Don't Let's Try to Break this Down: Teasing Apart Lexical Chunks

Zoë Bogart
University of Groningen
University of Malta

August, 2011

Supervisors:

Dr. Gertjan van Noord
University of Groningen
&
Dr. Mike Rosner
University of Malta

Submitted to the Erasmus Mundus European Masters Program
in Language & Communication Technologies (LCT)

In loving memory of my grandfather, Louis Bogart: a man of profound intelligence, perseverance, curiosity, and wit, but above all, a man of unlimited kindness.

Here's looking at you kid.

Many people have contributed in ways both direct and indirect to the work presented in this thesis, and I would like to acknowledge their help, without which this thesis would not have been possible.

First and foremost, I would like to thank my thesis advisor, Gertjan van Noord, without whose guidance this thesis might never have gotten off the ground, would likely not have developed nearly as much as it eventually did, and would certainly not have reached its completion in anything like a reasonable time period.

I would also like to thank Mike Rosner and Gosse Bouma, the program heads for the Erasmus Mundus Language and Communication Technologies program at the University of Malta and the University of Groningen, for all their assistance and work to make the program run as smoothly as possible for the students. I would further like to thank Mike Rosner for agreeing to be my second reader, even during his sabbatical year.

Thank you to the many wonderful teachers I had at the aforementioned universities, and particularly to Albert Gatt, who spurred my interest in statistical corpus-based linguistics. Many thanks also to Kaitlin Mignella and Ross Batchelor, who patiently rated hundreds of chunks (yes, I will buy you both dinner as promised).

To my family, who have enthusiastically supported me in all my endeavours, no matter how crazy, such as going off to study for one year in a tiny island in the middle of the Mediterranean and for another year in a country of seemingly endless (for this California-raised girl) rain, cold, and snow, or for choosing to study Computational Linguistics, an area whose mention produces blank stares at cocktail parties, and whose explanation produces even blanker stares.

Thank you to my parents for all their assistance with this thesis, and for all their assistance, love, and support in everything else as well. Y a mi querida abuelita, muchísimas gracias por todo tu apoyo y por el ejemplo que me has siempre dado con tu pasión por la vida.

Finally, I would like to thank the many students of English I had in Milan whose questions, conversations, and struggles with the English language taught me so much about my native tongue and about language in general. It takes courage to step outside the language you have been raised with and try to communicate in a foreign tongue. To all my students and to students of foreign languages everywhere, thank you for having the courage to try!

Contents

Chapter 1

Introduction

> *since feeling is first*
> *who pays any attention*
> *to the syntax of things*
> *will never wholly kiss you*
>
> e. e. cummings

1.1 Grammar and Lexicon: Problems with the Traditional Approach

How do people learn languages? How do people learn to comprehend and communicate complex ideas in a continuing, ever-changing stream of information? How do languages work? For years, research into questions like these has focused on grammar. Languages, especially as they are taught to non-native speakers, have been divided into two main areas: grammar and lexicon. The grammar is learnt through memorization and practice using different rules, and the words in the lexicon are simply inserted into the proper places, according to their grammatical categories. This approach is simple, and it can be applied to nearly any language, yet on closer inspection, it is full of difficulties, and not just minor ones.

Take for example, the English word *make*. This is one of the most common words in English, appearing over 200,000 times in the British National Corpus (BNC), a 100 million word collection of spoken and written British English[1]. As such, it is clearly an important word for the student of English to learn to use, yet many non-native speakers have great difficulty with this seemingly simple word. For one thing, many languages use the same verb to cover actions that English splits into those that you make and those that you do (e.g. French *faire*, German *machen*, Spanish *hacer*). When to use *make* and when to use *do* in English is not always clear: you make a mistake, but you do your homework; you make your bed, but you do the dishes (in American English at least - in British English, you do the washing up); you make someone an offer, but you do someone a favor; you do harm, but you make amends. In fact, amends are never anything but made in English.

---

[1] For comparison, the word *create*, a near synonym of *make*, appears in its inflected forms just over 20,000 times; thus it has only a tenth of the frequency of *make*.

As if this wasn't bad enough, *make* appears in all sorts of other constructions where its meaning seems totally different from the standard meaning of roughly "create, produce, cause to bring into existence". For example, constructions like *make up, make do, make out, make over,* and *make believe* seem to require separate lexical entries; certainly the phrase *make up*, meaning "to invent", cannot be derived from the standard meanings for either *make* or *up*. And besides this, *make* appears in a variety of idioms, which can also be learned only through memorization, e.g. *make hay while the sun shines, make or break,* and *make a mountain out of a molehill*.

All of these observations suggest that learning a language requires far more than just knowledge of grammatical rules and dictionary entries for vocabulary; in order to speak a language fluently, people must have knowledge of which words to use *when*, not only in grammatical terms, but in lexical terms. People learning English must know that *make* is used with *mistake, wish,* and *amends*, and not only this, but they must know that one makes *a* wish and very rarely *the* wish. They must know how to use and understand phrasal verbs like *make out* and idioms like *make or break*. Multiword constructions like these shall be referred to in the rest of this work as *lexical chunks*. As the examples above demonstrate, lexical chunks, which seem to fall somewhere in between grammar and lexicon, are of unique importance in both language acquisition and language use.

1.2 Lexical Chunks

Lexical chunks have been defined in numerous ways. Many of these definitions shall be explored in the following chapter, but for now they will simply be defined as groups of two or more words that tend to occur together and that often, though not always, are non-compositional (that is, the meaning of the chunk as a whole is not fully determinable from the meanings of its individual words and any meanings conveyed by the syntactic operations combining them). Lexical chunks have been implicated as playing an important role in human language processing and acquisition, and the automatic identification of lexical chunks is beneficial to many areas of Computational Linguistics, including Machine Translation, automatic parsing, and automatic text evaluation. Lexical chunk dictionaries are also useful for language teachers and learners, as knowledge of lexical chunks has been identified as a key factor distinguishing fluent from non-fluent speakers (Pawley and Syder 1983).

Despite their potential usefulness in a broad range of applications, lexical chunks, treated as a single category, have not received nearly as much attention from linguists as subcategories of lexical chunks have, perhaps because the full category of lexical chunks is difficult to define in concrete terms. While phrasal, idiom, and collocation dictionaries abound, no such equivalent dictionary of lexical chunks exists. Similarly, while much work has been done on the automatic extraction of specific types of chunks like collocations and named entities, little work has been done on the automatic extraction of lexical chunks as a whole category. In the past few years, multiword expressions have attracted more attention from computational linguists, and various workshops devoted to the theme of automatic extraction of these expressions have taken place. Yet even multiword expressions are often taken to be contiguous sequences of words, and as such, they form only a subset of the broader category of lexical chunks.

As the area of lexical chunks, as a single category, in computation has received little attention, the aim of this thesis is to review different methods for the automatic extraction of lexical chunks from text and determine which is the best. Five statistical methods were chosen for this task, and a corpus of roughly 5,800,000 words was used for training. Lexical chunks extracted from this corpus were then used to find chunks in an evaluation text. These chunks were then compared to a gold standard of chunks composed, for one part, of chunks found in various collocation and phrase dictionaries and, for the other part, of a list of chunks that had been judged by human raters to be good instances of lexical chunks.

The rest of the thesis is structured as follows: Chapter 2 gives an overview of lexical chunks as they have been defined in the literature and of evidence for the existence and usefulness of lexical chunks in language processing and acquisition. Chapter 3 goes into the details of previous computational work: automatic extraction methods that have been used for lexical chunks and related linguistic phenomena. Chapter 4 describes the materials and methods used in the current research, and Chapter 5 describes the method used for the evaluation process. Results are presented in Chapter 6, and a discussion of these results and their implications for future work is given in Chapter 7. Chapter 8 offers some concluding remarks and suggestions for directions for future research.

Chapter 2

Lexical Chunks: What are they and why should we care?

2.1 What is a Lexical Chunk?

Lexical chunks have received relatively little attention in much of the traditional literature on Formal and Theoretical Linguistics, with most of their biggest proponents coming from the areas of Applied Linguistics and Education, specifically Foreign Language Education[1]. One reason for this lack of discussion in Formal Linguistics may be that lexical chunks are a difficult phenomenon to pin down formally. Because they combine semantic, syntactic, lexical, and even pragmatic information, lexical chunks do not fit neatly into traditional linguistic categories. Additionally, and perhaps in part because of their cross-categorial nature, lexical chunks are difficult to define in simple, universally applicable terms. Despite this, some efforts have been made, and in this chapter, I will review some of the most influential of these definitions. I will then report on evidence for the existence of lexical chunks as distinct linguistic phenomena, and finally, I will offer some arguments for the importance of lexical chunks in a variety of real-world applications, not only in language teaching, but also in several NLP (Natural Language Processing) applications.

2.2 Defining Lexical Chunks

Lexical chunks were first introduced in the field of Applied Linguistics, and they have their biggest supporters among educators and linguists interested in Foreign Language Teaching. They arise out of the notion that language as it used - in spoken and written sentences - may best be viewed not as a collection of solitary words, transformed and joined together by morphological and syntactic procedures, but rather as groupings of words that tend to occur together and that are used across different situations to convey a similar pragmatic and semantic message. For a basic example, consider some of the phrases in table 2.1. These are all typical phrases of greeting, some of the first that might be taught in a foreign language class.

Though these are some of the first phrases a language learner would en-

---

[1] Though it should be noted that there are some theories, such as Construction Grammar, which place greater emphasis on lexical chunks and chunk-like structures.

Language	Phrases
English	*How are you, how's it going, what's up, how do you do*
French	*Comment ça va, comment vas tu, comment allez vous*
Spanish	*Que tal, que pasa, como va, como estás*
Turkish	*Nasılsın, ne haber, nasılsınız, ne var ne yok*

Table 2.1: *Phrases of greeting in different languages.*

counter in class, many of them are relatively complex in their morphology and syntax. For example, the French and English examples require subject-verb inversion, a phenomenon that proves quite difficult for second language learners to master (Pienemann 1998). Many of the phrases also require verbs to be conjugated to match the subject, and the Turkish phrases *nasılsın* and *nasılsınız* involve morphological operations that have a similar effect of subject-matching (*nasılsın* is 2nd-person singular informal, while *nasılsınız* is 2nd-person singular formal and 2nd-person plural). The complexity of the morphological and syntactic operations required to produce many very basic phrases such as these is part of the reason some linguists have posited that language may often be learned and processed in chunks as opposed to single words. The issue then is to determine what constitutes a chunk and what does not. In the remainder of this section, I shall review various definitions that have been put forth.

2.2.1 Definitions and Terminology

Definitions for lexical chunks abound, as does the terminology used to describe the phenomenon. Though this work shall exclusively use the term 'lexical chunks', the literature is full of alternative names, including: *conventionalized language forms, fixed expressions, formulaic expressions, formulaic language, formulaic sequences, institutionalized clauses, lexical bundles, lexical items, lexical phrases, lexicalized sentence stems, multiword expressions, multiword lexical units, multi-word sequences, patterned speech, phraseological expressions, recurrent phrasal constructions,* and *speech formulae*. Without even delving into the definitions behind these terms, we can already see some patterns emerging that give clues as to what the important characteristics of lexical chunks are. They are often defined as 'multiword', so a key feature of these chunks is that they are units longer than a single word. The word 'lexical' appears quite a bit, suggesting that words are grouped together lexically as opposed to semantically, syntactically, etc. We also see lots of terms referring to patterns, formulaicity, or conventionality, so it seems that these notions play an important role in defining what a lexical chunk is.

A look at some of the various definitions that have been posited for lexical chunks confirms these notions about what their important characteristics are. Table 2.2 gives a few of the many definitions that have been put forth, along with the corresponding terminology. The definitions listed here, just a small

Terminology	**Definition**	**Source**
lexical bundle	sequences of words that commonly go together in natural discourse	(Biber et al. 1999)
lexical item	a unit of description made up of words and phrases	(Sinclair 2004)
lexical phrase	multi-word lexical phenomena that exist somewhere between the traditional poles of lexicon and syntax, conventionalized form/function composites that occur more frequently and have more idiomatically determined meaning than language that is put together each time	(Nattinger and DeCarrico 1992)
lexicalized sentence stem	a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed; its fixed elements form a standard label for a culturally recognized concept, a term in the language	(Pawley and Syder 1983)
multiword lexical unit	a group of words that occur together more often than expected by chance	(Dias et al. 1999)
recurrent phrasal construction	combinations of lexis and grammar...which typically consist of a partly fixed lexical core plus other variable items	(Stubbs 2007)
speech formula	a multimorphemic phrase or sentence that, either through social negotiation or through individual evolution, has become available to a speaker as a single prefabricated item in his or her lexicon	(Peters 1983)

Table 2.2: *A sample of terminology and definitions*

sample, already cover a lot of ground, and while there is quite a bit of overlap, there are also areas of difference regarding both the form and the function of lexical chunks.

Formally, it seems clear that in almost all definitions, lexical chunks are groups of words. However, it is not clear how long these groups must be; some definitions require the groups to be of clause, phrase, or sentence length, while others make no such specification. Another area that is not clear is whether or not lexical chunks can contain gaps, as in a phrase like *as X as*, where the gap, represented by the 'X', can be filled by specific types of words or phrases. Some definitions, such as Biber et al.'s, require that the words be in a continuous sequence, while others either make no mention of gaps or, like Stubbs' definition, explicitly allow for their presence.

In terms of function, there is a strong tendency for definitions to note the conventionalized nature of lexical chunks, but different definitions emphasize this in different ways. For some, lexical chunks are defined mainly by their statistical properties, i.e., they are groups of words that occur frequently together. For others, idiomaticity is equally or even more important than frequency. In these definitions, a key feature of lexical chunks is that they are at least partly non-compositional, i.e., the meaning of the whole chunk is not fully determinable from the meanings of the individual words that combine to form it. Still other definitions of lexical chunks place greater emphasis on their pragmatic, social, cultural, or psychological functions, e.g., Peters' or Pawley and Syder's definitions.

This wide variety of ideas about what lexical chunks are stems in part from the fact that different people have looked at lexical chunks for different reasons. Quite naturally, researchers doing corpus analyses on lexical chunks have tended to focus on their statistical properties, while researchers interested in areas like aphasic speech have been more inclined to focus on the semantic properties and psychological representations of chunks. Lexical chunks then are perhaps best defined by the area of application; someone interested in creating a dictionary of lexical chunks for foreign language learners may want to use one definition, while someone building a system that automatically divides text into chunks may prefer to use a different definition. As my goal is to create a system that can find lexical chunks to be used for a variety of purposes, from dictionary creation to foreign language teaching, machine translation, and so on, I shall examine definitions of all these different types, focusing first on extensional definitions which list different linguistic categories of lexical chunks, and then on intensional definitions which focus on the sociological, pragmatic, psychological, and/or statistical properties of chunks.

2.2.2 Plotting the Territory: Linguistic Categories of Lexical Chunks

As noted in the beginning of this chapter, lexical chunks are not easily definable by their syntactic or morphological characteristics because they are so varied. Instead, most definitions that attempt to characterize lexical chunks by their linguistic features do so by dividing them into categories, which have as a common thread their formulaic, predictable nature. One of the most thorough and influential such categorizations is that of Nattinger (1980), adapted from Becker (1975). The Nattinger/Becker categories, ordered by phrase length from shortest to longest are described below.

1. **Polywords:** Small groups of words that function the same way a single word does. Examples that fall into this category include phrasal verbs (*wake up, turn off*), slang (*jump the gun, over the moon*), and euphemisms (*go to the bathroom, made redundant*).

2. **Phrasal Constraints:** Short phrases with more variability than polywords, but whose variability is generally constrained to a small set of words, as in: *two o'clock, twelve o'clock*, etc.

3. **Deictic Locutions:** Short to medium-length phrases which serve as pragmatic indicators that help direct the flow of conversation. These include phrases like: *by the way, on the other hand, for what it's worth*, and so forth.

4. **Sentence Builders:** Long, highly variable phrases (up to sentence length) which provide a framework for expressing an idea. They tend to have gaps which can be filled in with a large number of words, for example: *A is the new B* or *the X-er the Y-er*.

5. **Situational Utterances:** Long phrases, usually of sentence length, which are appropriate to very particular situations such as: *don't worry about it, pleased to meet you, have a good trip*.

6. **Verbatim Texts:** Memorized texts of any length - quotations, poems, song lyrics, parts of novels, etc.

