where you get
3792	14	3	white blood cells
3793	14	3	whole bunch of
3794	14	3	why aren't we
3795	14	4	will always be
3796	14	3	will take care
3797	14	3	with one of
3798	14	3	would be more
3799	14	3	would have had
3800	14	3	you can't say
3801	14	3	you come up
3802	14	3	you do for
3803	14	3	you do have
3804	14	4	you end up
3805	14	3	you find the
3806	14	3	you get this
3807	14	4	you go in
3808	14	3	you go out
3809	14	3	you have got
3810	14	3	you just take
3811	14	4	you like me
3812	14	3	you might need
3813	14	3	you needed to
3814	14	4	you put in
3815	14	3	you think they're

3816	14	3	you to look
3817	14	3	you try and
3818	14	3	you were all
3819	14	3	you will do
3820	14	5	you will see
3821	14	3	you would like
3822	14	3	you'd want to
3823	14	3	you've got no
3824	14	3	you've got some
3825	13	3	'cause if you
3826	13	3	a bag of
3827	13	4	a cure for
3828	13	3	a great deal
3829	13	3	a little bit,
3830	13	3	a long way
3831	13	3	a much more
3832	13	3	a whole bunch
3833	13	3	about it in
3834	13	3	about what you
3835	13	3	alone in the
3836	13	3	an increase in
3837	13	3	and as you
3838	13	5	and at the
3839	13	4	and because of
3840	13	3	and get out
3841	13	3	and get you
3842	13	4	and he said
3843	13	3	and i've got
3844	13	3	and it would
3845	13	3	and let the
3846	13	3	and that the
3847	13	3	and that will
3848	13	3	and that would
3849	13	3	and the one
3850	13	4	and the way
3851	13	4	and then it's
3852	13	3	and they can
3853	13	3	and you haven't
3854	13	3	and you might
3855	13	3	and you're just
3856	13	4	appear to be
3857	13	4	as i was
3858	13	4	as part of
3859	13	3	because we don't

3860	13	3	before you get
3861	13	3	before you know
3862	13	3	believe this is
3863	13	5	blood pressure is
3864	13	3	blood supply to
3865	13	3	but at the
3866	13	3	but i've got
3867	13	3	but if the
3868	13	3	but it didn't
3869	13	3	but the only
3870	13	4	but we do
3871	13	3	can help me
3872	13	3	can take the
3873	13	4	cancer cells are
3874	13	3	coming back to
3875	13	3	could i have
3876	13	4	could tell you
3877	13	3	couple of days
3878	13	4	couple of years
3879	13	3	did he just
3880	13	3	didn't think it
3881	13	3	do not need
3882	13	3	do that for
3883	13	4	do you still
3884	13	3	does anybody know
3885	13	3	does this mean
3886	13	3	don't do this
3887	13	3	don't even have
3888	13	3	don't give up
3889	13	3	don't know whether
3890	13	4	don't think this
3891	13	4	done in the
3892	13	4	enough to be
3893	13	3	even with the
3894	13	3	family history of
3895	13	3	for one of
3896	13	3	from the fact
3897	13	3	get to make
3898	13	3	get to your
3899	13	3	getting a little
3900	13	3	give her some
3901	13	4	go through that
3902	13	3	going through the
3903	13	3	going to use

3904	13	4	going to work
3905	13	3	gonna look at
3906	13	3	has no idea
3907	13	3	have got to
3908	13	3	have the right
3909	13	3	he came to
3910	13	3	he can get
3911	13	3	he might have
3912	13	3	he's still in
3913	13	3	heard about the
3914	13	4	here in a
3915	13	4	him to get
3916	13	3	how could i
3917	13	3	how long did
3918	13	4	i have done
3919	13	3	i just say
3920	13	3	i just wish
3921	13	3	i might as
3922	13	3	i put a
3923	13	3	i put in
3924	13	3	i put it
3925	13	3	i think i'll
3926	13	4	i think of
3927	13	4	i think we've
3928	13	3	i want them
3929	13	4	i went through
3930	13	4	i'll take you
3931	13	3	if there's no
3932	13	4	if you haven't
3933	13	3	if you remember
3934	13	5	in order for
3935	13	3	in the clinic
3936	13	3	in the early
3937	13	3	in the heart
3938	13	4	in the history
3939	13	3	in the in
3940	13	3	in the long
3941	13	3	in this case,
3942	13	3	is it so
3943	13	4	is sort of
3944	13	4	is that for
3945	13	3	isn't going to
3946	13	3	isn't that the
3947	13	3	it and then

3948	13	3	it and you
3949	13	3	it back to
3950	13	3	it because you
3951	13	3	it is possible
3952	13	3	it means the
3953	13	3	it on a
3954	13	3	it starts to
3955	13	3	it to your
3956	13	3	it up and
3957	13	3	it's a new
3958	13	3	it's a nice
3959	13	3	it's a simple
3960	13	3	it's not really
3961	13	3	it's probably a
3962	13	4	it's sort of
3963	13	3	just got to
3964	13	3	just in case
3965	13	4	just sort of
3966	13	3	kind of person
3967	13	3	like any other
3968	13	3	look at it.
3969	13	3	look for the
3970	13	5	look in the
3971	13	3	make sure they
3972	13	3	may want to
3973	13	3	me and i
3974	13	3	much do you
3975	13	5	need to look
3976	13	3	never going to
3977	13	3	no such thing
3978	13	4	not looking at
3979	13	4	now in the
3980	13	4	of a lot
3981	13	3	of blood in
3982	13	3	of course they
3983	13	3	of the time,
3984	13	4	of the tumor
3985	13	4	of them have
3986	13	4	of those things
3987	13	4	of us who
3988	13	3	on their way
3989	13	3	one hell of
3990	13	3	one of two
3991	13	3	one on the

3992	13	3	only thing you
3993	13	3	out all the
3994	13	4	out with the
3995	13	4	people in this
3996	13	3	percent of the
3997	13	3	pick up a
3998	13	3	put on the
3999	13	4	put them on
4000	13	3	put you in
4001	13	4	quality of life
4002	13	3	rest of it
4003	13	3	rest of you
4004	13	3	right at the
4005	13	3	said that i
4006	13	3	see if there's
4007	13	3	she had an
4008	13	3	show up in
4009	13	5	show you a
4010	13	3	so i would
4011	13	3	so now you
4012	13	3	so tell me
4013	13	4	so that's the
4014	13	3	so we should
4015	13	3	so we're just
4016	13	3	so you should
4017	13	3	something else to
4018	13	4	something in the
4019	13	4	sort of the
4020	13	3	starting to get
4021	13	3	such a big
4022	13	3	surface of the
4023	13	3	talk about it,
4024	13	3	thank you. so
4025	13	3	that all of
4026	13	3	that he has
4027	13	4	that i'm going
4028	13	3	that is going
4029	13	3	that no one
4030	13	3	that one of
4031	13	5	that they can
4032	13	4	that they're not
4033	13	3	that's a big
4034	13	3	that's when you
4035	13	4	that's where the

4036	13	3	the bad news
4037	13	3	the base of
4038	13	3	the doctor who
4039	13	5	the first two
4040	13	3	the important thing
4041	13	4	the lack of
4042	13	3	the one that's
4043	13	3	the ones you
4044	13	3	the patient and
4045	13	5	them in a
4046	13	4	there are things
4047	13	5	there would be
4048	13	3	there's a very
4049	13	3	there's no other
4050	13	3	there's not much
4051	13	4	they may have
4052	13	3	they're in the
4053	13	3	they've got a
4054	13	4	think if you
4055	13	3	think it might
4056	13	3	think that was
4057	13	3	think that's what
4058	13	4	this has to
4059	13	4	this in the
4060	13	5	this is actually
4061	13	3	this is something
4062	13	4	this to the
4063	13	4	time i was
4064	13	3	to any of
4065	13	3	to build a
4066	13	3	to explain to
4067	13	3	to tell us
4068	13	3	turn off the
4069	13	3	very important to
4070	13	5	was a very
4071	13	4	was not a
4072	13	3	way that you
4073	13	3	we are doing
4074	13	3	we can't take
4075	13	4	we have two
4076	13	3	we make a
4077	13	4	we needed to
4078	13	4	we'll be able
4079	13	3	we'll have a

4080	13	3	what about his
4081	13	3	what are your
4082	13	3	what i had
4083	13	3	what i told
4084	13	4	what i'm trying
4085	13	4	what we call
4086	13	3	whatever you want
4087	13	3	when i said
4088	13	5	when you look
4089	13	4	when you see
4090	13	3	will no longer
4091	13	3	you a lot
4092	13	3	you all have
4093	13	3	you are doing
4094	13	4	you can ask
4095	13	4	you come back
4096	13	3	you did what
4097	13	3	you don't make
4098	13	4	you from the
4099	13	4	you get all
4100	13	3	you have done
4101	13	3	you have one
4102	13	3	you just do
4103	13	4	you what i
4104	13	3	you who are
4105	13	3	you would expect
4106	12	3	a case of
4107	12	4	a change in
4108	12	4	a clinical trial
4109	12	4	a copy of
4110	12	3	a day or
4111	12	5	a disease that
4112	12	3	a glass of
4113	12	3	a moment to
4114	12	3	a point of
4115	12	3	a really nice
4116	12	3	a reason for
4117	12	3	a year and
4118	12	3	air in the
4119	12	3	almost impossible to
4120	12	3	and a little
4121	12	3	and i'd like
4122	12	3	and it all
4123	12	3	and it's all

4124	12	3	and not just
4125	12	3	and she didn't
4126	12	4	and so you
4127	12	4	and the blood
4128	12	3	and then what
4129	12	3	and we could
4130	12	4	and we know
4131	12	3	and what are
4132	12	3	and what does
4133	12	3	and when he
4134	12	3	and why do
4135	12	3	and you tell
4136	12	3	and you're going
4137	12	3	any of that
4138	12	3	anything else you
4139	12	4	are no longer
4140	12	3	as good a
4141	12	3	be talking to
4142	12	4	be the most
4143	12	3	because if i
4144	12	3	because it is
4145	12	3	because there's a
4146	12	3	been a little
4147	12	3	before you go
4148	12	4	but for the
4149	12	3	but he is
4150	12	3	but i could
4151	12	3	but i'm sure
4152	12	3	but it does
4153	12	3	but it's gonna
4154	12	3	but when you
4155	12	4	but you've got
4156	12	3	can come up
4157	12	4	change in the
4158	12	3	comes out of
4159	12	4	compared to the
4160	12	4	could talk about
4161	12	3	did you ask
4162	12	3	did you take
4163	12	3	different from the
4164	12	3	do is to
4165	12	3	do it with
4166	12	5	do this in
4167	12	4	doesn't matter if

4168	12	3	don't know. we
4169	12	3	don't take it
4170	12	3	each other and
4171	12	3	even a little
4172	12	3	everything else is
4173	12	4	fact that we
4174	12	3	flow to the
4175	12	3	get away with
4176	12	3	go and get
4177	12	4	go for the
4178	12	3	going to hurt
4179	12	3	gonna ask me
4180	12	3	gonna talk about
4181	12	3	got all the
4182	12	3	got in the
4183	12	4	has been a
4184	12	3	have a big
4185	12	3	have made a
4186	12	3	have to decide
4187	12	3	have to explain
4188	12	3	he had no
4189	12	3	he has the
4190	12	3	he was still
4191	12	3	he went into
4192	12	3	he's a little
4193	12	3	him up and
4194	12	3	i asked her
4195	12	3	i can't hear
4196	12	3	i just took
4197	12	3	i live in
4198	12	3	i make a
4199	12	3	i mean the
4200	12	3	i say that
4201	12	4	i think a
4202	12	3	i will find
4203	12	3	i'll take a
4204	12	3	i'm at the
4205	12	3	i'm gonna say
4206	12	3	i'm just making
4207	12	3	i'm not telling
4208	12	3	i'm sure you're
4209	12	3	i've been thinking
4210	12	3	i've got no
4211	12	3	if any of

4212	12	3	if there were
4213	12	4	if we find
4214	12	4	if you're in
4215	12	3	in a long
4216	12	3	in a row.
4217	12	4	in all the
4218	12	4	in fact the
4219	12	4	in the process
4220	12	4	in the wrong
4221	12	3	in with the
4222	12	5	interested in the
4223	12	3	is a bit
4224	12	5	is a nice
4225	12	4	is all about
4226	12	3	is important to
4227	12	3	is it about
4228	12	3	is looking at
4229	12	3	is not in
4230	12	3	is that an
4231	12	3	is that if
4232	12	3	is that it's
4233	12	3	is that so
4234	12	3	is that they
4235	12	3	is the reason
4236	12	3	it comes down
4237	12	3	it into a
4238	12	3	it into the
4239	12	3	it turns out,
4240	12	3	it was about
4241	12	3	it wasn't just
4242	12	3	it's all right
4243	12	3	it's all the
4244	12	4	it's coming from
4245	12	3	it's got a
4246	12	3	it's what we
4247	12	3	it, and you
4248	12	3	just for the
4249	12	3	just get the
4250	12	5	just have a
4251	12	3	just take the
4252	12	3	just talking about
4253	12	3	know i don't
4254	12	3	know where the
4255	12	5	let's look at

4256	12	3	like it or
4257	12	3	look a little
4258	12	3	look, this is
4259	12	3	lot of blood
4260	12	3	may need to
4261	12	4	me at the
4262	12	3	more complicated than
4263	12	4	more than the
4264	12	4	most of these
4265	12	4	most of us
4266	12	3	need to give
4267	12	4	need to learn
4268	12	3	not all of
4269	12	3	not having a
4270	12	4	not that we
4271	12	3	not to do
4272	12	3	nothing more than
4273	12	3	now, if you
4274	12	3	of course we
4275	12	5	of it is
4276	12	3	of that is
4277	12	3	of the disease
4278	12	3	of the heart.
4279	12	4	of the same
4280	12	3	of the time.
4281	12	4	of them were
4282	12	3	of these people
4283	12	3	of things that
4284	12	3	okay. this is
4285	12	3	one has to
4286	12	3	or in the
4287	12	5	out into the
4288	12	3	out of there
4289	12	3	over the next
4290	12	4	part of this
4291	12	5	problem is that
4292	12	4	put on your
4293	12	3	reaction to the
4294	12	3	really wanted to
4295	12	3	red blood cells
4296	12	3	remember what i
4297	12	4	said he had
4298	12	3	said that they
4299	12	3	she had the

4300	12	3	she might be
4301	12	3	she might have
4302	12	3	so it was
4303	12	3	so let me
4304	12	3	so there's no
4305	12	5	so we're going
4306	12	3	so what did
4307	12	3	so why is
4308	12	3	so you and
4309	12	3	so you were
4310	12	3	some of your
4311	12	3	soon as the
4312	12	3	start off with
4313	12	4	start with a
4314	12	4	still be able
4315	12	3	still don't know
4316	12	3	still working on
4317	12	3	take it to
4318	12	3	tell us about
4319	12	3	that and the
4320	12	4	that for a
4321	12	4	that have been
4322	12	3	that is why
4323	12	3	that she can
4324	12	3	that the immune
4325	12	3	that they have
4326	12	3	that we don't
4327	12	3	that's a little
4328	12	3	that's a very
4329	12	4	that's the first
4330	12	3	that's why the
4331	12	5	the body and
4332	12	3	the damage to
4333	12	3	the day you
4334	12	3	the father of
4335	12	4	the first day
4336	12	4	the last one
4337	12	3	the next step
4338	12	4	the opposite of
4339	12	4	the past three
4340	12	3	the patient doesn't
4341	12	4	the patients who
4342	12	4	the treatment of
4343	12	3	the way down

4344	12	3	the whole story.
4345	12	3	them at the
4346	12	3	then we'll have
4347	12	3	then why is
4348	12	3	then you do
4349	12	3	there any questions
4350	12	3	there are certain
4351	12	3	there can be
4352	12	4	there's a little
4353	12	4	there's a whole
4354	12	5	they do is
4355	12	4	they had a
4356	12	4	they should be
4357	12	3	they think they
4358	12	4	thing is that
4359	12	3	think i'm going
4360	12	3	think you might
4361	12	3	thinking about the
4362	12	3	this case is
4363	12	3	this isn't just
4364	12	3	this type of
4365	12	3	to be aware
4366	12	3	to be careful
4367	12	3	to be her
4368	12	4	to change the
4369	12	3	to check the
4370	12	3	to give the
4371	12	3	to give us
4372	12	3	to have one
4373	12	3	to look like
4374	12	3	to make your
4375	12	3	to the patient.
4376	12	3	to the same
4377	12	3	trauma to the
4378	12	3	try to save
4379	12	3	trying to give
4380	12	3	trying to understand
4381	12	4	turn on the
4382	12	3	up and down
4383	12	3	us in the
4384	12	3	want to check
4385	12	3	want to move
4386	12	3	want to share
4387	12	3	was a time

4388	12	3	was the best
4389	12	4	we did it
4390	12	3	we shouldn't be
4391	12	3	we were on
4392	12	3	we're dealing with
4393	12	3	we're looking for
4394	12	3	we're working on
4395	12	3	we've got an
4396	12	3	what did they
4397	12	3	what it means
4398	12	3	what you're going
4399	12	4	what's going to
4400	12	3	what's the point
4401	12	4	when i saw
4402	12	3	when you first
4403	12	4	when you go
4404	12	3	where is the
4405	12	3	which i think
4406	12	4	which is really
4407	12	3	which you can
4408	12	3	who don't know
4409	12	3	whole lot of
4410	12	3	why are they
4411	12	3	why do people
4412	12	4	why is that?
4413	12	3	with a little
4414	12	4	with the fact
4415	12	3	work as a
4416	12	3	worry about that
4417	12	3	you could say
4418	12	4	you don't actually
4419	12	3	you don't tell
4420	12	4	you find that
4421	12	3	you give a
4422	12	4	you go through
4423	12	3	you haven't heard
4424	12	3	you if i
4425	12	3	you know and
4426	12	3	you let them
4427	12	4	you open your
4428	12	5	you really have
4429	12	3	you see how
4430	12	3	you take your
4431	12	3	you tell the

4432	12	3	you up for
4433	12	3	you were saying
4434	12	3	you're doing it
4435	12	3	you're with me
4436	12	3	your ability to
4437	11	3	a doctor in
4438	11	3	a few of
4439	11	3	a good job
4440	11	4	a patient of
4441	11	3	a pile of
4442	11	3	a very bad
4443	11	5	a way that
4444	11	3	able to find
4445	11	3	able to stop
4446	11	4	all of them
4447	11	4	all of us
4448	11	4	an hour or
4449	11	4	and as i
4450	11	3	and as soon
4451	11	3	and his heart
4452	11	3	and in this
4453	11	3	and it has
4454	11	3	and it just
4455	11	3	and she's a
4456	11	3	and so if
4457	11	3	and that's when
4458	11	4	and then that
4459	11	3	and they go
4460	11	3	and they've got
4461	11	3	and we're not
4462	11	3	and why is
4463	11	3	any of them
4464	11	3	are all the
4465	11	3	as opposed to
4466	11	3	at least it
4467	11	5	at the right
4468	11	3	based on a
4469	11	3	be in that
4470	11	3	be working on
4471	11	3	because he's a
4472	11	3	because there's no
4473	11	4	because they don't
4474	11	3	because when you
4475	11	3	been working with