As can be seen, Nattinger and Becker's categories vary enormously in terms of length, form, and fixedness. Lengthwise, the lexical chunks can consist of any number of words greater than one, though only the verbatim text category allows for groups of words of greater than sentence length. Some categories, like the phrasal constraints and sentence builders, allow for gaps which can be filled in by a set of words, while other categories, like the polywords and verbatim texts, do not. When there are gaps, the group of words that can fit in the gaps can be large, as in the sentence builders, or small, as in the phrasal constraints.

Another important categorical definition is that of Lewis, who divides chunks into four categories, summarized below.

1. **Words and Polywords:** Words and short, idiomatic groups of words, e.g. *if you please, give up*

2. **Collocations:** Groups of words that occur together frequently, such as: *stormy weather, slippery slope*, etc.

3. **Institutionalized Utterances:** Medium to sentence-length phrases which tend to be highly idiomatic with low variability. They are mainly used in spoken discourse and stored as wholes in memory. Example include phrases like: *gotta go, what do you mean* and less 'phrase-like' chunks such as *if I were you, I'd....*

4. **Sentence Frames and Heads:** Quite variable in terms of length, these chunks generally help structure written discourse, e.g. sequencers like *firstly, ..., secondly, ...* , phrases like *as mentioned above*, and even longer frames which provide structure for an entire text.

Though Lewis' category names are quite similar to the names of Nattinger and Becker, this similarity is somewhat misleading, as the actual definitions differ quite a bit. For example, what Lewis terms 'polywords' seem to fit more into Nattinger and Becker's deictic locutions category, while Nattinger and Becker's sentence builders fall into both the institutionalized utterances and sentence frames and heads categories of Lewis. Another major difference in Lewis' definition is that it includes individual words as well as groups of words, and finally, it does not include any equivalent to Nattinger and Becker's verbatim texts category. Despite these differences, taken as a whole, Lewis's categories do cover most of the same territory as Nattinger and Becker's; the notable exceptions are the individual words in Lewis's definition and the verbatim texts in Nattinger and Becker's definition, which could be seen as two points at the end of a continuum of length, into which the other categories fall.

While other categorizations exist, the two detailed above are perhaps the most well-known, and they are also more detailed than most in their explanations. Other categories of linguistic expressions that have been noted as types of lexical chunk include: aphorisms, clichés, collocations, compound nouns and verbs, conventional expressions, epithets, euphemisms, exclamations, expletives, frozen collocations, frozen phrases, grammatically ill-formed collocations, greeting and leave-taking rituals, idioms, jargon, memorized sequences, prepositional and adverbial locutions (e.g., 'because of', 'once in a while'), proverbs, quotations, routine formulae, sayings, similes, swearing, small talk, social control phrases, and technical terms (Barker and Sorhus 1975) (da Silva et al. 1999) (Moon 1998) (Pawley and Syder 1983) (Peters 1983) (Sinclair 2004) (van Lancker-Sidtis 2009) (Yorio 1980).

The list of lexical chunk types shows how varied chunks can be, not just in form but also in function and even in definition. Some categories, such as compound nouns and verbs, are fairly easy to define in syntactic terms, while other categories, like memorized sequences, clichés, proverbs and quotations, are better defined in psychological and/or socio-cultural terms. Still other categories, such as greeting and leave-taking ritual, social control phrases, and small talk, are best defined pragmatically, while others fall somewhere in between these categories or must be defined in even different ways. Despite these differences, it is clear from the list that one of the key elements of lexical chunks is their formulaic nature. Words like 'idiomatic', 'non-compositional', 'pattern', 'routine', 'fixed', 'frozen', and 'memorized' appear frequently in the descriptions, and many non-categorical definitions focus on formulaicity as a defining property of lexical chunks. In the following section I will outline some of the most important of these non-categorial, intensional, definitions.

2.2.3 Honing in: Defining Properties of Lexical Chunks

As formulaicity can be defined many ways, lexical chunks can also be defined in many ways, including linguistically, pragmatically, psychologically, socio-culturally, and statistically. One common type of definition is psychological, in which chunks are generally defined as groups of words that are stored as a whole in the minds of speakers. We have already seen one such definition from Peters, who defines a speech formula as "a multimorphemic phrase or sentence that, either through social negotiation or through individual evolution, has become available to a speaker as a single prefabricated item in his or her lexicon" (1983: 2), but the concept of chunks which are stored as wholes in the minds of speakers goes back at least to Jespersen, who distinguished between formulas - memorized phrases that allow for very little lexical and intonational variation - and free expressions, which are built up from individual words (1924). Jespersen also noted that some formulas are freer than others, in that certain words can be substituted in certain places in the chunks. For example, in the phrase "Long live the King", various other subjects can be substituted for "the King", but the words 'long' and 'live' are invariable.

Other definitions that emphasize the psychological basis of lexical chunks include those of Wray, who describes formulaic sequences as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (2002: 9), and Wood, who also uses the term formulaic sequence to refer to "multiword units of language that are stored in long-term memory as if they were single lexical units" (2002: 2). While such psychological definitions are clear-cut, they are also difficult to apply in linguistically determining what is a lexical chunk and what isn't. In order to use such definitions, one would have to rely on evidence from either human evaluations, which is time-consuming to collect, or from actual neurological data, which is not only time-consuming to collect but also quite costly. Additionally, there is a problem of objectivity: what counts as a lexical chunk for one person may not count as a lexical chunk for another.

At the opposite end of the spectrum are definitions which use only statistical properties of lexical chunks to define them. These definitions are not only clear-cut but also easily applied to data to determine what is or is not a lexical chunk, and for this reason they are often used in corpus work on lexical chunks. In the *Longman Grammar of Spoken and Written English*, a corpus-based exploration of English grammar, Biber et al. define lexical bundles as sequences of three or more words that occur frequently (above a certain number of times per million words). Work that uses more sophisticated statistical measures similarly relies on the intuition that the key property of lexical chunks is that their chunk parts co-occur frequently, e.g. in Dias, who defines a multiword lexical unit as "a group of words that occur together more often than expected by chance" (1999, 1).

While statistical definitions offer an advantage in that they are easy to implement in computational systems, it is unclear that they correspond to actual psychological phenomena. Furthermore, frequently occurring word combinations often do not fit into neat syntactic, semantic, or pragmatic categories, as for example the combination *of the*, which occurs 83,417 times in the 100 million word British National Corpus.

Finally, some linguistic definitions seek to define properties of chunks - be they phonological, morphological, syntactic, semantic, or pragmatic - that set them apart from other pieces of language. Weinert (1995) offers the following criteria for identifying lexical chunks, or as she terms them, formulaic language:

1. Phonological coherence: lexical chunks are spoken without hesitations. The intonation contour is smooth.
2. Greater length and complexity of sequence as compared to other output
3. Non-productive use of rules underlying a sequence

4. Community-wide use of a sequence
5. Idiosyncratic/inappropriate uses of sequences (relating specifically to learner language)
6. Situational dependence: certain chunks are used only in certain situations.
7. Frequency and invariance in form

These criteria cover a range of areas, not just linguistic, but also sociological, psychological, and statistical. Linguistically, chunks are longer and more complex than other linguistic phenomena, they are phonologically fluent, and pragmatically related to specific contexts. Across different uses, they are relatively unchanging in form, and because the rules used to create them are non-productive, they may contain rare and archaic forms (as for example, the use of the subjunctive in the phrase "Long live the King"). Though Weinert herself admits these criteria are not exhaustive, they help give a picture of what sets lexical chunks apart linguistically, and because they cover a range of areas, they can be used in a variety of contexts and applications.

2.2.4 Lexical Chunk Definitions: A Summary

An examination into the literature on lexical chunks quickly reveals that there is no single agreed-upon definition of what constitutes a chunk. Definitions range from the purely statistical to the linguistic to the psychological, and many definitions include criteria from multiple areas. Most, though not all, linguistic definitions tend to be extensional, listing different categories that can fall under the heading of lexical chunks; meanwhile, statistical and psychological definitions are generally intensional, focusing on key properties of lexical chunks that set them apart from other linguistic phenomena. Broadly speaking, the most important of these properties are: frequency, non-compositionality, and being stored and retrieved as a unified whole in human memory.

2.3 Evidence of Lexical Chunks

Though, as we have seen, lexical chunks are difficult to define in precise terms, evidence from neurological, psychological, and linguistic studies confirms their existence and their importance in human language. In this section, I will review some of the evidence for the existence of lexical chunks as a distinct linguistic phenomenon, focusing on their use in language processing and language acquisition and finally looking at neuroscientific evidence for the existence of lexical chunks.

2.3.1 Lexical Chunks in Language Processing

An important argument for the existence and importance of lexical chunks has been that they allow language users to process language more efficiently, both in production and in comprehension. Given a natural language grammar and its corresponding lexicon, the set of sentences one could hypothetically generate is infinite, yet, as Pawley and Syder note, "native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting nativelike control of the language" (1983, 193). Not only would speakers not be judged as sounding non-nativelike, if they were to make full use of the combinatorial power of the rules and words available to them, the processing task they would face in using their everyday language would be enormous. If, as Pawley and Syder suggest, much of language is actually made up of prefabricated chunks which are either invariable or allow for limited transformations, substitution of certain words, etc., the processing load for speakers and listeners would be greatly decreased.

Supporting this theory, in a study of livestock auctioneer speech, Kuiper and Haggo (1984) found that this speech was almost entirely made up of chunks, (which they term *oral formulae*), and they attribute this to the high processing demands faced by the auctioneers. They hypothesize that by relying on a small set of low-flexibility phrases, auctioneers are able to speak fluently without pauses or hesitations for long periods of time and to meet the very specific demands of the high-pressure auction situation. Though people in everyday situations do not need to meet such demands, they do still need to speak fluently enough to hold their listener's attention and get across their ideas, and they need to be able to comprehend speech quickly in order to keep up with the conversation. The smaller the processing load they have to deal with, the easier these tasks will be.

If lexical chunks are in fact useful in decreasing the processing load on human memory, then they should be stored and retrieved faster than corresponding units of language that are not lexical chunks. Though there have not been extensive studies on the processing of lexical chunks, those that have been conducted have indeed found such evidence. In a measurement of reaction times to grammaticality judgments, Jiang and Nekrosova (2007) found that participants (both native and nonnative speakers) responded more quickly and made few errors when the sequences to be judged were formulaic than when they were nonformulaic. Similarly, in Conklin and Schmitt (2008), participant reading times were significantly faster for idioms than for control phrases of similar length and structure. The study found this effect even when the idioms were presented in a context which primed their literal interpretation as opposed to their idiomatic meaning, suggesting the effect is indeed a lexical one, and not just a semantic one.

In a self-paced reading task, Tremblay et al. (2011) found that lexical bundles were read more quickly than similar groups of words that did not make up lexical bundles. They also found that sentences containing lexical bundles were recalled accurately more often than sentences that did not contain lexical bundles, and participants judged them as making more sense. Millar (2011) found that sentences containing non-nativelike word choices were read more slowly by native speakers than sentences containing nativelike word choices (e.g., *ideal partner* vs. the non-nativelike *best partner*). All of these results suggest that lexical chunks are indeed processed more efficiently than groups of words that are not chunks, and thus they may aid in the production and comprehension of fluent language.

2.3.2 Lexical Chunks in Language Acquisition

Other evidence for the existence of lexical chunks comes from studies in language acquisition. Lexical chunks have been found to be used by children learning their native language (Lieven et al. 2009) and by children and adolescents learning a second language (Hakuta 1974) (Fillmore 1976) (Myles et al. 1998) (Perera 2001). Some researchers have suggested that lexical chunks are particularly useful in language acquisition because they are first learned as unanalyzed wholes and then eventually broken down into their constituent parts, enabling learners to figure out grammatical rules. For example, de Villiers and de Villiers (1978) note that the negative contractions *don't*, *can't*, and *won't* are among the first auxiliaries produced by children learning English as a first language, yet the forms *do*, *can*, and *will*, along with grammatical variants like *doesn't* do not appear until much later. When these forms do begin to appear, children develop the full system of English auxiliaries shortly thereafter (97). The widespread presence of this type of sequence in children's language development lends credibility to the idea that the breakdown and analysis of lexical chunks aids children in learning their language's grammars.

Further evidence of the importance of chunks in child language acquisition comes from a study by Lieven et al. (2009). Using corpus data from the speech of two-year-old children learning English, they found that a large proportion of the children's multiword utterances produced over a two-hour period could be traced back to utterances they had produced previously. For the four children examined, between 20 to 50% of the utterances produced in the testing period exactly matched previously produced utterances, while between 50 to 80% of the utterances could be traced back to previously produced utterances when one operation was allowed to change a multiword unit (allowable operations were the substitution of one word for another and the addition of a word to the beginning or end of an utterance).

Perera (2001) also found evidence of chunks in the language of children learning English as a second language. In a study of four Japanese children learning English, she found that the children used many prefabricated language chunks which were gradually broken down into more creative forms (for example, one child first learned the chunk *more cracker please* and then broke it down to create phrases like *more apple please* and *more salad please*). Her findings further support the hypothesis that chunks aid in acquisition not only because they help a learner achieve fluency, but also because they can help a learner internalize grammatical rules.

Additional support for this hypothesis comes from a dissertation by Wong Fillmore (1976). In a year-long study, the author recorded the speech of five children, all native Spanish speakers, learning English as a second language through mere exposure (without specific instruction). She then exhaustively analyzed the speech of these children and found evidence of both the heavy prevalence of formulaic speech in the children's language and of the usefulness of this speech in language learning. Fillmore notes that formulaic speech is useful in multiple ways. For one thing, it allows non-native speakers to communicate with native speakers before they have achieved the grammatical and lexical knowledge that would allow them to express themselves as completely as they might wish. This in turn encourages native speakers to continue to interact with the non-native speakers, thus providing them more opportunities for language practice and language learning. The other major function of formulaic speech is the one noted above - that this speech, once learned, is later broken down and analyzed into its parts, thus aiding in the acquisition of syntax and lexical items.

2.3.3 Lexical Chunks in the Brain

Evidence that lexical chunks are stored and processed separately in the brain comes particularly from studies of people with aphasia - language disorders caused by stroke or other injury to the brain. Van Lancker-Sidtis and Postman (2006) found that people with damage to the left hemisphere produced a greater proportion of formulaic expressions in their speech than a control group of non-aphasic subjects, while people with damage to the right hemisphere produced fewer formulaic expressions than the control group. Right hemisphere damage was also associated with a greater production of proper nouns, whereas subjects with left hemisphere damage produced relatively fewer proper nouns. These findings suggest that the right hemisphere is somehow involved in the processing of lexical chunks and furthermore, that lexical chunks are not processed in the same way as proper nouns.

Other studies have shown that right hemisphere damage is associated with an impaired ability to understand metaphor (Winner and Gardner 1977), idioms (Myers and Linebaugh 1981), familiar phrases (van Lancker and Kempler 1987), jokes (Brownell et al. 1983), and verbal irony (Molloy et al. 1990). Such findings are particularly interesting in light of the fact that most language processing is taken to be localized in the left hemisphere. Broca famously noted the correspondence between destruction of particular areas of the left frontal lobe and an inability to produce articulated language (1861, 1865), and a century later, experiments on split-brain patients revealed a similar inability of patients to verbally describe objects that had been presented solely to the left field of vision, that is, to the right hemisphere (Gazzaniga 1967). In addition to playing a dominant role in language production, the left hemisphere has been especially implicated in the processing of lexical-semantic and syntactic information (Gazzaniga et al. 2002).

Despite the acknowledged importance of the left hemisphere in language processing, research suggests that certain linguistic functions, particularly those related to prosody, broad semantic association, early acquisition, and pragmatic inference, are predominantly localized in the right hemisphere (see Beeman & Chiarello 1998 and Lindell 2006 for a review). Whether these functions are related to the processing of formulaic speech is an intriguing question - certainly prosody is a likely candidate, if we recall one of Weinert's criteria for determining formulaic language is phonological coherence. The role of the right hemisphere in language acquisition could also be related to its role in lexical chunk processing, as it has been shown that lexical chunks are important in language acquisition.

In summary, a wide body of evidence from aphasic and other brain-damaged patients, along with physiological data obtained from such means as fMRI and PET scans, suggests that the right hemisphere does play a role in language processing and that lexical chunks and related phenomena are mostly processed in this hemisphere. The separate loci of processing for these types of speech versus non-formulaic speech suggests that lexical chunks are indeed not just parts of ordinary language, but a unique phenomenon that deserve further study and attention.