4476	11	3	big part of
4477	11	3	but if she
4478	11	3	but if they
4479	11	3	but in a
4480	11	3	but it will
4481	11	3	but on the
4482	11	3	but they're not
4483	11	3	but we were
4484	11	3	can get it
4485	11	3	come here to
4486	11	3	coming to the
4487	11	3	could be in
4488	11	3	could you just
4489	11	3	didn't have the
4490	11	3	do for a
4491	11	4	do know that
4492	11	4	do that in
4493	11	3	do to get
4494	11	3	do what we
4495	11	3	do when you
4496	11	3	does that have
4497	11	3	doesn't have the
4498	11	3	don't get the
4499	11	3	don't like the
4500	11	3	don't need me
4501	11	3	don't think about
4502	11	3	don't try and
4503	11	4	done in a
4504	11	3	each of the
4505	11	3	either that or
4506	11	3	ever happened to
4507	11	3	exactly what you
4508	11	4	expect to see
4509	11	4	first thing that
4510	11	3	for a lot
4511	11	3	for all the
4512	11	3	for the fact
4513	11	3	for the whole
4514	11	3	for those of
4515	11	3	for years, and
4516	11	3	get one of
4517	11	4	get up to
4518	11	3	getting in the
4519	11	3	going to put

4520	11	3	got a few
4521	11	3	got it under
4522	11	3	had to put
4523	11	3	has a big
4524	11	3	has something to
4525	11	4	has to have
4526	11	3	have a meeting
4527	11	4	have at least
4528	11	3	have been so
4529	11	3	have to change
4530	11	3	he was so
4531	11	3	here are the
4532	11	4	here is that
4533	11	3	here's what i
4534	11	3	him to a
4535	11	3	his blood pressure
4536	11	4	how do they
4537	11	3	how it's gonna
4538	11	3	how long would
4539	11	4	how much time
4540	11	3	i do what
4541	11	3	i don't buy
4542	11	3	i get out
4543	11	3	i haven't done
4544	11	4	i know where
4545	11	3	i made you
4546	11	3	i put my
4547	11	3	i really wanted
4548	11	3	i said that
4549	11	3	i saw him
4550	11	3	i talked about
4551	11	5	i think in
4552	11	3	i think my
4553	11	3	i think your
4554	11	3	i use the
4555	11	3	i was your
4556	11	4	i went back
4557	11	3	i wondered if
4558	11	3	i'm saying is
4559	11	3	i'm sorry i'm
4560	11	4	i'm sure it
4561	11	3	i'm thinking of
4562	11	3	i'm working on
4563	11	3	if it comes

4564	11	3	if it had
4565	11	3	if it's all
4566	11	4	if we are
4567	11	4	if we just
4568	11	3	if we want
4569	11	5	if you put
4570	11	3	if you will
4571	11	3	in a different
4572	11	3	in a patient
4573	11	3	in her left
4574	11	3	in my head
4575	11	4	in my own
4576	11	3	in some ways
4577	11	3	in the sense
4578	11	3	in the world,
4579	11	3	in to the
4580	11	3	is a real
4581	11	3	is down to
4582	11	3	is how you
4583	11	4	is it possible
4584	11	3	is that how
4585	11	3	is what is
4586	11	3	is what you're
4587	11	3	it doesn't seem
4588	11	4	it doesn't work
4589	11	4	it is in
4590	11	3	it needs to
4591	11	3	it turned out
4592	11	3	it was for
4593	11	3	it's all in
4594	11	3	it's easier to
4595	11	3	it's never been
4596	11	3	it's not something
4597	11	3	it. what is
4598	11	3	just the way
4599	11	3	just wondering if
4600	11	3	know what your
4601	11	3	know when i
4602	11	3	know you can't
4603	11	3	let me put
4604	11	3	like a little
4605	11	5	look at these
4606	11	3	looks like it
4607	11	3	lot of work

4608	11	3	may be able
4609	11	4	may or may
4610	11	3	maybe it's a
4611	11	3	means that the
4612	11	4	not a real
4613	11	4	not at all
4614	11	3	not have to
4615	11	3	not really a
4616	11	3	not sure if
4617	11	3	not the same.
4618	11	3	not the way
4619	11	3	now what do
4620	11	3	now you don't
4621	11	4	now you need
4622	11	4	of like a
4623	11	3	of the left
4624	11	4	of the time
4625	11	4	of things to
4626	11	3	of you and
4627	11	3	off with a
4628	11	3	on his left
4629	11	3	on the next
4630	11	3	on the third
4631	11	3	one thing to
4632	11	4	one way or
4633	11	4	only in the
4634	11	3	only one that
4635	11	3	or any other
4636	11	3	order to do
4637	11	4	out in a
4638	11	3	out there and
4639	11	4	out with a
4640	11	4	outside of the
4641	11	3	part of her
4642	11	4	part of his
4643	11	4	patient in the
4644	11	4	pay attention to
4645	11	3	plenty of other
4646	11	4	point of view
4647	11	3	ran out of
4648	11	4	related to the
4649	11	3	response to the
4650	11	3	right now i'm
4651	11	3	right through the

4652	11	4	said there was
4653	11	3	say that to
4654	11	3	seem to have
4655	11	3	she does have
4656	11	3	shot in the
4657	11	4	should be doing
4658	11	3	shouldn't be a
4659	11	3	sit in the
4660	11	3	size of the
4661	11	3	so at the
4662	11	3	so if it's
4663	11	4	so now we
4664	11	3	so that means
4665	11	3	so why do
4666	11	4	sometimes you have
4667	11	4	sort of like
4668	11	3	stick to the
4669	11	3	still on the
4670	11	4	sure that i
4671	11	3	talk to a
4672	11	4	talking to a
4673	11	3	tends to be
4674	11	3	that came out
4675	11	3	that he can
4676	11	3	that he didn't
4677	11	3	that i've been
4678	11	4	that in your
4679	11	4	that it could
4680	11	3	that must be
4681	11	4	that needs to
4682	11	3	that she is
4683	11	3	that was an
4684	11	3	that we should
4685	11	3	that what this
4686	11	4	that you did
4687	11	3	that you're going
4688	11	3	that's a bad
4689	11	3	that's how we
4690	11	3	the best that
4691	11	3	the blood and
4692	11	4	the blood flow
4693	11	3	the blood in
4694	11	3	the body of
4695	11	3	the chance of

4696	11	3	the day before
4697	11	3	the death of
4698	11	5	the effects of
4699	11	3	the fact is
4700	11	3	the gold standard
4701	11	3	the last five
4702	11	4	the list of
4703	11	3	the lungs and
4704	11	4	the more you
4705	11	3	the one on
4706	11	3	the patient had
4707	11	3	the question is
4708	11	4	the reason for
4709	11	3	the reason that
4710	11	3	the sense that
4711	11	3	them on the
4712	11	4	them to the
4713	11	4	then we have
4714	11	3	there are the
4715	11	3	there are very
4716	11	3	there's not a
4717	11	3	there's not enough
4718	11	3	these are all
4719	11	4	these people are
4720	11	4	they could be
4721	11	5	they didn't have
4722	11	4	they were all
4723	11	5	things that we
4724	11	3	think i would
4725	11	3	this patient is
4726	11	3	through to the
4727	11	3	to be honest
4728	11	4	to be one
4729	11	3	to be talking
4730	11	3	to change your
4731	11	3	to check for
4732	11	3	to control the
4733	11	3	to cut into
4734	11	4	to do things
4735	11	3	to do, and
4736	11	3	to end up
4737	11	3	to get all
4738	11	3	to keep them
4739	11	3	to kill a

4740	11	3	to know a
4741	11	3	to make up
4742	11	3	to pick a
4743	11	3	to put it
4744	11	3	to say is
4745	11	3	to the fact
4746	11	4	to you is
4747	11	4	took her to
4748	11	4	top of your
4749	11	4	try to do
4750	11	4	up into the
4751	11	3	want to show
4752	11	3	want to sit
4753	11	3	was just thinking
4754	11	3	was like a
4755	11	3	was looking at
4756	11	3	was that a
4757	11	4	way of saying
4758	11	5	way that we
4759	11	3	we are so
4760	11	3	we are the
4761	11	3	we didn't do
4762	11	3	we do have
4763	11	3	we find out
4764	11	3	we get it
4765	11	4	we know about
4766	11	3	we were going
4767	11	3	we were trying
4768	11	3	we'll do the
4769	11	5	well as the
4770	11	3	were out of
4771	11	3	what she did
4772	11	3	what the patient
4773	11	3	what they do
4774	11	3	what we are
4775	11	4	when they were
4776	11	3	when you take
4777	11	4	who has a
4778	11	3	who should be
4779	11	3	why you have
4780	11	3	will have a
4781	11	4	with the same
4782	11	3	work in a
4783	11	3	would you please