2.4 Applications of Lexical Chunks

Data from many sources have shown that lexical chunks exist as distinct phenomena in the brain and in language as it is used by people. The usefulness of lexical chunks as aids in efficient language processing and language acquisition has also been demonstrated in a wide variety of contexts. However, it remains to be shown how lexical chunks can be of use in terms of concrete applications. Even if having a mental lexicon of these chunks helps people learn and use language more efficiently, would it help students of a foreign language to give them a list of such chunks to memorize? And are there other areas where having a corpus of language-specific lexical chunks could be useful? In the following section, I will outline some ways in which a lexical chunk corpus could indeed be useful, not only in language teaching, but also in areas of Natural Language Processing (NLP).

2.4.1 Should Lexical Chunks be Taught?

Second language learners face a huge task - the mastery of a complex system of grammar, thousands of new words to be learned, unfamiliar sounds, in some cases different writing systems. With limited lesson time and student attention spans, teachers would not want to spend valuable time teaching lexical chunks if knowledge of the chunks does little to improve students' ability to communicate in the new language. Wong Fillmore (1976) has already suggested two important ways in which use of lexical chunks can aid acquisition: by providing learners with grammatically well-formed wholes which they can break down and analyze to help them learn the syntax of the language and by giving learners a starting point for communication with native speakers, which in turn encourages the native speakers to interact more with learners, thus giving the learners more opportunities for language practice and improvement.

Unfortunately, many second language learners have limited or no access to native speakers, so the usefulness of lexical chunks as starting points for native-nonnative interaction may be irrelevant. However, the usefulness of lexical chunks in acquisition of syntax could still be helpful, as suggested in Wong Fillmore and other studies on second language acquisition discussed earlier, such as Perera (2001) and Myles et al. (1998). In this last, the researchers examined the production of spoken French from a group of native English-speaking adolescents learning the language. Learners' output was collected over a period of 2 years, and the researchers looked at three lexical chunks in particular. They found that these chunks, used extensively in early production, were indeed broken down later as their parts were combined with other words to form novel utterances.

Another reason to teach lexical chunks to foreign language learners is that higher production of such chunks has been associated with greater fluency in the target language. Zhao (2009) found a correlation between use of lexical chunks and proficient language production, as measured by a writing test, in native Chinese speakers learning English, and Hsu (2007) found a significant correlation between frequency of lexical collocations and oral proficiency scores for native Taiwanese participating in an impromptu speech contest in English. However, Zhao also found that the Chinese speakers had poor knowledge of English lexical chunks overall, and the general failure of adult learners to master lexical chunks of the foreign language being learned is well documented (see Wray 2002 for a review). As Wray notes: "the formulaic sequences used by native speakers are not easy for learners to identify and master, and... their absence greatly contributes to learners not sounding idiomatic (2002: 176). This sentiment is also reflected in Weinert's fifth criterion for lexical chunks: inappropriate use by learners.

Knowledge of lexical chunks is then desirable for language learners who wish to sound fluent, and it also appears to be difficult for adult learners to pick up. Targeted instruction of lexical chunks could be useful in improving proficiency and fluency. In a study designed to test such an hypothesis, Boers (2006) compared two groups of upper-intermediate/advanced learners of English in Brussels. Both groups were given the same language learning materials over a course of eight months (22 teaching hours). One group was specifically instructed to pay attention to "standardized word combinations", while the other group was not; apart from this, there were no differences in instruction method. Oral proficiency tests at the end of the course revealed significantly higher scores in the group of students who had been instructed to pay attention to word combinations. Analysis of the spoken output of all the students in interviews also revealed a correlation between frequency of formulaic sequences used and the oral proficiency test scores.

Other studies demonstrating the effectiveness of instruction in improving chunk and collocational knowledge include Chan & Liou (2005), Wood (2009), Fahim & Vaezi (2011), and Osman (2009). In this last, Osman found that Malaysian students who were taught a list of lexical phrases achieved improved scores on their ability to communicate in a group task. Additionally, the students reported feeling more confident and comfortable in communicating in English in response to questions about whether and how the phrases helped them in the group discussions. In a study on Turkish children learning English, Bircan (2010) found that teaching the children vocabulary items by presenting them in phrases and having the children practice those phrases led to increased vocabulary retention as compared to when the items were presented and practiced individually. i.e., not in phrases.

All of these studies suggest that explicit instruction on lexical chunks is helpful in improving student proficiency, fluency, and confidence in communicating in a foreign language. Instruction can be of many forms including drilling, *noticing*, i.e., instructing students simply to look out for chunks in reading or speech (as in Boers 2006), highlighting chunks in texts, and exercises designed specifically to help students practice memorizing and using chunks. All of these methods require a database of lexical chunks in the target language, and so it seems that for language learning at least, the automatic compilation of such a database is a useful task.

2.4.2 NLP Applications of Lexical Chunks

While most research on lexical chunks has been carried out within the framework of Second Language Education, this does not mean that lexical chunks are not of use in other areas. In particular, many Natural Language Processing (NLP) applications could be well-served by a database of lexical chunks. Knowledge of lexical chunks through a database has been shown to improve performance in NLP applications related to part-of-speech tagging (Constant and Sigogne 2011), parsing (Nivre and Nilsson 2004), Machine Translation (Ren et al. 2009), Word Sense Disambiguation (Finlayson and Kulkarni 2011), and Information Retrieval (Acosta et al. 2011)(Michelbacher et al. 2011)

(Vechtomova and Karamuftuoglu 2004).

For example, Nivre and Nilsson (2004) ran a deterministic dependency parser on Swedish text in two versions: one that had been trained to recognize multiword units (which had been manually annotated), and one that had not. They found that the multiword unit-aware parser produced more accurate parses than the non-aware parser. Of particular interest, Nivre and Nilsson found that parsing accuracy improved not only for the multiword units themselves, but also for the syntactic structures surrounding those units. By identifying multiword units as such, automatic parsers are more limited in the parses they can produce for surrounding structures, and this limitation should generally be favorable, as it reduces the number of possible incorrect parses (assuming of course that multiword units have been correctly identified).

In another example of the usefulness of lexical chunk knowledge in NLP applications, Ren et al. (2009) found that by adding a bilingual phrase table of multiword units to the phrase table normally produced in Moses, a phrase-based statistical Machine Translation system, BLEU scores improved for Chinese-to-English translation in two different domains: medicine and chemical industry. In yet another example, Michelbacher et al. (2011) adjusted an Information Retrieval system to recognize non-compositional phrases as single semantic units, and they found that this adjustment led to improvements in the number of relevant documents returned for queries.

Other areas where lexical chunk knowledge has been deemed an important component of NLP applications include bilingual dictionary building (Abu-Ssaydeh 2006) and automated essay scoring (de Oliveira Santos 2011). Abu-Ssaydeh notes that most translators working in Arabic-speaking countries are advanced, rather than native, speakers of English, and that though these translators may be quite advanced speakers of English, they often have difficulties with lexical chunks. He proposes that a bilingual Arabic-English dictionary of multiword units would be particularly useful in introducing native Arabic translators to English units they were previously unaware of and in improving the quality of their translations. De Oliveira Santos (2011) notes that identification of lexical chunks in essays is an important component of automatic essay scoring as number of lexical chunks used has been shown to correlate with language proficiency.

As interest in using lexical chunks in NLP applications increases, methods for automatically extracting these chunks have become more important, and a variety of methods have been explored, though many have only looked at specific types of lexical chunk. For the most part, these methods are statistical in nature, and they are trained on corpora from which they extract chunks in the target language by using different word association measures. In the following chapter, I will review the most widely used of these methods, and I will examine some of the results that have ben reported for automatic extraction of lexical chunks and related phenomena.

Chapter 3

Corpora and Computation

3.1 Lexical Chunks in Computation

With the advent of computers and the powerful processing and storage abilities they offer, Linguistics, and especially Corpus Linguistics, has undergone vast changes. Time-consuming experiments for gathering linguistic judgments from groups of native speakers and the type of intuitive theorizing referred to as "armchair Linguistics" are increasingly being replaced by corpus studies as a means to answer language-related questions. These studies, made possible through the widespread availability of large corpora of spoken and written language, seek to answer linguistic questions by examining the data of language as it has been used by thousands and even millions of speakers in everyday contexts. They are useful because they are drawn from a wide range of sources and because their data are real: actual language as it is used by actual people. Experimental settings and native-speaker reflections may be prone to a variety of biases that can lead researchers to false conclusions; while corpus studies are certainly not immune to bias (for instance, the way a particular corpus was created is certainly extremely influential on the type of language it contains), a well-selected corpus can provide troves of linguistic information that would otherwise be extremely difficult to come by.

The increased interest in corpora as sources of linguistic information has gone hand in hand with an explosion in the field of Computational Linguistics, an area that barely existed a few decades ago. Though computational tools are useful in many areas of Linguistics, they are of special use to corpus linguists because most thorough explorations of the gigantic corpora now commonly relied upon for corpus studies require the speedy processing and huge memory capacity of modern computers. For example, the British National Corpus (BNC), one of the most widely used natural language corpora in corpus linguistic studies, contains roughly 100 million words. Without computational tools, even simple queries would require humans to spend many tedious hours poring through the data, and the likelihood of errors would be high. Using computational tools, such queries can be made quickly and easily, and the likelihood of errors can be substantially reduced.

Because large corpora of natural language can now be accessed easily and efficiently with computational tools, they are a good source of information about the types of lexical chunks people commonly use. However, research into the extraction of lexical chunks from corpora has, until recently, been limited; most of the earlier work in this area deals with other types of language, particularly collocations. As lexical chunks have gained attention in the Computational Linguistics community, methods for their extraction have been employed, but many of these methods rely on the previous work on collocations. In the remainder of this section, I shall review some of these methods, and I will discuss some of the results that have been obtained for lexical chunk extraction by other researchers.

3.2 Statistical Methods used in Automatic Extraction

As noted in Chapter 2, an important characteristic of lexical chunks is their fixedness. This fixedness manifests itself in a many ways. For example, in a discussion of fixed expressions, Hudson (1998) lists four main criteria:

1. **Unexpected syntactic constraints on constituent parts**
   These include fixed word order, fixed article (compare *spill the beans* with **spill some beans*), and fixed number (for example *let the cat out of the bag* as opposed to **let the cats out of the bag*).
2. **Unexpected collocational restrictions within the expression**
   Fixed expressions do not allow for the substitution of lexical items with similar meanings (for example, **ill and tired* for *sick and tired*).
3. **Anomalous syntax or usage**
   This includes lexical items and grammatical constructions not normally used in the language, such as *handbasket* in *go to hell in a handbasket*, or the subjunctive in *long live the King*.
4. **Figurative meaning**
   Many fixed expressions do not receive a literal interpretation, as in the expressions *on pins and needles* (meaning anxious), *all broke up* (meaning very upset), and *grandfather clock* (referring to a specific type of clock.

Broadly speaking, all of these criteria have the result that, within fixed expressions, the particular lexical items in their particular order should occur more frequently than would be expected if the expressions were not fixed. Thus, one could expect to encounter *spill the beans* significantly more often than *spill some beans*, as compared to the relative frequencies of say *spill the cookies* and *spill some cookies*.

Statistical measures for the extraction of lexical chunks and related phenomena, such as collocations, rely on this notion that chunks will tend to be groups of words that appear together more often than would be expected by chance. However, there are multiple ways to translate this notion into mathematical terms. The most common measures for collocation and chunk extraction deal with raw frequency, mutual information, and hypothesis testing. These measures, and a few others that have proved useful, are discussed below.

3.2.1 Frequency

As noted in chapter 2, one of the main features of lexical chunks is their frequency. The lexical bundles found in Biber et al. (1999) are simply strings of 3 or more contiguous words that occur above a certain frequency. For their work, Biber et al. define strings as frequent if they occur at least ten times per million words in a given register (spoken or written), and if they occur in at least five different texts in that register. Five and six-word sequences need only occur five times per million words[1] (1999, 992-3). This technique of identifying chunks based only on frequency has since been used by a number of other researchers.

Chunks found using Frequency

The work by Biber et al. identified some interesting properties of lexical bundles. They found that 30% of the words in conversation occurred in such bundles, while only 21% of the words in written academic texts occurred in bundles[2]. Further, most bundles were short: words occurring in 3-word bundles made up 25% and 18% of the total words in conversation and academic prose respectively, while words occurring in 4-word bundles accounted for only 3% and 2% of the total words in the different registers.

The most common 3 and 4-word lexical bundles found in conversation and academic prose are listed in Table 3.1. These bundles exemplify patterns typical of the majority of lexical bundles found by Biber et al. For example, most of the lexical bundles did not form complete structural units; rather, they tended to bridge two structural units. Additionally, the most common structural types of bundles in conversation were quite different from the most common structural types in academic prose. Bundles of the form personal pronoun + lexical verb

	Conversation	**Academic Prose**
3-word bundles	*I don't know, I don't think, do you want, I don't want, don't want to, don't know what, and I said, I said to, I want to, you want to, you have to, do you know, you know what, have you got, what do you, I mean I, have a look*	*in order to, one of the, part of the, the number of, the presence of, the use of, the fact that, there is a, there is no*
4-word bundles	*I don't know what, I don't want to, I was going to, do you want to, are you going to*	*in the case of, on the other hand*

Table 3.1: *Most common lexical bundles (Biber et al. 1999: 994)*

---

[1] Biber et al. do not look at sequences of more than six words.

[2] When contractions such as *don't* were counted as two words, the percentage of words occurring in bundles in conversation increased to 45%.

phrase (+ complement clause), as in *I don't know what*, were by far the most common of the 4-word bundles in conversation, making up 44% of these bundles, whereas they hardly appeared at all in academic texts. By contrast, the most common structural types of 4-word bundles in academic prose were preposition + noun phrase fragment (e.g., *as a result of*), making up 33% of bundles, and noun phrase with post-modifier fragment (e.g., *the nature of the*), making up 30% of the bundles. These types made up only 3% and 4% of the 4-word bundles in conversation.

Problems with the Frequency approach

One of the issues with using frequency as a measure is that it only finds chunks that contain common words like *the, what,* and *of*. Chunks containing rare words, such as proper nouns or certain idioms, will not be found. Additionally, it has been suggested that many chunks - even chunks containing common words - appear infrequently despite their status as chunks. In an extremely thorough corpus examination, Moon (1998) found that 93% of all fixed expressions and idioms (identified from a previously assembled database) appeared fewer than 5 times per million words; in fact, 40% of these chunks appeared fewer than 5 times in the entire 18 million word corpus. The huge percentage of infrequent chunks suggests that using raw frequency to find chunks will be ineffective and that more sophisticated statistical measures are necessary.

3.2.2 Mutual Information

In a discussion of collocation, linguist J. R. Firth is famously quoted as saying that "You shall know a word by the company it keeps" (1968: 179). The notion that lexical items are best understood through the lexical items that commonly surround them has proved quite influential in studies of collocations, and, in a 1990 paper, Church and Hanks translate this notion into statistical terms by using Pointwise Mutual Information, a statistical measure drawn from Information Theory, to automatically extract collocations from a corpus. The principle behind this idea is that collocating words will be much more likely to occur together than would be predicted by chance. Formally, if we take a collocation like *squeaky clean*, and call *squeaky* word one (w_1) and *clean* word two (w_2), then the Mutual Information (I) between w_1 and w_2 is given by (3.1).

$$I(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{3.1}$$

If the two words are collocates, it is presumed that their joint probability, $P(w_1, w_2)$, will be higher than than the combined probabilities of observing the two words independently, and thus $I(w_1, w_2)$ will be greater than 0. If the words are not collocates, $I(w_1, w_2)$ should be approximately equal to 0.

Chunks found using Mutual Information

In practice, it is quite rare to find chunks with a Mutual Information score less than 0 because human language is regular: adjectives tend to precede nouns more than verbs, verbs precede prepositions more than articles, etc. Thus, it is necessary to find some cutoff above which word pairs can be considered actual collocations. Data from Church and Hanks and from other studies using Mutual Information suggest this cutoff should be somewhere in the range of 2-4. For example, Table 3.2 gives the MI scores found by Church & Hanks for phrasal verbs beginning with *set* in the 1988 AP Corpus (44 million words), and Table 3.3 gives the MI scores for bigrams of frequency 20 found by Manning and Schütze (1999) in a 14 million word corpus of text from the New York Times newswire.

verb + preposition	***I***
set up	7.3
set off	6.2
set out	4.4
set in	1.8
set on	1.1
set about	-0.6

Table 3.2: *Mutual Information scores for phrasal verbs using* set *(Church & Hanks 1990:25)*

bigram	***I***	$f(w_1)$	$f(w_2)$
Ayatollah Ruhollah	18.38	42	20
Bette Midler	17.98	41	27
Agatha Christie	16.31	30	117
videocassette recorder	15.94	77	59
unsalted butter	15.19	24	320
first made	1.09	14907	9017
over many	1.01	13484	10570
into them	0.53	14734	13478
like people	0.46	14093	14776
time last	0.29	15019	15629

Table 3.3: *Mutual Information scores for 10 bigrams of frequency 20 (Manning & Schütze 1999:167)*

Problems with Mutual Information

A big advantage of Mutual Information over raw frequency is that Mutual Information scores can be high for chunks or collocations that occur infrequently; thus, chunks involving rare words can still be found. On the other hand, Mutual Information scores for such chunks can often be overinflated, making some combinations appear to be chunks simply because the words they contain happen to only occur together in the dataset. For example, Manning & Schütze found that bigrams like *Schwartz eschews* and *fewest visits*, which occurred only once in the first 1000 documents of their corpus, received high MI scores because they contained words that also occurred infrequently in this subcorpus. Even when they extended the corpus to include all 23,000 documents, Manning & Schütze found that these bigrams still only occurred once and thus had overinflated MI scores. On the other hand, collocations involving very frequent words may receive scores that are too low.

3.2.3 Log Likelihood

Another measure that has been proposed for collocation-finding is the likelihood ratio, which is a measure of how likely one hypothesis is as an explanation for the data over another (Dunning 1993). For two hypotheses, H_1 and H_0, (the log of) this ratio is given by (3.2).

$$\log_2 \lambda = \log_2 \frac{L(H_0)}{L(H_1)} \tag{3.2}$$

In the case of deciding whether a bigram is a collocation or not, the two hypotheses being compared are the null hypothesis - that the second word's occurrence is independent of the first word's occurrence - and the hypothesis that there is a relation between the two words, i.e., the second word's occurrence is dependent on the first word's occurrence. As one can assume a binomial distribution (given in (3.3)) of words with a large enough corpus, these likelihoods are formalized for a corpus in (3.5) and (3.6), with f_1 = the frequency of the first word, f_2 = the frequency of the second word, f_{12} = the frequency of the bigram, N = the total number of words in the corpus, and p, p_1, and p_2 as given in (3.4).

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{3.3}$$

$$p = \frac{f_2}{N} \qquad p_1 = \frac{f_{12}}{f_1} \qquad p_2 = \frac{f_2 - f_{12}}{N - f_1} \tag{3.4}$$

$$L(H_0) = b(f_{12}, f_1, p) b(f_2 - f_{12}, N - f_1, p) \tag{3.5}$$

$$L(H_1) = b(f_{12}, f_1, p_1) b(f_2 - f_{12}, N - f_1, p_2) \tag{3.6}$$

The full log likelihood ratio is then given by (3.7).

$$\log_2 \lambda = \log_2 \frac{L(H_0)}{L(H_1)} = \log_2 \frac{b(f_{12}, f_1, p) b(f_2 - f_{12}, N - f_1, p)}{b(f_{12}, f_1, p_1) b(f_2 - f_{12}, N - f_1, p_2)} \tag{3.7}$$

Chunks found using Log Likelihood

The top-scoring bigrams involving the word *powerful* found by Manning & Schütze are shown in Table 3.4. Unlike the chunks found with Mutual Information, some of the highest scoring chunks found with Log Likelihood contain frequent words. At the same time, chunks containing infrequent words are also found, but their scores are not as inflated as with Mutual Information. These improvements result from the assumption of a binomial distribution of data, more appropriate for language than the assumption of a normal distribution. As Dunning notes: "Statistics based on the assumption of normal distribution are invalid in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only the very most common words" (1993: 71).

bigram	$-2 \log \lambda$	$f(w_1)$	$f(w_2)$	$f(w_1 w_2)$
most powerful	1291.42	12593	932	150
politically powerful	99.31	379	932	10
powerful computers	82.96	932	934	10
powerful force	80.39	932	3424	13
powerful symbol	57.27	932	291	6
powerful lobbies	51.66	932	40	4
economically powerful	51.52	171	932	5
powerful magnet	50.05	932	43	4
less powerful	50.83	4458	932	10
very powerful	50.75	6252	932	11
powerful position	49.36	932	2064	8
powerful machines	48.78	932	591	6
powerful computer	47.42	932	2339	8
powerful magnets	43.23	932	16	3
powerful chip	43.10	932	396	5
powerful men	40.45	932	3694	8
powerful 486	36.36	932	47	3
powerful neighbor	36.15	932	268	4
powerful political	35.24	932	5245	8
powerful cudgels	34.15	932	3	2

Table 3.4: *Log Likelihood scores and frequency for top-scoring bigrams (Manning & Schütze 1999:163)*

Problems with Hypothesis Testing

One of the features of hypothesis testing that Manning & Schütze point out is that many high-scoring chunks are subject-specific. Thus, bigrams relating to newsworthy events in 1989 such as *Prague Spring* and *East Berliners* had quite high relative frequencies in the subcorpus of New York Times newswire from that year, but they had low relative frequencies in the following year, 1990. This creates a problem: in a subject-specific corpus, the Log Likelihood measure will find many chunks, but the corpus will be smaller, and chunks relating to other subjects will not be found. In a larger, balanced corpus, some of the chunks that could have been found in the smaller, specific corpus may no longer be found, due to low overall frequency (but high local frequency). This issue will be discussed further in the following chapters.

3.2.4 Other Methods of Chunk Extraction

Though many other statistical methods have been employed for the extraction of chunks (see Pecina 2005 for a thorough review of collocation extraction methods in particular), I will only detail two additional methods. Most methods have been used to find only specific types of chunks, such as certain types of collocations, but the two methods I describe below have both been used with some success in the extraction of broad-coverage multiword units. These methods are Mutual Expectation and Symmetric Conditional Probabilities, used in Dias et al. (1999) and da Silva et al. (1999). Their respective formulas are given below, in (3.8) and (3.9).

$$\frac{2f(w_1, w_2)}{f(w_1) + f(w_2)} \cdot P(w_1, w_2) \tag{3.8}$$

$$\frac{P(w_1, w_2)^2}{P(w_1)P(w_2)} \tag{3.9}$$

Mutual Expectation for a two-word chunk is given by the product of the probability of the chunk and the arithmetic mean of the marginal probabilities of the chunk. Though da Silva et al. and Dias et al. used a modified version of this measure, they found that it outperformed several other measures in the extraction of multiword units. Da Silva et al. also used SCP in the extraction of contiguous multiword units with some success. Symmetric Conditional Probability (SCP) for a two-word chunk is simply the product of the two conditional probabilities for each word appearing in the chunk. That is, it is the product of the probability of the second word in its position, given the first word, and the probability of the first word in its position, given the second word.

3.3 State of the Art in Automatic Extraction

Evaluation of methods for lexical chunk extraction is a tricky task, due to the fact that no single definition of the phenomenon exists. Experiments in lexical chunk extraction by different researchers often differ quite a bit in both the types of chunks they extract and the ways they determine whether these chunks are valid or not. This makes comparison between experiments very difficult. Many methods extract only very specific types of chunk, such as such as verb-noun collocations, or domain-specific compound nouns. Methods also vary in the length of chunks extracted, with several experiments reporting data for bigrams only. Though methods that are geared towards extracting specific types of chunks, such as verb-noun collocations, often extract chunks containing gaps, the vast majority of methods for extracting a broader range of chunk types restrict these chunks to only those containing contiguous words. A prominent exception is the work of Dias and colleagues, which shall be reviewed below.

The difficulties of comparing methods that extract different types of chunks are compounded by the differences in evaluation methods used by researchers. In similar areas of NLP, it is traditional to have a 'gold standard' reference body, against which results can be compared. For example, automatic parsing applications are compared to manually produced parses of a test corpus, or automatically translated texts are compared to translations previously produced by humans. However, as no gold standard of lexical chunks exists, most researchers have had to resort to either having humans check by hand all the lexical chunks found by their methods, a laborious and time-consuming task that cannot be easily repeated, or to only reporting qualitative results.

Another indirect, but often useful, method of lexical chunk evaluation is to use the chunks in some other application, such as parsing or Machine Translation, and see how much the application's performance is improved when different methods of chunk extraction are employed. This strategy is advantageous in that it does not need to rely on difficult-to-obtain human evaluations or chunk lists gathered from dictionaries, which tend to be incomplete. On the other hand, specific application-based evaluation is difficult for other researchers to repeat, unless they have access to the exact same application used by the original researchers.

In sum, it is not easy to say what the state of the art is for lexical chunk extraction because of the many differences in evaluation methods used and types of chunks extracted. A better idea of the current state of automatic extraction methods can be obtained by simply looking at some of the different results that have been reported and the methodology employed in those experiments. In the remainder of this section, I will review some of these results, explaining for each what type of chunks were extracted and how chunks were evaluated.

3.3.1 Restricted Chunk Types

When lexical chunks are restricted to very specific types, gold standard chunk lists can more easily be compiled, and so the standard measures of precision (percentage of chunks found that were correct) and recall (percentage of total possible chunks that found chunks account for) can be reported. It is common practice to report precision results for the n-best chunks (i.e., chunk with the highest score, according to whatever measure was used to extract them), and with relatively low n, precision can be quite good. In looking only at adjacent bigrams of adjective-noun combinations evaluated manually, Evert & Krenn (2001) obtained a maximum precision of 65% (using Log Likelihood) for the 100 highest-scoring combinations. However, when the number of combinations examined was increased to 500, precision for this measure dropped to 42.80%. Further, the chunks were evaluated by two human raters, and any chunk accepted by either of the annotators was considered a good chunk. This broad allowance for combinations to be accepted as chunks may thus have led to inflated precision scores.

Ngomo (2008) was able to use a previously existing gold standard of chunks in evaluation by extracting highly domain-specific medical terminology, for which the MESH (Medical Subject Headings) vocabulary is available. Using a measure called Smoothed Relative Expectation, Ngomo achieved a maximum precision of 29.40% for the 500 best terms, but recall was only 1.05%. Though the evaluation method is quite solid for this experiment, the chunk types are so restricted that it is difficult to generalize the results to most other types of chunks.

Another researcher who relied on a previously compiled chunk list for evaluation is Lin (1999), who used an algorithm based on Mutual Information to extract three types of collocation which were expected to be involved in idioms, namely: object-verb, noun-noun, and adjective-noun. Collocations for which the mutual information between the two words was significantly higher than the mutual information that resulted from replacing one of the words with a semantically similar word (obtained from a thesaurus) were extracted as likely chunk candidates. For evaluation, all the extracted collocations involving ten specific words (five high-frequency words and five lower-frequency words) were compared against idioms taken from two idiom dictionaries, the NTC's English Idioms Dictionary and the Longman Dictionary of English Idioms. Idioms were selected if their head word was one of the ten words that had been selected and if the idiom contained an object-verb, noun-noun, or adjective-noun relationship. Lin's results are displayed in Table 3.5.

As can be seen, recall and particularly precision scores differ noticeably between the two dictionary lists; this suggests that even gold standard lists can be unreliable in lexical chunk evaluation, unsurprising given the extent to which definitions of chunks and chunk-like phenomena differ and given the extremely wide range of items to be covered.

	Precision	Recall
NTC English Idioms Dictionary	15.7%	13.7%
Longman Dictionary of English Idioms	39.4%	20.9%

Table 3.5: *Precision and Recall for three types of collocation (Lin 1999: 320)*

This brief survey of results and evaluation methodology for automatic extraction methods geared towards extracting only specific types of chunks reveals difficulties that only become more pronounced when more types of chunks are considered. Namely, the wider the range of acceptable chunks, the less likely it is that human judgments and previously compiled lists of chunks will be reliable. Perhaps because of this, few results have been reported for methods which extract chunks of unrestricted type. The main exception is a series of experiments run by Dias and colleagues, described below.

3.3.2 Unrestricted Chunk Types

Dias and Guilloré (1999) used five different association measures to extract chunks of both contiguous and non-contiguous words. They determined precision scores through manual evaluation, counting chunks as good if they formed either grammatical or meaningful units (their terminology). Using this method, Dias and Guilloré obtained a maximum precision of roughly 90% using the Mutual Expectation measure. Instead of recall, the extraction rate is given, and a maximum of 3.5% is achieved (using Log Likelihood). Here, precision is quite good, but by counting everything that forms a grammatical unit as a chunk, Dias and Guilloré cannot distinguish between lexical chunks and merely grammatical chunks.

Dias and Vintar (2005) used Mutual Expectation to extract chunks in English and Slovene, and they again relied on manual evaluation, but they used a more specific definition of chunks. In this case, raters were asked to determine whether extracted chunks fell into one of the following categories: set phrases, phrasal verbs, adverbial locutions, compound determinants, prepositional locutions, and institutionalized phrases. Using this evaluation method, Dias & Vintar obtained a maximum precision of 14.5% for English chunks and 29.8% for Slovene chunks.

Similarly, da Silva et al. (1999) used several different measures to extract Portuguese chunks containing contiguous and non-contiguous words and counted as good all chunks which fell into one of the following categories: proper nouns, compound nouns, compound verbs, frozen forms, and "other n-grams occurring relatively frequently and having strong "glue" among the component words" (123). Using this methodology, da Silva et al. obtained a maximum precision of 81% for contiguous-word chunks (using SCP), and a maximum precision of 90% for non-contiguous-word chunks (using Mutual Expectation). As in (Dias et al. 1999), these results are quite good, but the fact that the experimenters themselves performed the evaluation and the broad definition of chunks may have contributed to the high precision scores.

Overall, few results have been reported for the automatic extraction of unrestricted chunk types, and even fewer of these results have used evaluation methods that rely on external sources, such as previously compiled chunk lists or ratings gathered from humans other than the experimenters. The practical aspects of finding chunk lists or asking humans to rate huge numbers of chunks naturally play a role in the scarcity of reliable evaluation metrics. As I aim to extract chunks of any type, these challenges will come up in evaluation, and I will address the method by which I attempt to deal with them in Chapter 5. Before this, I will describe the methodology I used to extract chunks and the materials used to train my system.

Chapter 4

Materials and Method

4.1 Materials

Lexical chunks were acquired from a subsection of the British National Corpus (BNC), consisting of 365 texts, with a total of 5,874,133 words. Of the 365 texts, 169 (approximately 46%) were spoken, and the other 196 were written. The spoken texts accounted for approximately 27% of the total words (1,574,651 words). The subsection used for lexical chunk acquisition consisted of a considerably greater percentage of spoken text than the full BNC, in which approximately 22% of the texts and 11% of the total words are spoken. The inclusion of relatively more spoken texts was purposeful, as it was hypothesized that lexical chunks would be more prevalent in these texts, following the observations of Biber et al. (1999).

4.1.1 Some Notes about the Data

Due to the specific format of BNC texts and other processing requirements, certain conventions were followed with respect to chunking which might not have otherwise been the most obvious choices. I outline these conventions below:

- All contracted forms (e.g., *we'll, I'm, gonna*) are treated as two or more separate words, with the break between words appearing in the location of the apostrophe. The possessive ending 's is also treated as a separate word.
- All punctuation other than apostrophes was omitted.
- All capital letters were switched to lowercase.

4.2 Method

4.2.1 Algorithm

Chunking was performed through an iterative algorithm, similar to one proposed by Wible et al. (2006). The chunker first scans the corpus for all chunks of exactly two words. What counts as a chunk is determined by the particular statistical measure being used and a pre-determined cutoff. For example, using a measure of raw frequency with a cutoff of 10, all two-word chunks that appear ten or more times in the corpus are selected and used as input for the next iteration. In that next iteration, the chunker makes three-word chunks by taking the two-word chunks as a base and adding the surrounding words to make new chunks of exactly three words[1]. The three-word chunks that score above the cutoff are then used as input for the next iteration and so on. Though chunking could in theory go on until no further chunks could be added, very long chunks tended to be idiosyncratic and not particularly useful, so I chose to look at only those chunks consisting of five words or fewer, following Smadja (1993).

Following the acquisition of all chunks consisting of up to five words, the chunker performed a post-processing step to remove subchunks, i.e., chunks that appeared in longer chunks. Subchunks were removed only if they both began with the same word as a longer chunk and had a lower score than the longer chunk. This was done to prevent the accidental removal of chunks that could function both on their own and as parts of other chunks. For example, the chunk *white silk shirt* was found to occur twice in the corpus, while its subchunk *white silk* occurred 14 times. Since the subchunk was more frequent than the full chunk, it is likely that both chunk and subchunk are acceptable lexical chunks, and thus neither should be eliminated. On the other hand, the subchunk *the X states* was found to occur exclusively within the chunk *the united states*, and it is thus a good candidate for elimination.