4784	11	3	you all know
4785	11	3	you can either
4786	11	3	you can leave
4787	11	3	you can sit
4788	11	3	you could look
4789	11	3	you could see
4790	11	3	you don't go
4791	11	3	you get in
4792	11	3	you give it
4793	11	3	you go back
4794	11	3	you heard about
4795	11	3	you heard of
4796	11	3	you hit the
4797	11	4	you in this
4798	11	3	you just get
4799	11	3	you know he
4800	11	4	you know they
4801	11	3	you know your
4802	11	3	you know, there's
4803	11	3	you leave the
4804	11	3	you like it
4805	11	3	you like your
4806	11	3	you might like
4807	11	3	you need the
4808	11	3	you understand what
4809	11	3	you wanna do
4810	11	3	you were to
4811	11	3	you'll have a
4812	11	3	you're doing this
4813	11	3	you're on a
4814	10	3	a change of
4815	10	3	a decision that
4816	10	3	a good chance
4817	10	3	a lot in
4818	10	4	a much better
4819	10	3	a pretty big
4820	10	3	a time when
4821	10	4	able to make
4822	10	3	about a little
4823	10	3	about a year
4824	10	4	about all the
4825	10	3	about it, but
4826	10	3	about the other
4827	10	3	all the rest

4828	10	3	all the things
4829	10	3	all three of
4830	10	3	all you need
4831	10	4	allow me to
4832	10	3	an infection in
4833	10	3	and as long
4834	10	5	and at this
4835	10	3	and find out
4836	10	3	and he'll be
4837	10	3	and i shouldn't
4838	10	4	and if that
4839	10	3	and put it
4840	10	3	and so are
4841	10	3	and that makes
4842	10	3	and that's where
4843	10	3	and the patient
4844	10	3	and the thing
4845	10	3	and they're not
4846	10	3	and you made
4847	10	3	and you told
4848	10	3	any number of
4849	10	3	any of my
4850	10	3	anything else i
4851	10	3	are people who
4852	10	4	area of the
4853	10	3	as much of
4854	10	3	as we can
4855	10	3	as we get
4856	10	3	as well just
4857	10	3	as you get
4858	10	4	at all the
4859	10	3	at least two
4860	10	3	at me with
4861	10	3	at risk of
4862	10	4	back to his
4863	10	4	back to you
4864	10	3	be a way
4865	10	3	be done in
4866	10	3	be so much
4867	10	3	be some kind
4868	10	3	be something else.
4869	10	3	because if it
4870	10	3	because of his
4871	10	3	been doing this

4872	10	4	been going on
4873	10	3	before i go
4874	10	3	best we can
4875	10	3	better than the
4876	10	4	bit of the
4877	10	3	bleeding into his
4878	10	3	blood vessels that
4879	10	3	bottom of the
4880	10	3	but i couldn't
4881	10	3	but i still
4882	10	3	but i've never
4883	10	4	but in this
4884	10	3	but it also
4885	10	3	but these are
4886	10	4	but they do
4887	10	3	but what about
4888	10	4	but when the
4889	10	3	call it the
4890	10	3	can cause a
4891	10	3	can find the
4892	10	3	can get back
4893	10	3	can go and
4894	10	3	can put it
4895	10	4	can see how
4896	10	4	can't see the
4897	10	4	control of the
4898	10	3	could at least
4899	10	3	could be caused
4900	10	4	could do the
4901	10	3	could go to
4902	10	3	could have gone
4903	10	3	difference between the
4904	10	4	do it because
4905	10	3	do something to
4906	10	3	do you take
4907	10	3	does that sound
4908	10	3	doing this for
4909	10	3	don't get any
4910	10	3	don't go to
4911	10	4	don't have enough
4912	10	3	don't see how
4913	10	3	don't think of
4914	10	4	due to a
4915	10	3	end of this

4916	10	3	even though i
4917	10	3	even try to
4918	10	3	find out where
4919	10	4	find out why
4920	10	3	focused on the
4921	10	3	found out that
4922	10	3	gave us a
4923	10	3	getting out of
4924	10	3	gives you an
4925	10	3	go in the
4926	10	3	go into a
4927	10	4	go through all
4928	10	4	going to call
4929	10	3	going to keep
4930	10	4	had a very
4931	10	3	have a great
4932	10	3	have more than
4933	10	3	he is an
4934	10	3	he was having
4935	10	4	heard of the
4936	10	3	her blood pressure
4937	10	4	here at the
4938	10	3	here because i
4939	10	3	here is a
4940	10	3	how is it
4941	10	3	how many patients
4942	10	3	how much is
4943	10	3	how should i
4944	10	3	i became a
4945	10	3	i beg your
4946	10	3	i can assure
4947	10	3	i can never
4948	10	3	i do that
4949	10	3	i hope it
4950	10	3	i like it
4951	10	3	i said we
4952	10	5	i think one
4953	10	4	i think people
4954	10	3	i thought about
4955	10	3	i wasn't sure
4956	10	3	i'd be more
4957	10	3	i'll take that
4958	10	3	if i ever
4959	10	3	if i start

4960	10	3	if it gets
4961	10	3	if it goes
4962	10	3	if they can
4963	10	3	if we give
4964	10	4	if we're going
4965	10	3	if you believe
4966	10	3	if you cut
4967	10	3	if you try
4968	10	3	if you're just
4969	10	4	immune system and
4970	10	3	in in a
4971	10	3	in response to
4972	10	3	in such a
4973	10	3	in the best
4974	10	3	in the chest
4975	10	3	in the country
4976	10	3	in the dark
4977	10	4	in the real
4978	10	3	in the second
4979	10	3	in the upper
4980	10	3	in time for
4981	10	3	in time to
4982	10	3	in touch with
4983	10	3	into one of
4984	10	4	is actually a
4985	10	4	is also a
4986	10	3	is causing the
4987	10	4	is still a
4988	10	4	is where i
4989	10	3	is why we're
4990	10	3	it doesn't really
4991	10	3	it from a
4992	10	3	it out in
4993	10	4	it really is
4994	10	3	it was more
4995	10	4	it was only
4996	10	3	it's a bit
4997	10	3	it's a real
4998	10	3	it's a waste
4999	10	4	it's actually a
5000	10	3	it's all over
5001	10	3	it's exactly what
5002	10	3	it's on the
5003	10	3	just can't get

5004	10	3	just do it
5005	10	3	just say the
5006	10	3	just think of
5007	10	3	just thinking about
5008	10	4	know that if
5009	10	3	know where i
5010	10	3	know where we
5011	10	3	less of a
5012	10	3	let you go
5013	10	3	like to try
5014	10	4	look for a
5015	10	3	look on your
5016	10	3	lot of other
5017	10	3	make a lot
5018	10	3	may have a
5019	10	3	me a couple
5020	10	3	me for my
5021	10	4	me on a
5022	10	3	might have to
5023	10	3	might just be
5024	10	3	more like a
5025	10	3	must be very
5026	10	3	need to bring
5027	10	3	no point in
5028	10	3	no reason why
5029	10	3	no way you
5030	10	4	not all the
5031	10	3	not in this
5032	10	3	not quite sure
5033	10	4	not sure how
5034	10	3	not until you
5035	10	3	now i think
5036	10	4	of a patient
5037	10	3	of the cells
5038	10	3	of the drugs
5039	10	3	of the few
5040	10	4	of the human
5041	10	3	of the year
5042	10	4	of thousands of
5043	10	4	of trying to
5044	10	3	of what you
5045	10	3	off to a
5046	10	3	off to the
5047	10	4	okay if i

5048	10	3	on a patient
5049	10	3	on my own
5050	10	4	on the top
5051	10	4	one part of
5052	10	4	or is that
5053	10	3	or the other.
5054	10	4	or would you
5055	10	3	out at the
5056	10	3	out of it
5057	10	3	patient is a
5058	10	3	patients in the
5059	10	4	place in the
5060	10	3	put him back
5061	10	3	put your hands
5062	10	3	quite a bit
5063	10	4	really going to
5064	10	3	remind you of
5065	10	3	rule out a
5066	10	3	say hello to
5067	10	3	say it was
5068	10	4	say this is
5069	10	4	see where the
5070	10	3	show you how
5071	10	3	side of his
5072	10	3	so i want
5073	10	3	so i'd like
5074	10	3	so if the
5075	10	3	so in other
5076	10	3	so let's go
5077	10	3	so much better
5078	10	4	so that they
5079	10	4	so that's why
5080	10	3	so we are
5081	10	3	so what happens
5082	10	3	some of her
5083	10	3	some things you
5084	10	4	something a little
5085	10	3	something we can
5086	10	3	soon as they
5087	10	3	spending time with
5088	10	3	stop with the
5089	10	4	sure that it
5090	10	4	talked to you
5091	10	3	tell them how

5092	10	3	than any of
5093	10	3	thank you all
5094	10	3	thank you. this
5095	10	3	that because i
5096	10	3	that for the
5097	10	3	that means is
5098	10	3	that means the
5099	10	3	that means you
5100	10	3	that part of
5101	10	5	that we can't
5102	10	3	that we had
5103	10	4	that we know
5104	10	3	that we need
5105	10	3	that we're going
5106	10	3	that will be
5107	10	3	that won't be
5108	10	3	that would have
5109	10	3	that you've been
5110	10	3	that's how it
5111	10	3	that's not for
5112	10	3	that's not going
5113	10	3	that's why they
5114	10	3	the car and
5115	10	4	the first place
5116	10	4	the importance of
5117	10	4	the infection is
5118	10	3	the left side
5119	10	4	the liver and
5120	10	3	the pain and
5121	10	4	the power of
5122	10	3	the question is,
5123	10	4	the same for
5124	10	3	the state of
5125	10	3	the symptoms are
5126	10	3	the thought of
5127	10	3	then i'm going
5128	10	3	then you know
5129	10	3	there on the
5130	10	4	these are not
5131	10	3	these are your
5132	10	3	they are in
5133	10	4	they call it
5134	10	3	they get to
5135	10	3	they might be