4.2.2 Gaps

The question of how to deal with gaps is always a tricky one. Many computational methods for extracting lexical chunks avoid the issue by only extracting contiguous sequences, as in Biber et al. (1999). While computationally simpler, this method is somewhat unsatisfactory, as one of the distinctive features of lexical chunks is their ability to contain gaps. Approaches that do allow for gaps are thus more desirable, but such approaches must deal with several issues that do not arise when chunks are defined as contiguous sequences of words. Chief among these issues are the number of gaps to be allowed and the question of whether chunks containing the same words but with different placement with respect to each other should count as separate or equivalent (for example the chunks *strands X hair* and *strands X X hair*, where X's represent gaps).

My chunker does allow gaps to appear within chunks, but the number of these gaps was limited to two (with each missing word counting as one gap,

---

[1] Surrounding words were defined as those words either following or internal to the chunk, if the chunk contained a gap. For example, a two-word chunk with one gap in it, such as *cup X tea* could be expanded with the word in the gap between *cup* and *tea*, or with one of the words following *tea*, either directly after it or after other gaps.

regardless of whether the missing words are contiguous or not). Though I originally set the chunker to search for chunks with up to four gaps, I found that this resulted in a lot of long and very unintuitive chunks. As the greater number of gaps was also computationally quite expensive, for the evaluation, I allowed for only two or fewer gaps. Number of gaps was varied from one to two in different trials to test whether it made a difference in performance.

I also chose to count chunks as different if they contained the same words but with different placement relative to each other, as in the *strands X hair* examples above. Though many approaches count such chunks as the same, I believe that this approach is better suited for the extraction of collocations than actual chunks. The line between collocations and lexical chunks is not a clear one, but if they are to be distinguished from each other, perhaps the key difference is that the placement of collocating words relative to each other is more variable than it is in chunks. For example, in the classic example of collocation, the words *strong* and *tea* can appear in many different positions relative to each other, as in phrases like “This is very **strong tea**”, “This **tea** is **strong**”, and “This **tea** is quite **strong**”. By contrast, standard examples of lexical chunks like *How do you do?* and *as X as* involve a much more fixed word order.

The iterative nature of my algorithm posed another issue for dealing with gaps, namely, when expanding chunks with gaps, should the words in the gaps be considered as possible chunk parts, or should the expansion only look at words following the last word of the chunk? I chose to consider words both internal to and following such chunks in the expansion in order to find as many chunks as possible. The chunking algorithm is summarized below.

Chunking Algorithm

1. Scan the text for all two-word chunks, and select those chunks that score above the cut-off.
2. Make all possible three-word chunks from the two-word chunks, and again select only those chunks that score above the cut-off.
3. Repeat step (2) to make chunks of four, five, and six words (or however many words is desired).
4. Remove chunks that are fully contained in other chunks which both begin with the same word and which received a score higher than the subchunk’s score.

4.2.3 Statistical Measures

I compared five different statistical methods for chunk acquisition, namely: Raw Frequency, Pointwise Mutual Information, Log Likelihood, Symmetrical Conditional Probability, and Mutual Expectation, discussed in depth in the previous chapter and summarized in Table 4.1[2]. These methods were chosen because they are among the most frequently and successfully used methods for

---

[2] The Log Likelihood formula given here is a simplified one, as the full formula was too complex to be displayed in the space of the table. See Chapter 3 for the full formula.

lexical chunk and collocation acquisition. Many of the measures have generally been used for the acquisition of two-word chunks only, but in an iterative algorithm they are easily modified for creating longer chunks if the base chunk is treated as a single word. Thus, where one would normally input the frequencies of word one and word two, one instead inputs the frequencies of the base chunk and the new word.

Statistical Measure	**Formula**
Raw Frequency	$f(w_1, w_2)$
Pointwise Mutual Information	$\log \frac{P(w_1,w_2)}{P(w_1)P(w_2)}$
Log Likelihood	$\log \frac{b(f_{12},f_1,p)b(f_2-f_{12},N-f_1,p)}{b(f_{12},f_1,p_1)b(f_2-f_{12},N-f_1,p_2)}$
Symmetrical Conditional Probability	$\frac{P(w_1,w_2)^2}{P(w_1)p(w_2)}$
Mutual Expectation	$\frac{2f(w_1,w_2)}{f(w_1)+f(w_2)} \cdot P(w_1, w_2)$

Table 4.1: *Formulae for different statistical methods*

Each statistical method was run with a variety of different cutoffs: generally, the lower the cutoff, the more chunks were found, but due to computer memory and processing limitations, a lower bound was almost always found to be necessary. This bound varied for different methods, but once it was found through trial-and-error, it was used, along with a range of higher cutoffs for comparison purposes. In evaluation, the lowest cutoff was always used, except in some cases where different varieties of the same measure were being compared, in which case the lowest cutoff common to both varieties was used. The lower-bound cutoffs used in the standard evaluation are shown in Table 4.2

Statistical Measure	**Cutoff**
Raw Frequency	10
Pointwise Mutual Information	2
Log Likelihood	10
Symmetrical Conditional Probability	5×10^{-6}
Mutual Expectation	5×10^{-9}

Table 4.2: *Cutoffs for different statistical methods*

I also experimented with using different word frequency cutoffs, i.e., only considering words as candidates for being in a chunk if those words occurred above a certain frequency in the corpus. Such a cutoff is particularly useful when using Mutual Information, which otherwise gives very high scores to chunks containing very infrequent words. However, in the end, it was found that a cutoff of 1 (meaning chunk words had to appear at least two times) was sufficient for methods that found chunks with a maximum of 1 gap, and a cutoff of 5 was sufficient for methods that found chunks with a maximum of 2 gaps.

A final variable in my chunking experiments was corpus division. To my knowledge, almost all methods of chunk and collocation acquisition that rely on corpora treat the corpus as a unified whole in training. Such an approach capitalizes on the computational capabilities that allow for the analysis of huge quantities of data. Though large corpora are certainly important in data-driven statistical methods, an important feature of lexical chunks is their high **local** frequency, as noted by Manning and Schütze (1999). This is particularly true of things like jargon and technical terminology, but it also applies to chunks like proper nouns, and more generally, I hypothesized that many chunks are likely to only appear in certain registers or in reference to specific topics. For this reason, I tested two methods of lexical chunk acquisition: the first relied on using the whole corpus as a base in chunk search, while the second split the corpus into the original 365 texts and looked for chunks in each text. Chunks found in the different texts were then compiled into one large list of chunks, so the large amount of data was still utilized, but in a different way from traditional methods.

Chapter 5

Evaluation Methodology

Chunk acquisition from the BNC data resulted in chunk lists of hundreds of thousands and occasionally over 1 million chunks. Because of the impracticability of evaluating such a large number of chunks for each measure, evaluation was performed in the following manner: an evaluation text was selected, and all 'good' chunks in that text were determined through a combination of human judgments and comparison with dictionaries, as described later on in this chapter. The chunk databases found by each statistical measure were then used by a program which went through the article and extracted all lexical chunks that occurred in the database. The final list of chunks extracted in this way was then compared to the pre-determined 'good' chunks for each measure, and values of precision, recall, and f-measure were calculated.

5.1 Evaluation Materials

The text used in evaluation was an article taken from the front page of the New York Times online version (Lichtblau et al. 2011). The article contained a total of 1488 words in 52 sentences. As stated above, for each statistical measure, the list of chunks that had been acquired from the BNC corpus was used by a program which went through the New York Times article and found all the lexical chunks in each sentence. Chunk-finding allowed the same word to appear in multiple chunks (so that overlapping chunks could be found), and chunks were listed multiple times if they occurred more than once in the article. Each output thus came in the form of a list of all tokens of chunks that had been found in the article by a given method.

It should be noted that because chunks were acquired from a corpus of British English, certain differences may have existed in the types of chunks found in acquisition and those present in the evaluation article, which was taken from an American newspaper. For the purpose of comparison of different statistical measures against each other, these differences were not particularly relevant, as all measures acquired chunks from the same data and extracted chunks from the same evaluation text. However, if one wishes to look at the values for precision, recall, etc. in an absolute sense, one should keep in mind that it is quite likely that all the measures performed worse in chunk extraction on the article in American English than they would have on an article written in British English. Worse performance could be attributed to several differences between American and British English including differences in word spellings, vocabulary (e.g., American *cookie* vs. British *biscuit*), phrases and common expressions used (e.g., American *parking lot* vs. British *car park*, and relative frequencies of those phrases and expressions that are common to both dialects.

5.2 Methodology

5.2.1 What Counts as a Chunk?

As noted in Chapter 3, because there is no clear definition for lexical chunks, the question of how to evaluate a measure that is meant to find lexical chunks is problematic. Past efforts at finding lexical chunks have dealt with this problem in various ways. Some simply avoid the issue by defining lexical chunks as exactly those that their system finds and then analyzing those chunks to see what sorts of conclusions can be drawn about them. This is the method employed by Biber et al. (1999), who predefine lexical bundles as contiguous strings of words occurring above a certain frequency and then give an in-depth analysis of the sorts of strings one finds in a corpus using such a definition. However, because the purpose of the present research is to compare different methods for finding lexical chunks, such an approach is clearly ill-suited.

The ideal means to evaluate the methods would of course be to compare the chunks they found against some agreed-upon standard database of lexical chunks. Previous attempts at evaluation of this sort were described in Chapter 3, and it was noted that many difficulties were encountered. Mainly, where previously existing databases can be found, they exist for only very specific types of chunks. When extraction is performed for unrestricted chunk types, databases cannot be used, and the chunks found must be evaluated by hand, either by the experimenters themselves or by external raters. When experimenters evaluate the chunks themselves, the reliability of ratings must come into question; using external raters solves this problem, but a method meant to find chunks of unrestricted type will generally return thousands and thousands of chunks, so time and rater fatigue become problems.

One approach to evaluation that attempts to reconcile these problems is that of Simpson-Vlach and Ellis (2010). They evaluated formulas found in academic text with different statistical measures by asking a group of language instructors and testers to rate a subset of the formulae on three characteristics, namely:

1. Whether the formula constituted 'a formulaic expression, or fixed phrase, or chunk'.
2. Whether the raters thought the formula had 'a cohesive meaning or function, as a phrase'.
3. Whether the raters thought the formula was 'worth teaching, as a bona fide phrase or expression'.

From these ratings, Simpson-Vlach and Ellis were then able to derive scores for the other formulae that represented how likely instructors would be to judge those formulae as worth teaching (the primary goal of the research in this case). By this method, Simpson-Vlach and Ellis were able to avoid the problems of wholly internal evaluation and also make external evaluation a realistic task[1]. Additionally, the characteristics by which formulae were judged in this experiment have the advantage of not pointing to any specific subtype of chunk, but rather indicating broader properties of lexical chunks, in particular, pragmatic and psychological properties. As lexical chunks have been most succinctly described not in grammatical terms but in pragmatic and psychological terms, framing the question of what a chunk is in these terms seems more likely to produce reliable judgments.

Following Simpson-Vlach and Ellis, I also used human judgments on a subset of plausible potential chunks to determine what should count as a chunk or not, and I preceded the rating task with a brief description of lexical chunks designed to give as accurate a definition as possible while allowing for a broad range of chunk subtypes to be possibly accepted (see Appendix A for the description text). However, unlike Simpson-Vlach and Ellis and almost every other experiment involving human evaluations, I did not give human raters chunks found by the measures themselves to rate; instead, I manually extracted all plausible chunks from the test data - the New York Times article - and had raters judge a subset of these plausible chunks for their acceptability. Scores were then derived for the remaining plausible chunks, and these scores were then used to create a gold standard of chunks, which could be used in combination with the token output lists from the running of different statistical measures to determine precision, recall, and f-measure scores. Details of the compilation of this gold standard list are given below.

5.2.2 Chunk List Compilation

Extracting Plausible Chunks

The first step in my evaluation process was to manually extract all plausible chunks from the New York Times article. Following Dias et al., I defined plausible chunks grammatically, but I did not use the notion of meaningful unit to come up with plausible chunks - this was to be part of the job of the human raters. I also attempted to allow for a broad range of chunk types by including many more grammatical categories than have typically been used in such tasks. Overall, I came up with 25 grammatical categories likely to produce chunks (described below) and extracted all word combinations that fell into these categories.

---

[1] 108 of the formulae found were rated, but it should be noted that the total number of formulae found was around 1500, which is very small compared to the number of chunks found by many other methods.

The first 10 categories used were taken from Nesselhauf (2005), who listed them as grammatical categories of collocations. These categories appear in Table 5.1. Because lexical chunks are not limited to just collocations, I enriched this list with an additional 15 categories of grammatical units that were likely to produce chunks. The full 25 categories are listed in the first column of Table 5.2. It should be noted that all of these categories could apply to contiguous and non-contiguous strings of words, but order was a discriminating factor. Thus a chunk could be counted in the adjective-noun category if it was of the form *adjective noun* or *adjective X Noun* but not of the form *noun adjective*.

Grammatical Category
Adjective + Noun
Noun (subject) + Verb
Noun + Noun
Adverb + Adjective
Verb + Adverb
Verb + Noun (object)
Verb + Preposition + Noun (object)
Noun + Preposition
Preposition + Noun
Adjective + Preposition

Table 5.1: *Grammatical types of collocation (from Nesselhauf 2005)*

After the categories had been selected, the New York Times article was parsed using the Stanford PCFG Parser (Klein and Manning 2003), and the parse was then reviewed and corrected where necessary. From the parsed version, all groups of words that fell into the 25 categories were then extracted manually. Groups of words were counted as falling into a category if they contained the correct parts of speech in the correct order for a category, and if those parts of speech were in a direct grammatical relation with each other. Thus, for the 'Verb + Noun' category, the noun had to be the object of the verb. Up to 2 gaps were allowed for each plausible chunk, so if a noun appeared as the direct object of a verb with two intervening words, the plausible chunk was extracted, but if there were three or more intervening words, the plausible chunk was not extracted.

Using this method, a total of 1,026 plausible chunks were extracted from the article. In parallel with this, a group of already-standard chunks was extracted manually using a selection of dictionaries containing current English phrases, collocations, and idioms. These resources are described below.

- **The Free Dictionary - Idioms and phrases:** A listing of English idioms compiled from the Cambridge International Dictionary of Idioms and the Cambridge Dictionary of American Idioms (Farlex 2011).
- **Oxford Collocations dictionary for students of English:** A searchable database of English collocations (*Oxford Collocations dictionary for students of English* 2011).

- **Academic Formulas List:** A listing of the top-scoring formulas found in academic spoken and written speech (Simpson-Vlach and Ellis 2010).
- **A Dictionary of American Idioms:** A dictionary of idiomatic words, expressions, regionalisms, and informal English expressions (Makkai et al. 2004).

Altogether, a total of 229 chunks were extracted from the combined dictionaries. Of the 1,026 plausible chunks extracted based on grammatical form, 188 (18.32%) overlapped with the chunks in the dictionary list (DL). The total number of chunks per category and the number of plausible chunks that overlapped with the DL chunks per category is shown in Table 5.2. After the 188 chunks found in the dictionaries were removed, 838 plausible chunks remained to be evaluated by humans. Of these, an additional 138 were removed due to their similarities to other plausible chunks, leaving a total of 700 plausible chunks to be evaluated.

Grammatical Category	**Chunks**	**% Total Chunks**	**Chunks in DL**	**% in DL**
Preposition + Noun	115	11.21%	10	8.70%
Adjective + Noun	108	10.53%	42	38.89%
Determiner + Noun	107	10.43%	0	0.00%
Noun + Preposition	79	7.70%	34	43.04%
Verb + Noun (object)	67	6.53%	15	22.39%
Verb + Preposition	58	5.65%	23	39.66%
Noun + Preposition + Noun	53	5.17%	1	1.89%
Noun (subject) + Verb	51	4.97%	2	3.92%
Preposition + Determiner + Noun	45	4.39%	9	20.00%
Noun + Noun	41	4.00%	18	43.90%
Verb + Verb	40	3.90%	0	0.00%
Preposition + Noun + Preposition	35	3.41%	2	5.71%
Determiner + Noun + Preposition	34	3.31%	1	2.94%
Verb + Preposition + Noun (object)	30	2.92%	1	3.33%
Proper noun	25	2.44%	8	32.00%
'to' + Verb	25	2.44%	1	4.00%
Preposition + Determiner + Noun + Preposition	19	1.85%	4	21.05%
Adverb + Verb	17	1.66%	2	11.76%
Quantity	16	1.56%	0	0.00%
Determiner + Preposition	15	1.46%	0	0.00%
Adjective + Preposition	12	1.17%	6	50.00%
Verb + Adverb	12	1.17%	1	9.09%
Time phrase	10	0.97%	2	20.00%
Verb + Adjective	9	0.88%	7	77.78%
Adverb + Adjective	4	0.39%	0	0.00%
Total	1026	100.00%	188	18.32%

Table 5.2: *Plausible chunks found per category and overlap with Dictionary List (DL) chunks*

The 700 plausible chunks to be evaluated were presented to two raters through a series of online surveys at http://www.surveymonkey.com. These raters were both native English speakers with some background in Linguistics, though they did not have specific knowledge of what lexical chunks were prior to the study. The raters first read an introductory text (see Appendix A) which explained the basic concept of lexical chunks, and they were then asked to rate the chunks on a 7-point scale, reproduced below.