5136	10	3	they need a
5137	10	3	they think that
5138	10	3	they're not going
5139	10	3	thing i want
5140	10	3	think about is
5141	10	3	think i did
5142	10	3	think it could
5143	10	5	think one of
5144	10	3	think that the
5145	10	3	think that's the
5146	10	4	this as a
5147	10	4	this is in
5148	10	5	this is like
5149	10	3	three of us
5150	10	3	to be better
5151	10	4	to create a
5152	10	3	to go up
5153	10	3	to have been
5154	10	4	to help us
5155	10	3	to hold your
5156	10	4	to know what's
5157	10	3	to look out
5158	10	4	to me in
5159	10	3	to pull the
5160	10	4	to realize that
5161	10	3	to remember that
5162	10	3	to respond to
5163	10	3	to say a
5164	10	3	to see this
5165	10	3	to test for
5166	10	5	to the brain
5167	10	3	to the brain.
5168	10	3	to the end
5169	10	3	to the rest
5170	10	3	to work in
5171	10	3	told her that
5172	10	3	told you you
5173	10	3	too weak to
5174	10	3	tried to get
5175	10	3	turns out to
5176	10	4	use the word
5177	10	3	wait to see
5178	10	3	was at a
5179	10	3	was going on

5180	10	3	was no way
5181	10	3	we are talking
5182	10	3	we call it
5183	10	3	we could use
5184	10	4	we do not
5185	10	3	we don't stop
5186	10	3	we find a
5187	10	3	we going to
5188	10	3	we got it
5189	10	3	we haven't even
5190	10	3	we live in
5191	10	3	we may be
5192	10	3	we put in
5193	10	3	we take a
5194	10	3	we think it's
5195	10	3	we were all
5196	10	3	we were looking
5197	10	3	we'll do it
5198	10	3	went down to
5199	10	4	went to see
5200	10	4	were about to
5201	10	5	were talking about
5202	10	3	what i just
5203	10	3	what i thought
5204	10	4	what was going
5205	10	3	what we need
5206	10	3	what would happen
5207	10	3	what's gonna happen
5208	10	3	when he got
5209	10	3	when she had
5210	10	3	when you give
5211	10	3	when you need
5212	10	3	when you're not
5213	10	3	where do we
5214	10	3	where you are
5215	10	3	where you were
5216	10	3	which can be
5217	10	4	which is an
5218	10	3	who cares if
5219	10	4	who had a
5220	10	4	why do they
5221	10	3	why does it
5222	10	3	why would they
5223	10	4	will tell you

5224	10	3	with a good
5225	10	3	with each other.
5226	10	3	work in the
5227	10	3	working on a
5228	10	3	would be better
5229	10	3	would like you
5230	10	3	would make you
5231	10	3	would you give
5232	10	3	wouldn't be the
5233	10	3	wouldn't let me
5234	10	3	you an idea
5235	10	3	you are very
5236	10	3	you can only
5237	10	3	you come in
5238	10	3	you cut the
5239	10	3	you do is
5240	10	3	you don't take
5241	10	4	you get up
5242	10	3	you guys know
5243	10	3	you had an
5244	10	3	you happen to
5245	10	4	you have this
5246	10	3	you have what
5247	10	3	you hear what
5248	10	3	you know we
5249	10	4	you look for
5250	10	3	you looked at
5251	10	3	you mean the
5252	10	3	you might wanna
5253	10	3	you put your
5254	10	3	you tell them
5255	10	3	you what you
5256	10	3	you're dealing with
5257	10	3	you're starting to
5258	10	3	your patient is

Appendix C

Spoken Corpus: Full List of 4-gram Sequences

Spoken Corpus
4-gram MWUs

#Total No. of N-Gram Types: 823
#Total No. of N-Gram Tokens: 20626

#	Freq.	Ran.	MWU
1	473	4	i don't want to
2	294	4	you want me to
3	219	3	i want you to
4	211	3	i don't know what
5	206	4	you don't have to
6	187	4	do you want to
7	170	3	you don't want to
8	166	4	if you want to
9	141	4	i don't know how
10	139	3	i just want to
11	130	4	in the middle of
12	124	3	i don't know. i
13	110	5	the end of the
14	103	3	i want to be
15	98	3	i just wanted to
16	97	4	at the end of
17	96	3	what do you think
18	94	4	do you want me
19	93	5	take a look at
20	87	4	do you have any
21	86	4	for the rest of
22	86	3	want to talk about
23	86	3	you don't get to
24	83	4	i don't know if
25	78	4	i'm not going to
26	77	3	need to talk to
27	77	3	what are you talking
28	76	3	and i don't want
29	75	3	i just need to
30	73	3	are you talking about?
31	72	4	this is not a

32	71	4	do you know what
33	71	4	the middle of the
34	70	3	don't want to be
35	70	4	you want to go
36	68	3	are you going to
37	67	4	the rest of the
38	67	4	what do you want
39	66	5	a little bit of
40	65	4	to go to the
41	62	3	do you know how
42	62	5	i would like to
43	62	4	talk to you about
44	62	3	you need to get
45	61	3	don't know what to
46	61	4	to talk to you
47	61	3	why do you think
48	60	3	do you think i
49	60	3	nothing to do with
50	60	5	to be able to
51	59	3	i don't know why
52	59	3	i have to go
53	59	4	you know what i
54	57	4	i don't have to
55	57	4	is going to be
56	56	3	don't know how to
57	56	3	i don't think i
58	56	3	you want to do
59	55	3	and i don't know
60	55	3	you want to know
61	54	5	to make sure that
62	54	3	you don't need to
63	54	4	you have to do
64	53	3	and i want to
65	53	4	i think we should
66	52	4	i don't have time
67	50	3	going to talk about
68	49	3	i want to know
69	49	3	you want to talk
70	48	3	have no idea what
71	48	3	the two of you
72	47	3	i need to talk
73	47	3	i thought it was
74	47	3	thank you so much.
75	47	4	want to talk to

76	47	3	won't be able to
77	46	5	have a lot of
78	46	3	i have to get
79	46	5	is one of the
80	46	5	you have to be
81	46	3	you might want to
82	46	3	you want to get
83	45	3	don't want to talk
84	45	4	going to have to
85	45	4	the rest of your
86	45	3	to get out of
87	45	3	you need to be
88	44	3	as soon as i
89	44	3	don't want to go
90	44	4	i am trying to
91	44	4	i was going to
92	44	4	it has to be
93	44	4	it's going to be
94	44	4	we don't have to
95	44	3	what is wrong with
96	44	3	you really want to
97	43	3	i have no idea.
98	43	4	we need to do
99	42	4	don't have to do
100	42	3	don't want to do
101	42	4	i want to do
102	42	4	the only thing that
103	42	4	there's a lot of
104	42	4	we have to do
105	42	5	you're not going to
106	41	4	i have no idea
107	41	4	you are going to
108	40	3	and i have a
109	40	3	i thought we were
110	40	3	want to know what
111	40	4	you can see that
112	39	4	how do you know
113	39	4	i have to do
114	39	3	if you don't want
115	39	3	in a lot of
116	39	4	now i have to
117	39	4	the only way to
118	39	3	to go back to
119	39	3	when i was a

120	38	3	by the end of
121	38	3	i need to be
122	38	3	the fact that you
123	38	4	want to make sure
124	38	4	we were able to
125	37	4	a lot of people
126	37	4	i wish i could
127	37	3	need to do a
128	37	5	one of the most
129	37	3	we are going to
130	37	3	you don't have a
131	37	3	you just have to
132	37	4	you're going to be
133	36	3	i don't think you
134	36	3	i have to be
135	36	3	want to go to
136	36	3	you have to tell
137	35	4	going to be a
138	35	5	should be able to
139	35	3	so what do you
140	35	3	would you like to
141	34	5	and this is the
142	34	4	are going to be
143	34	3	figure out how to
144	34	3	go back to the
145	34	4	i have to tell
146	34	3	in a couple of
147	34	3	the rest of his
148	34	3	this is going to
149	34	3	trying to figure out
150	34	3	which is why i
151	33	3	and i have to
152	33	3	and i know that
153	33	4	to get to the
154	33	4	to take a look
155	32	3	have a problem with
156	32	3	i don't care if
157	32	3	i don't have any
158	32	4	i don't think it's
159	32	3	i was in the
160	32	3	know how to do
161	32	3	the middle of a
162	31	5	a look at the
163	31	3	get out of the

164	31	4	i think it's a
165	30	4	a bit of a
166	30	4	and i'm going to
167	30	3	as soon as we
168	30	3	but you have to
169	30	3	come up with a
170	30	4	have to do is
171	30	3	i know that you
172	30	4	i want to make
173	30	4	if you have a
174	30	3	not supposed to be
175	30	4	the top of the
176	29	4	and you can see
177	29	3	did you get the
178	29	4	for the first time
179	29	3	i'd like you to
180	29	3	what do you mean
181	29	5	when it comes to
182	29	3	you don't want me
183	29	3	you to tell me
184	28	5	a lot of the
185	28	3	do you know who
186	28	4	don't have time to
187	28	3	i think i'm gonna
188	28	3	i'm not talking about
189	28	4	if you have any
190	28	3	is that what you
191	28	4	might be able to
192	28	3	right in front of
193	28	4	there are a lot
194	28	3	you just need to
195	27	3	a hell of a
196	27	4	a little bit more
197	27	4	and this is a
198	27	3	as soon as you
199	27	3	do you think that
200	27	3	i don't see any
201	27	3	that's a lot of
202	27	3	to figure out how
203	27	4	to get rid of
204	27	4	what do we do
205	27	4	what i want to
206	27	4	you need to go
207	27	4	you need to know