1. Definitely not a chunk
2. Probably not a chunk
3. Maybe not a chunk
4. Could go either way
5. Maybe a chunk
6. Probably a chunk
7. Definitely a chunk

One of the raters gave much higher ratings than the other rater, so the ratings were normalized using the decoupling method, recommended in (Jin and Si 2004). Normalized ratings were then averaged for all plausible chunks, and plausible chunks scoring above a cutoff of 50% were accepted as chunks favored by the raters. This resulted in the addition of 330 chunks to the dictionary list. Scores for the 138 plausible chunks that had been removed from the evaluation list were then extrapolated from the scores for similar plausible chunks that had been rated, and an additional 90 received scores higher than 50%. The total list of acceptable chunks thus consisted of 649 chunks. Of these, three were removed because they involved words spelled differently in British and American English (as the evaluation text was American English while the training corpus was British English), for a total of 646 'gold standard' chunks which could be used as a basis for evaluation of the different statistical methods. 47 of these contained two gaps, so the remaining 599 chunks were used for the evaluation of methods that found chunks containing a maximum of one gap. These gold standard chunks are listed in full in Appendix C.

Chapter 6

Results

6.1 Five Methods Compared

6.1.1 Types and Tokens

Prior to comparison with the chunks in the gold standard list, the five statistical methods were compared to each other on three measures: total number of chunks extracted in training, number of tokens found in the evaluation text, and number of types found in the evaluation text. These data appear in Table 6.1[1].

Statistical Method	Total Chunks	Tokens	Types	Type/Token (%)	Type/Total (%)
SCP	1165598	289	269	90.66	0.02
Mutual Information	659902	115	102	88.70	0.02
Log Likelihood	222379	174	148	85.06	0.07
Mutual Expectation	220536	791	633	80.03	0.29
Raw Frequency	204996	1355	902	66.57	0.44

Table 6.1: *Chunks, types, and tokens found by different statistical measures*

Of the methods, Raw Frequency is clearly the most ‘efficient’, in that the largest proportion of the chunks found in training appear in the evaluation text. By this measure, SCP and Mutual Information are the least efficient, but they also extracted many more chunks in training than the other measures. SCP and Mutual Information also had the highest type/token ratios, meaning more of the chunks they found in the evaluation text appeared only once in that text. Unsurprisingly, the type/token ratio for Raw Frequency was much lower than for the other methods. This is natural, since the method extracts the most frequent chunks.

6.1.2 Precision and Recall

Though the information in Table 6.1 gives us some idea of how the different methods are working, it is unclear whether the chunks being found are any good, in the sense that they fit the criteria for lexical chunks. A method that

---

[1] Unless otherwise noted, the cutoffs used for each statistical measure in evaluation are those in Table 4.2.

extracts lots of word groups that are not useful chunks (i.e., has high recall but low precision) may be less desirable than a method that extracts fewer groups, but of those groups, most are chunks (i.e., has lower recall but higher precision). To that end, the chunk types found by the statistical methods in the evaluation text were compared against the gold standard list determined by a combination of dictionary compilation and human evaluation, as discussed in the previous chapter. Methods were evaluated on precision (% of correct types found out of the total number of types found), recall (% of correct types found out of the total number of correct types that could have been found), and f-measure ($\alpha = 1$, thus, the harmonic mean of precision and recall). These data are displayed in Table 6.2.

Raw Frequency and Mutual Expectation, the measures which found the greatest number of overall types in the evaluation text, also had the highest f-measure scores, due mainly to their greater recall. In fact, their precision scores were the lowest of the five measures. In precision, Log Likelihood was the clear leader, but it had relatively low recall. SCP and Mutual Information had the lowest f-measure scores, with Mutual Information coming in last, due to its low recall of only 6%.

Statistical Method	**Types**	**# correct**	**Precision (%)**	**Recall (%)**	**F-measure**
Raw Frequency	902	219	24.28	36.56	0.29
Mutual Expectation	633	152	24.01	25.38	0.25
Log Likelihood	148	77	52.03	12.85	0.21
SCP	262	69	26.34	11.52	0.16
Mutual Information	102	36	35.29	6.01	0.10

Table 6.2: *Precision, recall, and f-measure for different statistical measures*

It is typical to see an inverse relationship between precision and recall scores: as recall increases, precision tends to decrease. However, this pattern was not evidenced for all of the methods tested. The relationship between precision and recall for the five measures (each run with three different cutoffs) is shown in Figure 6.1. It would be expected that lower cutoffs would result in greater recall and lower precision, and this pattern does indeed occur for Mutual Information, SCP, and to a lesser extent Log Likelihood. The measures of Raw Frequency and Mutual Expectation however, do not show this pattern. Though their recall increases as the cutoff is lowered, their precision scores remain more or less constant.

6.1.3 Grammatical Types

To aid in the performance analysis of the different measures, the types found by each measure were categorized by the grammatical types of the words that occurred in them. The grammatical type categories used are given in Table 6.3. Each chunk type found by a method was given a code made up of the

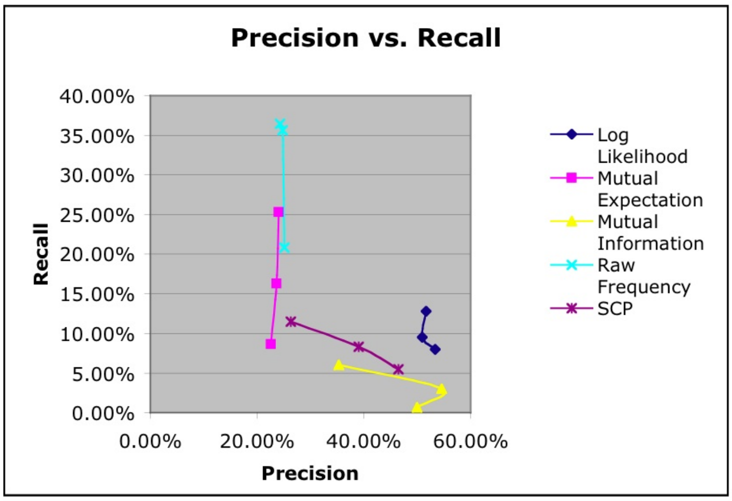

Figure 6.1: *Relationship between precision and recall for 3 different cutoffs*

Code	Type	Example
A	Adverb	*finally*
C	Conjunction	*and*
D	Determiner	*the*
EX	Existential *there*	*there* used existentially (as in *there is*)
J	Adjective	*uncommon*
N	Noun	*company*
P	Preposition	*of*
PN	Proper Noun	*muammar*
POSS	Possessive Pronoun	*his*
PRO	Pronoun	*it*
Q	Quantity	*seven*
REL	Relative Pronoun	*who*
S	Possessive *'s*	*'s*
T	Time word	*2008*
TO	Infinitival *to*	*to* preceding infinitive (e.g., in *to be*)
V	Verb	*provide*
X	Gap	Indicates gap in chunk

Table 6.3: *Codes for Grammatical Type Categories*

codes for all the grammatical types occurring in it. Thus, a chunk like *could not be* was given the code V + A + V, and a chunk with a gap, like *few X ago*, was given the code J + X + A. Frequently occurring types for the different methods were then examined. Table 6.4 shows the five most frequent grammatical types found by each method, along with the percentage of overall types found by the method that the grammatical type accounted for.

Several trends are evident from these data. First of all, the methods find a wide variety of grammatical types, with even the most frequent types accounting for only around 5% of the overall types found. A notable exception is Log Likelihood, where the types are dominated by determiner-noun combi-

Log Likelihood		**Mutual Information**		**Raw Frequency**		**SCP**		**Mutual Expectation**	
type	**%**	**type**	**%**	**type**	**%**	**type**	**%**	**type**	**%**
D + N	16.89	J + N	8.82	N + P	5.65	N + P	4.46	P + D	4.90
TO + V	8.78	N + N	7.84	D + N	5.21	J + N	3.35	V + P	4.58
J + N	6.76	V + P	2.94	V + P	3.44	N + N	3.35	N + P	3.79
P + X + N	6.08	N + P + N	2.94	P + D	2.66	V + P	2.97	P + X + P	2.21
N + N	6.08	V + V + P	2.94	D + X + N	2.55	P + D + N	2.60	D + N	2.05
				N + X+ D	2.55			D + X + P	2.05

Table 6.4: *Top 5 grammatical types found by each method*

nations, followed by infinitival verb clauses and adjective-noun combinations, these three together accounting for nearly a third of the total types found. In fact, the top 10 types found by Log Likelihood account for over 60% of the total grammatical types found, whereas the top 10 types found by all the other methods account for roughly 25-35% of the total grammatical types found.

The most commonly found types across methods are two-word types ending in either nouns or prepositions, namely determiner-noun, adjective-noun, noun-noun, verb-preposition, and noun-preposition combinations. Types can also be categorized according to whether they contain only content-word types (nouns, adjectives, verbs), only function-word types (prepositions, determiners, conjunctions, etc.), or some combination of these. In the top 5 types found by the different methods, only those for Raw Frequency and Mutual Expectation contain types consisting of just function words (e.g., preposition-determiner chunks, or chunks with two prepositions separated by a gap). By contrast, Mutual Information finds a particularly large proportion of chunks consisting of only content words.

Other interesting comparison points between methods include length (number of words) and gaps. Mutual Information and SCP are the only methods for which chunks of more than two words are among the top 5 types found, and none of the methods find many chunks of more than three words. Mutual Information and SCP are also the only methods which do not find many chunks with gaps in them. When chunks with gaps in them are found, the most common types include either a determiner or a preposition, and the other word is often a noun.

6.2 Adding Gaps

The results discussed above are all for chunks involving no more than one gap. Methods were also run with the maximum number of gaps set to two, but no major differences in performance - precision or recall - were found, as shown in Figures 6.2 and 6.3[2].

---

[2] For comparison between chunks, methods were run with a word frequency cutoff of 5, as opposed to 1, because without the higher cutoff, the number of 2-gap chunks extracted in

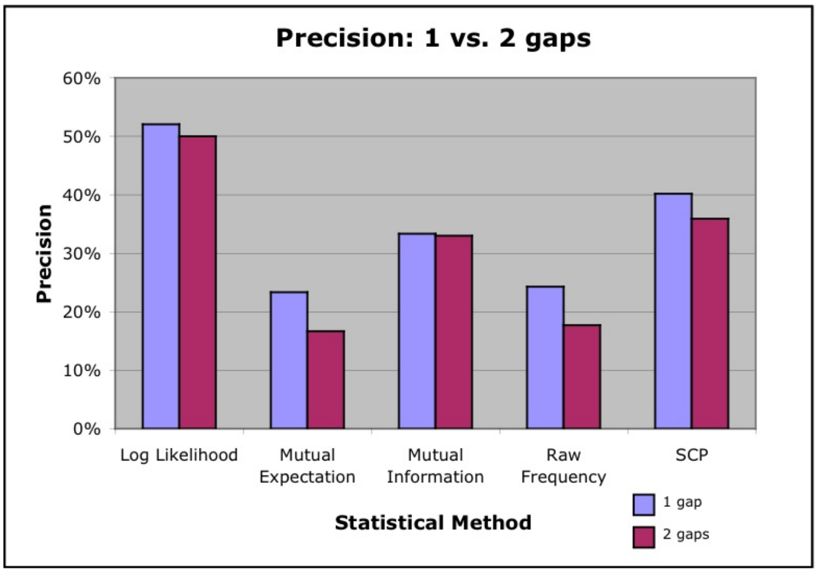

Figure 6.2: *Precision for 1 vs. 2-gap chunks*

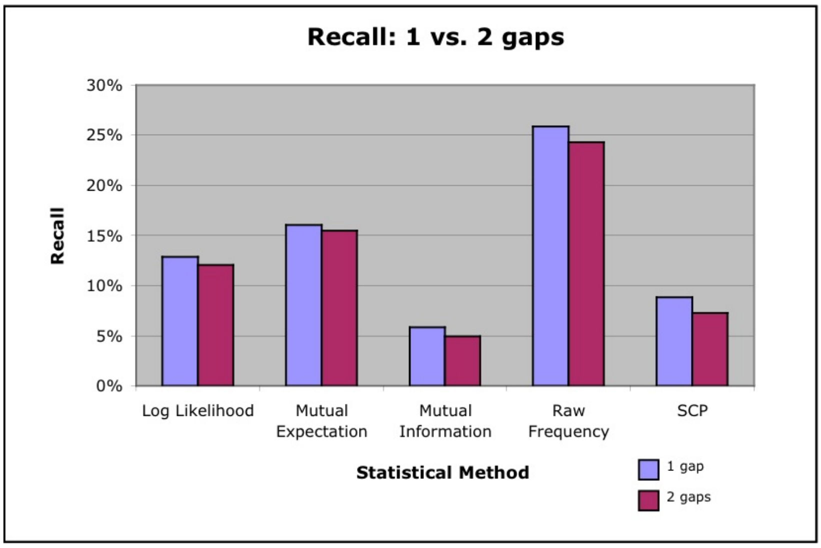

Figure 6.3: *Recall for 1 vs. 2-gap chunks*

Though differences in performance were minor, the 1-gap versions had slightly better precision and recall than the 2-gap versions. This held for all five methods tested with no exceptions. Further investigation into the numbers of types and tokens found in the evaluation text by the 2-gap versions, as shown in Table 6.5, showed significantly larger numbers only for Mutual Expectation and Raw Frequency. Other measures found only slightly more types and tokens when chunks could have a maximum of 2 gaps, and in some cases, they actually found fewer types and tokens. This could occur because of the

---

training was extremely large.

subchunk removal: 2-gap versions that found chunks with higher scores than subchunks found by both 1-gap and 2-gap versions would remove those subchunks, while 1-gap versions might not find the longer chunks and so would not remove the subchunks.

As the 2-gap versions did not show significant improvements over the 1-gap versions, further analyses were performed only on data from the 1-gap versions.

Statistical Method	Gaps	Total Chunks	Tokens	Types	Type/Token (%)	Type/Total (%)
Mutual Information	1	645077	115	105	91.3	0.02
SCP	1	299677	148	132	89.19	0.04
Log Likelihood	1	217336	174	148	85.06	0.07
Mutual Expectation	1	30114	556	412	74.10	1.37
Raw Frequency	1	26950	809	638	78.86	2.37
Mutual Information	2	961276	111	97	87.39	0.01
SCP	2	455584	148	131	88.51	0.03
Log Likelihood	2	277350	183	156	85.25	0.06
Mutual Expectation	2	40436	781	601	76.95	1.49
Raw Frequency	2	37687	1131	886	78.34	2.35

Table 6.5: *Chunks, types, and tokens found by different statistical measures, 1 vs. 2 gaps*

6.3 Splitting the Corpus

6.3.1 Types and Tokens

One of the main research questions of this work was what effect, if any, splitting the corpus would have on the number and type of lexical chunks found. The type/token results of splitting the corpus, in the manner detailed in Chapter 4, are shown in Table 6.6, along with the results previously presented for the unsplit corpus (from Table 6.1).

A few major trends are evident from these data. First of all, when the corpus is split, the total number of chunks found in training decreases for all measures except Log Likelihood. The reduction in total number of chunks found is particularly drastic for SCP and Raw Frequency. By contrast, the number of tokens and types found in the evaluation text increases for all measures except Raw Frequency. These increases are quite large, except in the case of Mutual Expectation[3].

---

[3] It should be noted that different cutoffs were used for Mutual Expectation and SCP when the corpus was split. This was because these two measures are highly dependent on overall corpus size and the cutoffs used for the full corpus run were simply too small to use in the split corpus run. The cutoffs used for Mutual Expectation and SCP in the split corpus runs were 1×10^{-5} and 0.01, respectively.

Statistical Method	Corpus	Total Chunks	Tokens	Types	Type/Token (%)	Type/Total (%)
SCP	split	292886	761	596	78.32	0.20
Mutual Information	split	638777	1048	878	83.78	0.14
Log Likelihood	split	565006	1001	812	81.12	0.14
Mutual Expectation	split	203261	828	654	78.99	0.32
Raw Frequency	split	21370	573	418	72.95	1.96
SCP	full	1165598	289	269	90.66	0.02
Mutual Information	full	659902	115	102	88.70	0.02
Log Likelihood	full	222379	174	148	85.06	0.07
Mutual Expectation	full	220536	791	633	80.03	0.29
Raw Frequency	full	204996	1355	902	66.57	0.44

Table 6.6: *Chunks, types, and tokens found by different statistical measures, full vs. split corpus*

6.3.2 Precision and Recall

As with the full corpus data, the types found in the split corpus method were compared against the gold standard chunk types, and measurements of precision, recall, and f-measure were obtained. These data are summarized in Table 6.7, which also gives the scores from the full corpus runs. All f-measure scores improved, except for those of the Raw Frequency method. The differences in f-measure scores for the full vs. split corpus runs are shown in Figure 6.4.

Statistical Method	Corpus	Types	# correct	Precision (%)	Recall (%)	F-measure
SCP	split	596	223	37.42	37.23	0.37
Log Likelihood	split	812	263	32.39	43.91	0.37
Mutual Information	split	878	263	29.95	43.91	0.36
Mutual Expectation	split	654	185	28.29	30.88	0.30
Raw Frequency	split	418	107	25.60	17.86	0.21
SCP	full	262	69	26.34	11.52	0.16
Log Likelihood	full	148	77	52.03	12.85	0.21
Mutual Information	full	102	36	35.29	6.01	0.10
Mutual Expectation	full	633	152	24.01	25.38	0.25
Raw Frequency	full	902	219	24.28	36.56	0.29

Table 6.7: *Precision, recall, and f-measure for different statistical measures using split corpus*

In general, precision scores showed only small improvements or even decreased when the corpus was split; the improvements in f-measure were thus driven mainly by large improvements in recall. Precision scores decreased for both Log Likelihood and Mutual Information, and the only measure for which precision substantially improved in the split corpus run was SCP. Recall, on the other hand, improved substantially for SCP, Log Likelihood and Mutual Information, and it improved moderately for Mutual Expectation. The only measure for which recall decreased was Raw Frequency, but this decrease was large, from 36.56 to 17.86% - over a 50% decrease.

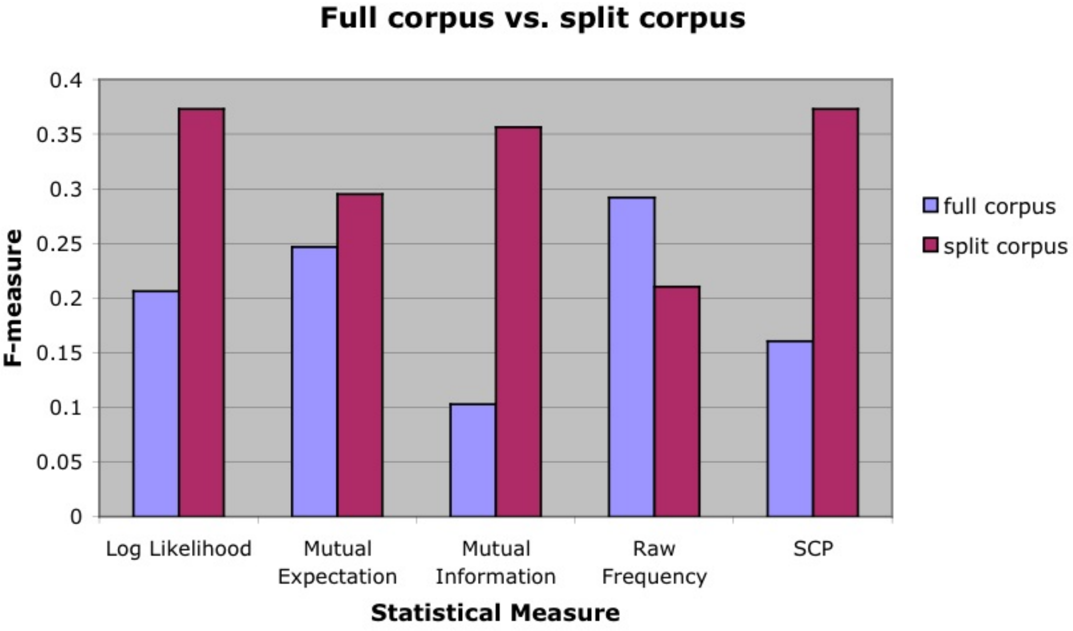

Figure 6.4: *F-measure scores for full vs. split corpus runs*

The % change that occurred in precision, recall, and f-measure values for all five measures when the corpus was split is shown in Figure 6.5. Mutual Information showed the biggest gains in f-measure and recall, followed by SCP

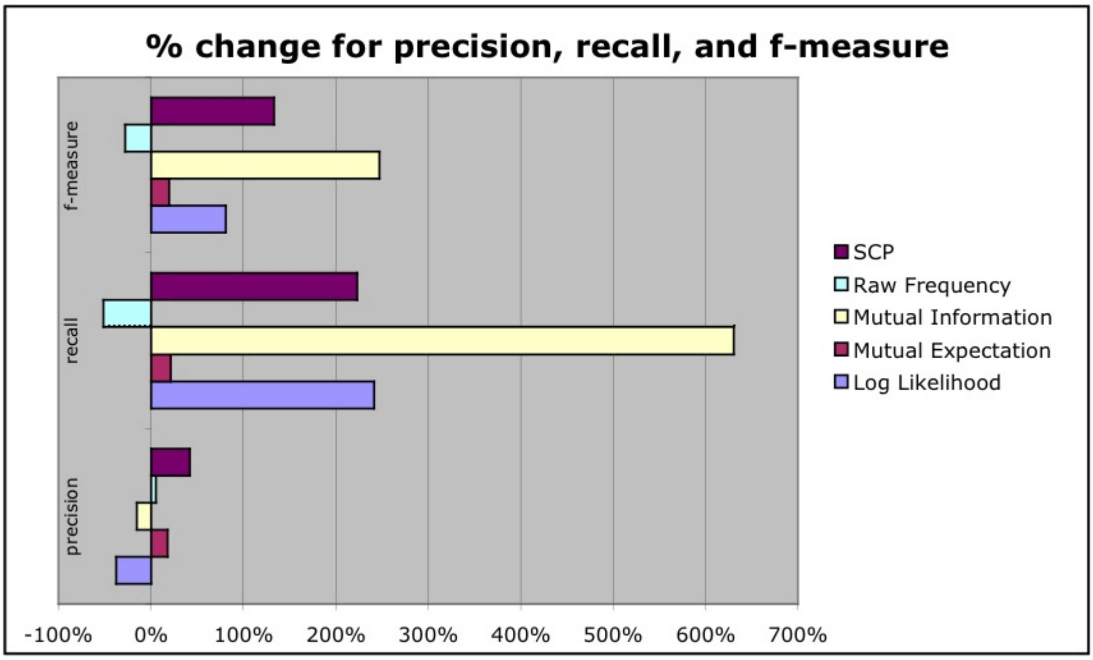

Figure 6.5: *% change in precision, recall, and f-measure values in split corpus runs*

and Log Likelihood. Raw Frequency was the only measure which showed worse performance on these measures in the split corpus run.

6.3.3 Grammatical Types

As with the full corpus chunks, an analysis of the grammatical types of words making up the split corpus chunks was performed. The five most frequent grammatical types for each method are shown in Table 6.8, along with the five most frequent grammatical types found by the full-corpus runs, for comparison.

In the case of Log Likelihood and Mutual Information, the most frequent types found in the split corpus runs differed quite a bit from the frequent types found in the full corpus runs. For both these methods, only one grammatical type appeared in the top five for both runs. On the other hand, Mutual Expectation and Raw Frequency were fairly consistent, with four types appearing in the five most frequent types for the split and full corpus runs.

Log Likelihood		**Mutual Information**		**Raw Frequency**		**SCP**		**Mutual Expectation**	
type	**%**	**type**	**%**	**type**	**%**	**type**	**%**	**type**	**%**
N + P	6.03	N + P	5.92	P + D	7.18	N + P	6.54	D + N	4.89
D + N	5.17	D + N	5.13	D + N	6.46	V + P	5.20	P + D	4.59
V + P	4.31	V + P	4.10	V + P	3.11	D + N	5.03	N + P	4.28
P + D	3.57	P + D	2.85	TO + V	3.11	P + D	4.87	V + P	3.67
V + D	2.22	P + X + N	2.73	N + P	2.87	TO + V	2.68	P + D + N	1.99
								P + X + N	1.99
D + N	16.89	J + N	8.82	N + P	5.65	N + P	4.46	P + D	4.90
TO + V	8.78	N + N	7.84	D + N	5.21	J + N	3.35	V + P	4.58
J + N	6.76	V + P	2.94	V + P	3.44	N + N	3.35	N + P	3.79
P + X + N	6.08	N + P + N	2.94	P + D	2.66	V + P	2.97	P + X + P	2.21
N + N	6.08	V + V + P	2.94	D + X + N	2.55	P + D + N	2.60	D + N	2.05
				N + X+ D	2.55			D + X + P	2.05

Table 6.8: *Top 5 grammatical types found by each method: split in top row, full in bottom row*

Unlike with the full corpus runs, split corpus runs gave more consistent types across the five methods. Four types appeared in the top five most frequent types of every single statistical method in the split corpus run, while no type appeared in the top five for all five methods in the full corpus run. The four types that were found frequently by every method in the split corpus run were: determiner-noun, noun-preposition, verb-preposition, and preposition-determiner combinations. The other types that were found by at least two methods were infinitival verb clauses and preposition-noun combinations with a gap. No other combinations with a gap occurred in the top five chunk types for the split corpus run, and the only three-word type in the top five was a preposition-determiner-noun chunk.

In another difference from the full corpus types, the most frequent split corpus types did not contain any types consisting of only content words. All the frequent types consisted of a combination of content and function words, except for the preposition-determiner combination. The percentage of overall types accounted for by the top five types did not change much from the full to the split corpus runs, generally staying between 15-20%. The one exception was Log Likelihood, for which the top five full corpus types accounted for a much greater percentage (∼45%) than the top five split corpus types. These data are shown in Figure 6.6.

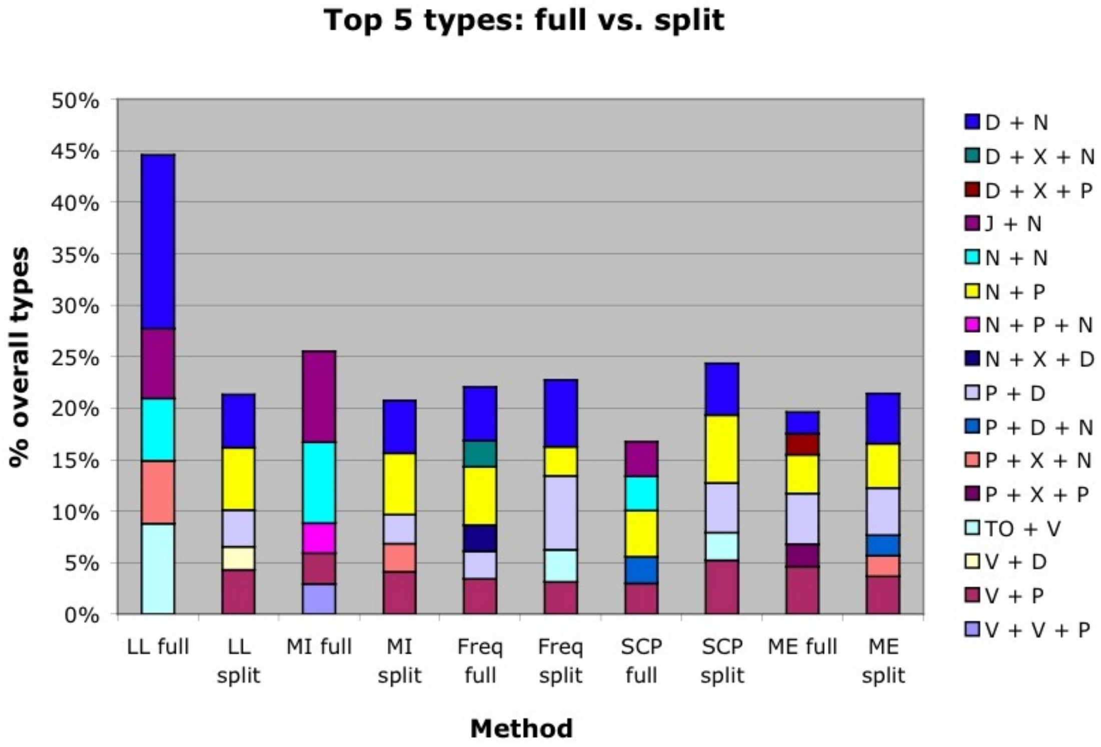

Figure 6.6: *Top five types found by full and split corpus runs*

Chapter 7

Discussion

7.1 Which Method is Best?

Data from the first set, in which all methods were trained on an unsplit corpus, indicate that Raw Frequency is the best method for finding lexical chunks. It outperforms all the other methods on recall, but it has the second-lowest precision score, so its advantage is mainly in the fact that it finds lots of chunks. In some cases it may be preferable to use a method which, though it may not find as many chunks, finds 'better' chunks, i.e., has high precision. If this is the case, Log Likelihood is the clear winner with a precision score roughly twice as high as precision scores for other measures. However, with a recall of close to 13%, this method fails to find many chunks.

One of my original hypotheses was that many lexical chunks might only occur locally: they would occur with high frequency in a few texts and not at all or rarely in the majority of other texts. Splitting the corpus would thus enable these chunks to be found more easily, as they would be likely to stand out statistically in the texts where they occurred. Indeed, splitting the corpus led to vast improvements in recall for Log Likelihood, SCP, and particularly for Mutual Information. Splitting the corpus actually resulted in fewer overall chunks being extracted in training for all methods except Log Likelihood, for which more than twice as many chunks were extracted when the corpus was split. Despite the smaller chunk lexicons, all the methods except Raw Frequency actually found more types in the evaluation text with the split corpus chunks, and type/total ratios increased across the board.

Though type/total ratios increased in the split corpus runs, this does not necessarily indicate that a higher percentage of good chunks were found. However, examination of the data showed that this was indeed the case for all the measures except SCP. The percentage of the overall chunks which were also correctly extracted from the evaluation text by different methods is given in Figure 7.1. As can be seen from these data, splitting the corpus generally resulted in more useful chunk lexicons, in the sense that although they contained fewer chunks overall, they were more likely to contain good chunks. This supports the hypothesis that a split corpus method is better for finding lexical chunks, which may appear with high frequencies in certain texts and not at all in others.