208	26	3	for a long time.
209	26	3	got a lot of
210	26	3	have to do it
211	26	3	i asked you to
212	26	3	i just need a
213	26	4	i want to talk
214	26	3	i'm going to be
215	26	3	i'm trying to get
216	26	3	not going to be
217	26	4	so what do we
218	26	4	this is one of
219	26	5	to do with the
220	26	3	to get back to
221	26	4	we just have to
222	25	4	as long as you
223	25	3	but i don't think
224	25	3	don't want me to
225	25	3	don't want to get
226	25	3	don't want to have
227	25	3	if you could just
228	25	3	it's not going to
229	25	3	one of the things
230	25	3	thank you very much
231	25	4	the other side of
232	25	4	to tell you that
233	25	3	what are you going
234	24	3	all you have to
235	24	4	are a lot of
236	24	3	can i ask you
237	24	3	do you think it
238	24	3	find a way to
239	24	3	have to go to
240	24	4	i don't know about
241	24	3	i don't think that
242	24	4	i think this is
243	24	4	i was wondering if
244	24	4	i'm going to do
245	24	3	know what it's like
246	24	3	the right thing to
247	24	3	this is why i
248	24	3	used to be a
249	24	3	you want to see
250	23	3	and i want you
251	23	4	and you have to

252	23	3	and you need to
253	23	3	don't know how you
254	23	3	don't need to be
255	23	3	he doesn't have a
256	23	4	it's not a good
257	23	3	so what are you
258	23	3	so you need to
259	23	3	that you need to
260	23	3	the last time you
261	23	3	the one with the
262	23	4	to be in the
263	23	4	to look at the
264	23	3	want to do it
265	23	3	what do you do
266	23	3	which is why you
267	23	4	you need to do
268	23	3	you were going to
269	23	4	you're going to have
270	22	3	about the fact that
271	22	3	at the top of
272	22	3	do you think the
273	22	3	have to worry about
274	22	3	it looks like a
275	22	4	not be able to
276	22	3	so we need to
277	22	4	there's got to be
278	22	5	this is what i
279	22	3	what it's like to
280	22	4	you can see the
281	22	3	you want to give
282	22	3	you were supposed to
283	21	3	a little bit about
284	21	3	and i don't have
285	21	4	for a couple of
286	21	4	get rid of the
287	21	3	have to do a
288	21	4	have to go back
289	21	3	i don't have the
290	21	3	i need to take
291	21	3	i'm not sure i
292	21	3	in front of a
293	21	3	need to know what
294	21	4	on the other side
295	21	4	other side of the

296	21	3	so if you want
297	21	4	something to do with
298	21	3	the fact that you're
299	21	4	this is a very
300	21	3	to tell me what
301	21	3	want to be the
302	21	3	want to have a
303	21	3	we need to be
304	21	3	what i'm going to
305	21	3	you didn't have to
306	21	3	you might as well
307	20	3	a lot of time
308	20	4	at the same time
309	20	4	i think that you
310	20	3	i want to get
311	20	4	if you're going to
312	20	3	in the back of
313	20	4	it could be a
314	20	3	need to figure out
315	20	3	on my way to
316	20	3	this is the last
317	20	4	to get to know
318	20	4	to give you a
319	20	3	turned out to be
320	20	3	want to tell me
321	20	3	you have to get
322	20	3	you want to make
323	20	3	you'll be able to
324	19	3	and a lot of
325	19	3	and i think i
326	19	4	and i think that
327	19	4	and you have a
328	19	3	and you want to
329	19	3	but i want to
330	19	3	but you need to
331	19	3	has to be a
332	19	3	i can't tell you
333	19	3	i don't think she
334	19	3	i have a lot
335	19	3	i know it's not
336	19	4	if you look at
337	19	4	if you need to
338	19	3	if you've got a
339	19	3	it's one of the

340	19	3	just want to know
341	19	4	let me tell you
342	19	3	need to take a
343	19	3	need to talk about
344	19	3	rest of your life.
345	19	3	see if you can
346	19	3	the first time in
347	19	4	the size of a
348	19	3	there are plenty of
349	19	4	think of it as
350	19	3	this is what you
351	19	4	those of you who
352	19	3	we don't even know
353	19	4	you have to go
354	18	3	a pain in the
355	18	3	all we have to
356	18	3	by the time you
357	18	3	have to deal with
358	18	4	i can tell you
359	18	3	i don't think we
360	18	3	i need to do
361	18	3	i think it was
362	18	3	i'll give you a
363	18	3	i'm going to go
364	18	3	it had to be
365	18	3	need to make sure
366	18	4	one of the best
367	18	3	she didn't want to
368	18	4	so i'm going to
369	18	3	so that you can
370	18	3	so you can see
371	18	3	tell me about the
372	18	3	thank you very much.
373	18	4	that this is a
374	18	3	the back of the
375	18	3	the last time i
376	18	3	there's nothing we can
377	18	3	to talk about the
378	18	3	wanted to make sure
379	18	3	we have to be
380	18	3	who do you think
381	18	5	you look at the
382	18	3	you out of your
383	17	3	a lot of things

384	17	5	as far as i
385	17	3	but we need to
386	17	5	do a lot of
387	17	3	don't know if i
388	17	3	going to need to
389	17	3	have to be the
390	17	3	he's not going to
391	17	4	i don't think that's
392	17	3	i really need to
393	17	3	i would love to
394	17	3	if you don't get
395	17	5	it turns out that
396	17	3	it's got to be
397	17	3	just a little bit
398	17	3	thank you for your
399	17	4	that i need to
400	17	3	that you want to
401	17	3	the first time i
402	17	3	the head of the
403	17	3	there has to be
404	17	4	there's no sign of
405	17	3	there's no way to
406	17	3	this is the best
407	17	3	this part of the
408	17	4	thought it was a
409	17	3	to be on the
410	17	3	we need to find
411	17	3	you give me a
412	17	3	you going to do
413	17	3	you were trying to
414	17	3	you're not supposed to
415	16	3	a couple of weeks
416	16	4	all the way up
417	16	3	and we need to
418	16	4	anything to do with
419	16	3	can i have a
420	16	3	had nothing to do
421	16	3	how do you think
422	16	3	how long do you
423	16	3	i can get a
424	16	3	i think i need
425	16	3	i'm not sure what
426	16	4	if you think about
427	16	3	is gonna be a

428	16	3	is supposed to be
429	16	4	not going to get
430	16	3	so i have to
431	16	3	think this is a
432	16	3	this is a big
433	16	3	this is a good
434	16	3	want to do is
435	16	3	we don't know what
436	16	4	we're going to do
437	16	4	we're not talking about
438	16	3	what does that tell
439	16	4	what you need to
440	16	5	what's going on in
441	16	3	you don't have the
442	16	3	you have a good
443	16	4	you have to give
444	16	3	you want to keep
445	15	3	a good idea to
446	15	3	all the time you
447	15	3	all the way to
448	15	3	and i think it's
449	15	3	and if you don't
450	15	3	be able to tell
451	15	3	can i help you
452	15	3	do you think i'm
453	15	3	don't you want to
454	15	3	end of the day
455	15	3	got a problem with
456	15	3	how can you be
457	15	3	how did you get
458	15	4	i think we can
459	15	3	i thought it would
460	15	4	i told you that
461	15	3	if you don't do
462	15	3	is not going to
463	15	4	it doesn't have to
464	15	4	it would be a
465	15	3	know this is a
466	15	3	me to tell you
467	15	3	or it could be
468	15	4	so how do we
469	15	4	so this is a
470	15	3	so this is the
471	15	3	so you don't have

472	15	4	the best way to
473	15	3	think it would be
474	15	4	this is the first
475	15	3	want me to get
476	15	3	want me to take
477	15	3	want to know the
478	15	3	we have to go
479	15	3	we need to go
480	15	3	what you can see
481	15	3	you have to have
482	15	3	you know what this
483	15	3	you said that you
484	15	3	you want to come
485	15	3	you want to have
486	14	3	a problem with the
487	14	3	all over the place.
488	14	3	am i going to
489	14	3	as much as i
490	14	3	but i think i
491	14	3	but we have to
492	14	4	can see that the
493	14	3	can you give me
494	14	3	do you know about
495	14	3	do you think this
496	14	3	go in there and
497	14	4	going to be the
498	14	4	how do we know
499	14	3	i thought i had
500	14	3	i wanted to talk
501	14	3	i was thinking about
502	14	3	i won't be able
503	14	3	is there anything else
504	14	3	just need to know
505	14	3	just want to go
506	14	3	just wanted to make
507	14	3	she was trying to
508	14	5	the fact that the
509	14	4	the right side of
510	14	3	the source of the
511	14	3	this is what we
512	14	3	to do it in
513	14	3	trying to find a
514	14	3	want me to go
515	14	3	want you to do

516	14	3	what you have to
517	14	3	when you think about
518	14	3	where do you think
519	14	4	you can do the
520	14	3	you need to stop
521	14	3	you so much for
522	14	3	you tell me what
523	13	5	a look at this
524	13	3	a look at your
525	13	3	and this is what
526	13	3	are not going to
527	13	3	be a part of
528	13	3	be able to see
529	13	3	because you want to
530	13	4	do you know if
531	13	3	do you need to
532	13	4	doesn't have to be
533	13	3	don't even know how
534	13	3	don't have to go
535	13	3	don't know anything about
536	13	4	for a long time
537	13	3	for you to be
538	13	4	got to be a
539	13	3	had a lot of
540	13	4	has to do with
541	13	4	have to look at
542	13	4	have to tell you
543	13	4	he's going to be
544	13	4	how do we get
545	13	3	how much do you
546	13	3	i don't know yet.
547	13	3	i don't see why
548	13	3	i have to ask
549	13	3	i know that i
550	13	3	i might as well
551	13	3	i think i might
552	13	3	i'd like to do
553	13	3	i've been trying to
554	13	3	if you don't have
555	13	3	if you don't know
556	13	5	in order to get
557	13	4	it may not be
558	13	3	it might not be
559	13	3	it would have been

560	13	3	know what to do
561	13	4	lots and lots of
562	13	3	need to know that
563	13	3	of you who are
564	13	3	on the surface of
565	13	4	on top of the
566	13	3	only way to get
567	13	3	she's going to be
568	13	4	so how do you
569	13	3	so you have to
570	13	3	thank you so much
571	13	3	that would be a
572	13	3	the back of your
573	13	3	the rest of it
574	13	3	the side of the
575	13	4	this is just a
576	13	3	this is the way
577	13	3	thought it would be
578	13	3	to ask you to
579	13	3	to deal with the
580	13	3	to figure out what
581	13	3	to make sure you
582	13	3	want to know how
583	13	3	we don't have time
584	13	3	we need to make
585	13	3	we were in the
586	13	4	we'll be able to
587	13	3	what happens when you
588	13	4	what i'm trying to
589	13	3	will take care of
590	13	3	would you like me
591	13	3	you like me to
592	13	3	you should be able
593	12	3	a hole in the
594	12	3	a lot of blood
595	12	3	a patient with a
596	12	3	a whole bunch of
597	12	3	and i'm trying to
598	12	3	and the only way
599	12	3	and then there was
600	12	3	and what do you
601	12	3	but there is a
602	12	3	didn't have to do
603	12	3	do you think it's

604	12	3	don't know what's going
605	12	3	don't think i know
606	12	3	don't want to take
607	12	3	get a lot of
608	12	3	get out of here
609	12	3	going to show you
610	12	3	going to talk to
611	12	3	have to be in
612	12	3	have to find a
613	12	3	how do you do
614	12	3	i hope you don't
615	12	3	i know you're not
616	12	3	i was talking to
617	12	3	i will give you
618	12	3	i'm going to tell
619	12	3	if i told you
620	12	4	if i want to
621	12	3	if you can get
622	12	4	if you had a
623	12	3	if you'd like to
624	12	3	is a little bit
625	12	3	is going to take
626	12	3	it comes down to
627	12	3	it out of the
628	12	3	it's all about the
629	12	3	it's the only way
630	12	4	just a couple of
631	12	4	need to be a
632	12	3	need to do an
633	12	3	one hell of a
634	12	3	so i don't have
635	12	3	so we have to
636	12	3	so you want to
637	12	5	that i have to
638	12	4	that i'm going to
639	12	3	the kind of thing
640	12	3	the only thing you
641	12	3	the rest of you
642	12	3	there is no way
643	12	3	there was a time
644	12	3	there's no such thing
645	12	4	to be a little
646	12	3	to find a way
647	12	3	to go to a

648	12	3	to take you to
649	12	3	waiting for me to
650	12	3	was on my way
651	12	4	we might be able
652	12	3	we're going to need
653	12	3	what i need to
654	12	3	you can do it
655	12	3	you need to make
656	12	4	you really need to
657	11	4	a way to get
658	11	3	and hope for the
659	11	3	and i know i
660	11	3	and one of the
661	11	3	and the rest of
662	11	3	and then i will
663	11	3	and you're going to
664	11	3	are you talking about
665	11	3	as a result of
666	11	3	as soon as the
667	11	4	be able to get
668	11	4	been a lot of
669	11	3	but this is a
670	11	3	but you can see
671	11	4	by the time we
672	11	3	can come up with
673	11	3	can i have the
674	11	4	can't do anything about
675	11	3	didn't know what to
676	11	3	don't know how much
677	11	4	don't know if you
678	11	3	don't know what it's
679	11	3	don't want to know
680	11	3	fell in love with
681	11	3	get them out of
682	11	4	going to try to
683	11	4	have to have a
684	11	3	he was going to
685	11	3	hell of a lot
686	11	3	how can i help
687	11	3	how do we do
688	11	3	i can give you
689	11	3	i don't know. we
690	11	3	i don't think it
691	11	3	i get out of

692	11	4	i think it's the
693	11	3	i want to keep
694	11	3	i wanted to say
695	11	3	i was just thinking
696	11	3	i'm not sure that
697	11	3	if she has a
698	11	3	in and out of
699	11	3	in order to do
700	11	3	it doesn't matter what
701	11	5	it was the first
702	11	3	may be able to
703	11	3	need to tell you
704	11	3	not gonna do it.
705	11	3	nothing you can do
706	11	3	now i have a
707	11	5	on to the next
708	11	3	one of them is
709	11	3	one of you is
710	11	3	she was in the
711	11	3	that is going to
712	11	3	that you have a
713	11	3	the size of the
714	11	3	the surface of the
715	11	3	the wall of the
716	11	3	then i have to
717	11	3	there's something wrong with
718	11	4	think we need to
719	11	4	to be a better
720	11	3	to do with it.
721	11	3	to give her a
722	11	3	to go back in
723	11	3	to go into the
724	11	3	to stay in the
725	11	3	want to look at
726	11	4	was going to say
727	11	3	was one of the
728	11	3	was wondering if you
729	11	4	we can do about
730	11	3	we can talk about
731	11	3	we don't know how
732	11	3	what do you need
733	11	3	wish i could tell
734	11	4	with the fact that
735	11	4	you could have a

736	10	3	a couple of days
737	10	3	a lot of work
738	10	4	a part of the
739	10	3	and as soon as
740	10	4	and i don't think
741	10	3	and look at the
742	10	3	and the fact that
743	10	3	and the only thing
744	10	4	and then you get
745	10	3	and this is one
746	10	3	and you know what
747	10	3	are there any questions
748	10	3	be some kind of
749	10	3	blood flow to the
750	10	3	but you don't know
751	10	3	do you think is
752	10	3	does it look like
753	10	3	don't know how many
754	10	3	don't know what that
755	10	3	don't know what the
756	10	3	get out of this
757	10	3	going back to the
758	10	3	going to give you
759	10	4	going to tell you
760	10	3	have a chance to
761	10	3	have a couple of
762	10	3	have to tell me
763	10	3	have to tell them
764	10	3	i can think of
765	10	3	i can't think of
766	10	3	i could tell you
767	10	3	i don't see how
768	10	3	i don't think this
769	10	3	i gave you a
770	10	3	i have to say
771	10	3	i should be able
772	10	3	i think i got
773	10	3	i think that's a
774	10	5	i think we have
775	10	4	i think we need
776	10	3	i want to give
777	10	3	i was talking about
778	10	3	i'm going to say
779	10	3	i've been thinking about

780	10	3	if it was a
781	10	4	in the united states
782	10	4	is part of the
783	10	3	it was a bad
784	10	3	it's a good thing.
785	10	3	it's a waste of
786	10	5	it's not just the
787	10	4	let's go back to
788	10	3	more likely to get
789	10	3	my way to the
790	10	3	need to find a
791	10	3	need to know the
792	10	3	not a very good
793	10	3	not going to talk
794	10	4	of it as a
795	10	3	on one of the
796	10	4	put it in the
797	10	3	she doesn't have any
798	10	3	so you can get
799	10	4	sometimes you have to
800	10	3	that needs to be
801	10	3	that there was a
802	10	3	the fact that we
803	10	3	the way to the
804	10	3	they have to be
805	10	4	to be one of
806	10	3	to come out of
807	10	3	to do is to
808	10	3	to find out what
809	10	3	to make sure i
810	10	3	to the fact that
811	10	3	to the point where
812	10	3	to the rest of
813	10	3	want me to talk
814	10	4	was a lot of
815	10	3	we still have to
816	10	3	we were trying to
817	10	3	why is it so
818	10	3	would like you to
819	10	3	you have to make
820	10	3	you know what the
821	10	3	you look like a
822	10	3	you want to stop
823	10	3	you've got to be

Appendix D

Spoken Corpus: Full List of 5-gram Sequences

Spoken Corpus
5-gram MWUs

#Total No. of N-Gram Types: 72
#Total No. of N-Gram Tokens: 1447

#	Freq.	Ran.	MWU
1	86	4	do you want me to
2	68	3	what are you talking about?
3	55	4	at the end of the
4	55	4	in the middle of the
5	51	3	i don't know what to
6	45	3	i need to talk to
7	43	3	and i don't want to
8	36	4	to talk to you about
9	31	3	by the end of the
10	29	4	to take a look at
11	28	3	what do you want to
12	27	3	i don't want to go
13	25	3	what are you going to
14	23	4	for the rest of your
15	23	3	i don't want to do
16	23	4	there are a lot of
17	23	3	what do you want me
18	23	3	you don't want me to
19	21	3	i don't know how you
20	21	3	if you don't want to
21	21	4	take a look at the
22	21	4	the other side of the
23	20	3	do you want to go
24	20	3	i have no idea what
25	18	3	the rest of your life.
26	18	3	to figure out how to
27	17	3	are you going to do
28	17	4	on the other side of
29	17	3	you have to do is
30	16	4	i don't have time to
31	16	3	i don't know if i

32	16	3	know what it's like to
33	16	3	need to talk to you
34	15	3	at the top of the
35	15	3	the end of the day
36	14	3	and i don't know if
37	14	3	for the rest of the
38	14	3	i have a lot of
39	14	3	i just want to know
40	14	3	i want to make sure
41	14	3	i want to talk to
42	14	3	so if you want to
43	14	3	this is one of the
44	13	3	i won't be able to
45	13	3	so you don't have to
46	13	3	thank you so much for
47	13	3	want to make sure that
48	13	3	we need to do a
49	13	3	what do you think the
50	13	3	would you like me to
51	12	4	and you can see that
52	12	3	i just want to go
53	12	3	i just wanted to make
54	12	3	this is going to be
55	12	3	those of you who are
56	12	4	we might be able to
57	12	3	you're going to have to
58	11	3	do you know what i
59	11	4	i don't know if you
60	11	4	i was going to say
61	11	3	i was wondering if you
62	11	3	i wish i could tell
63	11	3	just wanted to make sure
64	11	3	the only way to get
65	11	3	want to talk to you
66	11	4	you can see that the
67	11	3	you should be able to
68	11	3	you want me to get
69	10	4	i have to tell you
70	10	3	if you look at the
71	10	4	think of it as a
72	10	3	why do you think i

Appendix E

Written Corpus: Full List of 3-gram Sequences

Written Corpus
3-gram MWUs

#Total No. of N-Gram Types: 282
#Total No. of N-Gram Tokens: 7873

#	Freq.	Ran.	MWU
1	269	4	the presence of
2	168	5	as well as
3	153	3	the number of
4	124	3	the absence of
5	117	3	in order to
6	106	3	in response to
7	104	3	in the absence
8	100	5	one of the
9	93	3	in the presence
10	92	3	been shown to
11	89	3	due to the
12	84	6	a number of
13	81	3	shown to be
14	80	3	it has been
15	71	3	the development of
16	70	4	in addition to
17	69	3	the role of
18	66	5	a role in
19	66	4	involved in the
20	65	5	according to the
21	64	4	% of the
22	63	5	likely to be
23	62	4	the effect of
24	62	4	the use of
25	61	3	in the case
26	61	3	the case of
27	61	3	the production of
28	59	5	based on the
29	58	4	a variety of
30	57	4	part of the
31	56	3	the fact that

32	53	3	was found to
33	50	5	in the same
34	49	6	some of the
35	48	3	responsible for the
36	48	3	the formation of
37	46	3	a total of
38	46	3	found in the
39	46	4	most of the
40	46	4	the level of
41	45	3	the majority of
42	45	6	the risk of
43	44	3	role in the
44	43	3	the importance of
45	42	4	to be a
46	41	4	is associated with
47	40	3	. % of
48	40	4	the lack of
49	40	3	was associated with
50	38	4	in which the
51	38	4	play a role
52	37	3	have been shown
53	37	3	it is possible
54	37	4	the proportion of
55	36	3	a result of
56	36	5	increase in the
57	35	3	as a result
58	35	6	be used to
59	35	5	for example, the
60	35	3	large number of
61	35	3	remains to be
62	35	4	the university of
63	34	3	appears to be
64	34	4	it is not
65	33	4	there is a
66	32	5	in the first
67	32	3	is involved in
68	32	3	s and s
69	32	4	such as the
70	32	5	to be the
71	31	4	associated with a
72	30	3	as part of
73	30	5	can be used
74	30	4	led to the
75	30	3	of the genes

76	30	4	suggesting that the
77	30	3	the amount of
78	29	3	in terms of
79	29	3	we found that
80	28	5	have shown that
81	28	4	in patients with
82	28	4	need to be
83	28	3	the course of
84	27	3	depending on the
85	27	3	is possible that
86	27	5	this is the
87	27	3	well as the
88	26	4	at least one
89	26	4	be able to
90	26	3	contribute to the
91	26	3	increased risk of
92	26	6	that can be
93	26	5	the effects of
94	26	4	the impact of
95	25	4	all of the
96	25	3	an increase in
97	25	3	are able to
98	25	5	found that the
99	25	4	have not been
100	25	4	may not be
101	25	3	that of the
102	24	4	at the university
103	24	3	compared with the
104	24	4	of the most
105	24	3	suggest that the
106	23	3	because of the
107	23	3	been found to
108	23	3	in addition, the
109	23	3	is one of
110	23	5	more likely to
111	23	5	part of a
112	23	4	suggested that the
113	23	3	the basis of
114	23	3	the potential to
115	23	3	there was no
116	22	3	it may be
117	22	4	known to be
118	22	5	of the first
119	21	4	parts of the

120	21	3	the evolution of
121	21	3	were found to
122	20	4	are likely to
123	20	3	be used in
124	20	3	has been found
125	20	3	high levels of
126	20	3	in the last
127	20	4	is known to
128	20	3	it should be
129	20	4	led to a
130	20	3	needs to be
131	20	3	quality of life
132	20	4	seems to be
133	20	3	to identify the
134	19	3	and it is
135	19	3	any of the
136	19	3	are thought to
137	19	5	at the same
138	19	3	considered to be
139	19	3	have been found
140	19	4	is an important
141	19	3	is the first
142	19	3	out of the
143	19	4	region of the
144	19	4	that it is
145	19	4	the need for
146	19	4	were able to
147	19	3	were associated with
148	19	3	which can be
149	18	6	a range of
150	18	3	genes in the
151	18	3	suggests that the
152	18	3	that have been
153	18	3	the time of
154	18	4	to explore the
155	18	3	to have a
156	17	3	associated with an
157	17	3	at the time
158	17	3	be associated with
159	17	3	body mass index
160	17	3	could be used
161	17	4	one or more
162	17	3	than in the
163	17	5	the end of

164	17	3	two or more
165	16	3	are known to
166	16	3	be related to
167	16	4	cells in the
168	16	3	in the study
169	16	3	is based on
170	16	3	it would be
171	16	3	levels of the
172	16	3	lower risk of
173	16	3	may be a
174	16	3	results suggest that
175	16	3	there is no
176	15	3	be important for
177	15	3	been found in
178	15	4	could be a
179	15	3	could not be
180	15	4	different types of
181	15	4	in the past
182	15	4	is that the
183	15	3	many of the
184	15	3	of the total
185	15	4	studies have shown
186	15	3	the control group
187	15	3	the extent of
188	15	3	the quality of
189	15	3	there was a
190	15	3	thought to be
191	14	3	at least two
192	14	3	be explained by
193	14	3	due to a
194	14	4	end of the
195	14	4	for the first
196	14	3	in the human
197	14	4	lead to a
198	14	4	more than %
199	14	4	of more than
200	14	3	of the study
201	14	3	over the last
202	14	4	risk of death
203	14	3	serve as a
204	14	3	that they are
205	14	3	the study of
206	14	3	up to %
207	13	3	a combination of

208	13	3	a subset of
209	13	3	account for the
210	13	3	an increased risk
211	13	3	approved by the
212	13	3	at least in
213	13	3	be used for
214	13	3	in the current
215	13	3	in the form
216	13	3	in the genome
217	13	4	or in the
218	13	3	regions of the
219	13	3	than % of
220	13	3	the form of
221	13	3	the increase in
222	13	3	to be important
223	13	3	to show that
224	13	3	was able to
225	12	3	as a result,
226	12	3	based on a
227	12	3	believed to be
228	12	3	but it is
229	12	3	have led to
230	12	3	in the early
231	12	3	it was not
232	12	3	of a gene
233	12	4	that % of
234	12	3	that could be
235	12	3	the immune system
236	12	3	this is a
237	12	4	this type of
238	12	3	version of this
239	12	3	will not be
240	12	3	with an increased
241	12	5	women in the
242	11	3	a link between
243	11	3	a loss of
244	11	3	a very low
245	11	3	and they are
246	11	3	cells of the
247	11	3	details of the
248	11	3	effect of the
249	11	3	is a key
250	11	3	is needed to
251	11	4	it is a

252	11	3	it remains to
253	11	4	less likely to
254	11	5	made up of
255	11	3	of the human
256	11	3	that there are
257	11	3	the complexity of
258	11	4	the first time
259	11	3	the release of
260	11	3	to be associated
261	11	3	to create a
262	11	3	to examine the
263	11	4	which is a
264	10	3	a lack of
265	10	3	during the first
266	10	3	for patients with
267	10	3	have the potential
268	10	3	impact on the
269	10	3	in the uk
270	10	4	of the disease
271	10	3	of the university
272	10	4	published in the
273	10	3	quality of the
274	10	3	stages of the
275	10	4	such as those
276	10	3	the range of
277	10	5	the same time
278	10	3	there have been
279	10	3	to the development
280	10	4	whether or not
281	10	3	which in turn
282	10	3	which may be

Appendix F

Written Corpus: Full List of 4-gram Sequences

Written Corpus
4-gram MWUs

#Total No. of N-Gram Types: 21
#Total No. of N-Gram Tokens: 597

#	Freq.	Ran.	MWU
1	94	3	in the absence of
2	85	3	in the presence of
3	59	3	in the case of
4	41	3	been shown to be
5	36	4	play a role in
6	35	3	have been shown to
7	34	3	as a result of
8	27	3	as well as the
9	27	3	it is possible that
10	24	4	at the university of
11	17	4	can be used to
12	15	4	more likely to be
13	15	3	one of the most
14	13	3	in the form of
15	12	3	this is the first
16	11	3	an increased risk of
17	11	3	it remains to be
18	11	3	to be associated with
19	10	3	associated with an increased
20	10	3	have the potential to
21	10	3	to the development of

Appendix G

Written Corpus: Full List of 5-gram Sequences

Written Corpus
5-gram MWUs

#Total No. of N-Gram Types: 1
#Total No. of N-Gram Tokens: 17

#	Freq.	Ran.	MWU
1	17	3	have been shown to be