The best-performing methods overall were SCP and Log Likelihood us-

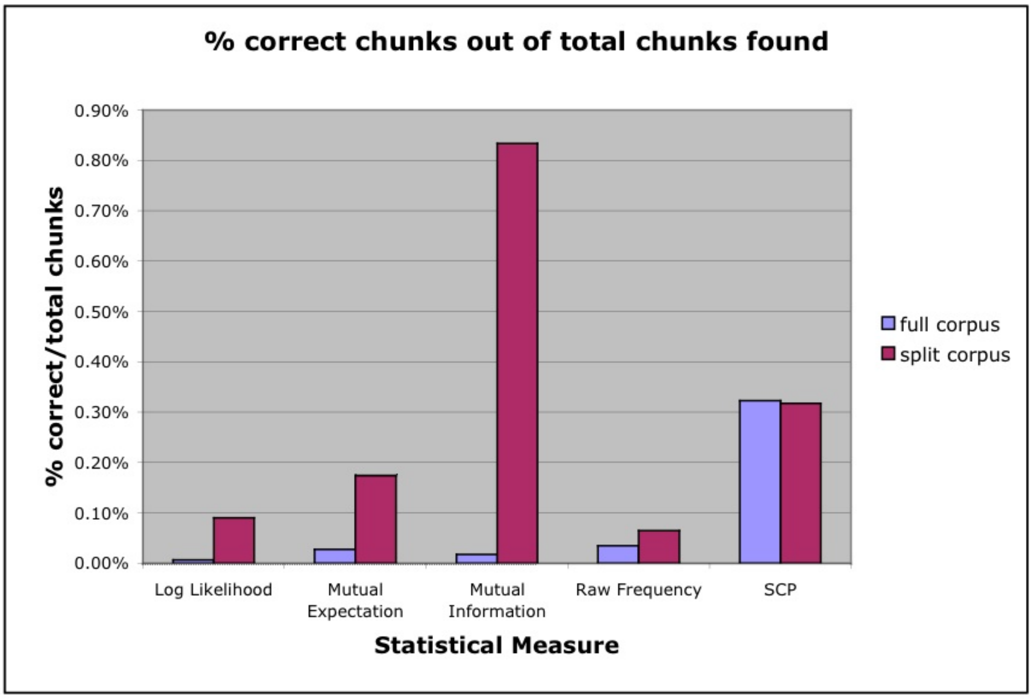

Figure 7.1: *% correct types over total types found, full vs. split corpus*

ing a split corpus, though Mutual Information using a split corpus performed nearly as well. All three of these measures had higher precision and higher recall than the full corpus Raw Frequency method, which had the highest f-measure score of the full corpus methods. The highest precision score in the split corpus run was achieved by SCP, and this precision score was better than all full-corpus scores, except that of Log Likelihood. For the researcher interested in good overall performance, it would seem that SCP, Log Likelihood or Mutual Information on a split corpus are the best options, while a researcher interested only in precision would do better to use Log Likelihood on a full corpus. If high recall is the goal, the split corpus methods are assuredly better, except in the case of Raw Frequency.

7.2 Investigating Differences in Chunks Found: Full vs. Split

7.2.1 Differences in Chunk Word Frequencies

As noted above, a variety of differences were observed in the performance of different methods when they were trained on a full vs. a split corpus. Differences in the types of chunks, as categorized by the part of speech (POS) type of their constituent words, were also observed both across measures and across the full/split corpus divide. Performance and chunk type differences are summarized for each measure in Table 7.1.

One of the main differences in chunk types found using a full corpus as opposed to a split one is that purely content-word chunks (adjective-noun, noun-noun, etc.) were common in the full corpus chunks for Log Likelihood, Mutual Information, and SCP, but they were much less common in the split corpus chunks. Because content words generally have a much lower frequency than function words (see Table 7.2), it seems that full corpus training biases the above three methods to find chunks that do not contain high-frequency words. To see if this was the case, an analysis of the frequency of the chunk words was carried out. The analysis was based on a random sample of 20 chunks found by each measure. Each word in each chunk was classified by its frequency and the measures were then examined to see what percentage of the words fell below different frequency cutoffs. The results of this analysis are shown in Figure 7.2.

Statistical Method	Corpus	# chunks	Precision	Recall	F-measure	Dominant Chunk Types
Log Likelihood	full	avg	very high	low	avg	D+N, TO+V, J+N
	split	avg	high	high	high	N+P, D+N, V+P
Mutual Expectation	full	avg	avg	avg	avg	P+D, V+P, N+P
	split	avg	avg	avg	avg	D+N, P+D, N+P
Mutual Information	full	high	high	very low	low	J+N, N+N
	split	high	avg	high	high	N+P, D+N, V+P
Raw Frequency	full	avg	avg	high	avg	N+P, D+N, V+P
	split	very low	avg	low	avg	P+D, D+N
SCP	full	very high	avg	low	low	N+P, J+N, N+N
	split	avg	high	high	high	N+P, V+P, D+N

Table 7.1: *Summary of performance data for full and split runs of different statistical measures*

These data bear out the hypothesis that Log Likelihood and Mutual Information find a greater percentage of chunks containing low frequency words in the full corpus runs, but SCP does not seem to follow the pattern. The differences between the full corpus versions of Log Likelihood and Mutual Information and all the other methods run are striking. Over 40% of the words in the sample of chunks found by these two measures occurred fewer than 1000 times in the 5.87 million word corpus, and nearly three quarters of the words for these measures occurred fewer than 10,000 times. The correspondence between word frequency and word type (content or function) can be seen in Table 7.2, where a sample of 20 different chunk words with their frequencies is given.

In general, words occurring more than 10,000 times in the corpus were function words, modal verbs, or a few very common verbs, such as *said* and *came*. Low-frequency words were nearly all nouns, verbs, adjectives, and adverbs.

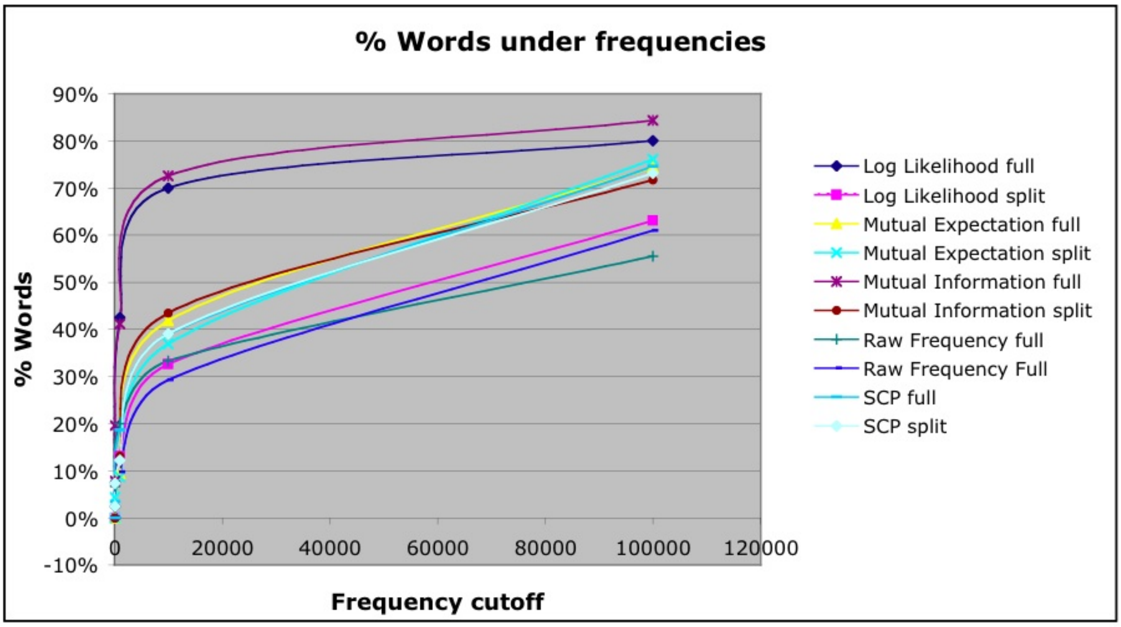

Figure 7.2: *% chunk words below different frequencies*

Word	POS type	Frequency	Word	POS type	Frequency
an	D	20218	*indicated*	V	247
business	N	3198	*it*	PRO	74787
company	N	2943	*many*	J	3621
finally	A	700	*not*	A	28981
for	P	49831	*of*	P	167694
foreign	J	495	*pay*	V	1472
had	V	21900	*said*	V	10529
he	PRO	37337	*the*	D	346118
his	POSS	22176	*to*	P/TO	165851
in	P	106552	*was*	V	47312

Table 7.2: *20 chunk words with their corpus frequencies*

7.2.2 Interactions between Chunk Word Frequencies and Performance

The word-frequency analysis suggests one explanation for why the full corpus versions of Log Likelihood and Mutual Information find such different chunks from the split corpus versions, but it does not do much to elucidate the differences in performance. One might hypothesize that finding chunks containing high-frequency words leads to greater recall and thus better performance. To a certain extent this is likely true, but an analysis of the association between chunk word frequencies and recall shows that this only holds for the full corpus versions of methods. The associations between word frequencies and recall are given for the full and split corpus versions in Figure 7.3.

While methods that found more high-frequency words had greater re-

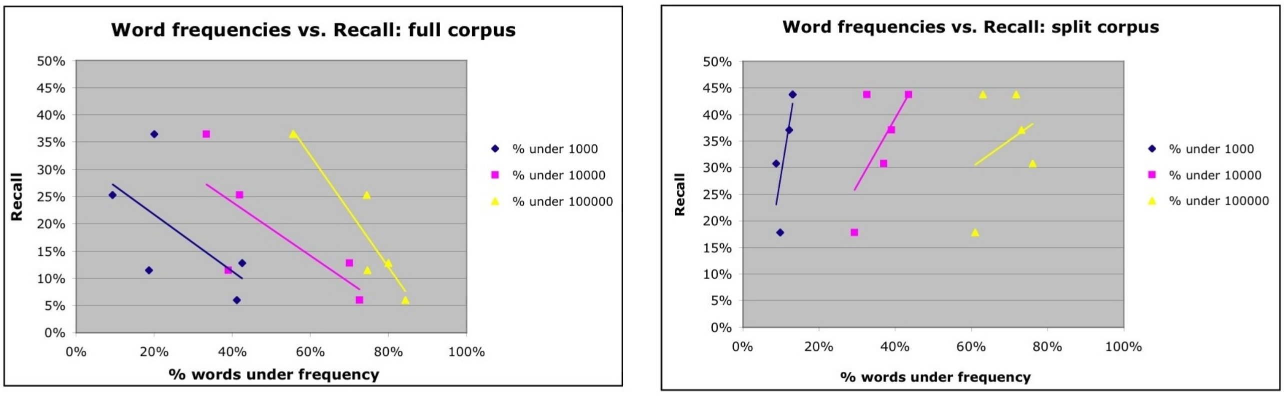

Figure 7.3: *Association between % chunk words below different frequencies and recall*

call in the full corpus versions, the reverse held true for the split corpus versions. Associations between word frequency and precision, shown in Figure 7.4, showed that measures that found more high-frequency words generally had worse precision than measures that found more low-frequency words. This held true for both full and split corpus versions, though the association was more prominent for the full corpus versions. The explanation that higher recall was a result of methods finding more chunks with high-frequency words fails to account for the split corpus trends. Further investigations into the different types of chunks found across measures and across the full vs. split corpus runs were thus necessary.

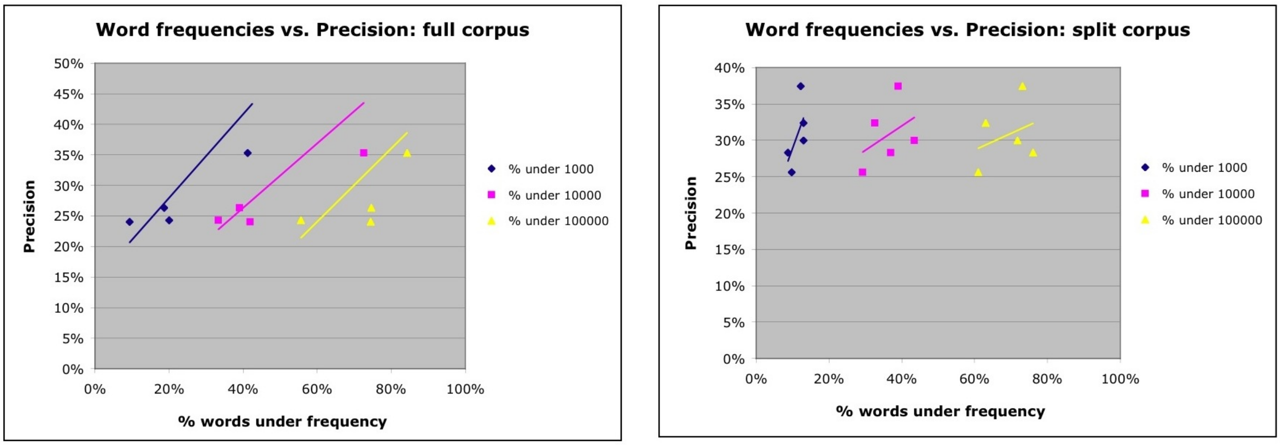

Figure 7.4: *Association between % chunk words below different frequencies and precision*

7.2.3 Interactions between Chunk Word Type and Performance

Because changes in the percentages of chunk word frequencies could not account for the differences in performance alone, the POS types of chunk words were examined more thoroughly and in comparison with the POS types of the gold standard 'good' chunks. In particular, the good chunks that were not found by any of the five methods were examined to see if any patterns could be found. The most frequent types for the gold standard chunks are displayed in Table 7.3, along with the number of chunks of each POS type that were not found by any method.

One of the most striking patterns in these data is that the chunks that were most difficult to find were almost always content-word chunks: adjective-noun, noun-noun, noun-verb, proper noun combinations, etc. These types of chunks were also among the most common types of chunks found by Log Likelihood, Mutual Information, and SCP when trained on a full corpus. The worse performance of these measures as compared to the split-corpus-trained versions is then surprising. An examination into type-by-type recall for all the measures, as shown in Figures 7.5 through 7.9, shows why this is the case.

Chunk type	**Total**	**# not found**	**% not found**	**Chunk type**	**Total**	**# not found**	**% not found**
N + P	65	16	24.62	P + D + N	15	2	13.33
J + N	59	47	79.66	V + X + N	15	13	86.67
D + N	43	8	18.60	P + X + N + P	14	11	78.57
V + P	38	2	5.26	PN + PN	14	11	78.57
N + N	28	17	60.71	V + V	13	1	7.69
D + X + N	24	8	33.33	P + D + N + P	11	5	45.45
D + N + P	23	10	43.48	J + P	10	2	20.00
TO + V	23	2	8.70	D + X + P	9	1	11.11
P + X + N	19	6	31.58	N + P + N	9	6	66.67
P + N	17	8	47.06	V + N	9	7	77.78
J + X + N	15	14	93.33				

Table 7.3: *Most frequent POS types of gold standard chunks*

It appears that the split corpus versions of Log Likelihood, Mutual Information, and SCP actually find nearly as many content-word-only chunks as the full corpus versions; they just find a lot more *other* types of chunks, which the full corpus versions do not find. In particular, the split corpus versions find a lot more determiner-noun combinations, infinitival verb clauses, and content-word+preposition combinations, which include phrasal verbs - an important category of lexical chunk in English. Raw Frequency recall is higher for nearly all of the top 10 grammatical chunk types for the full corpus version, the one exception being infinitival verb phrases, which are found exclusively by the split corpus version (possibly due to greater deletion of subchunks in the split corpus version). Mutual Expectation shows much less difference between full and split recall for different types; the main noticeable difference is that the split version finds chunks containing gaps while the full version does not.

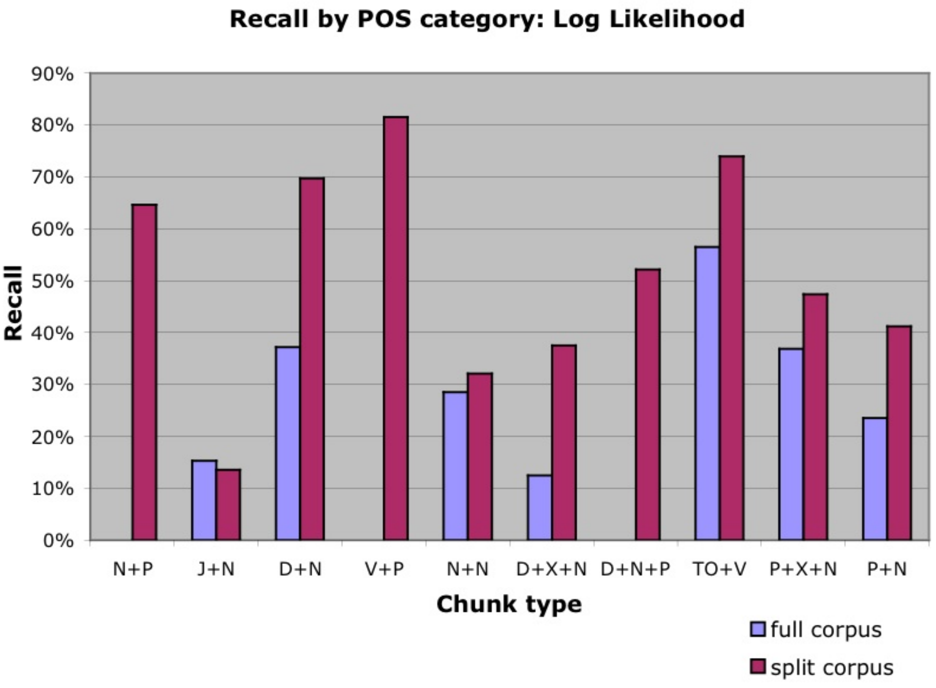

Figure 7.5: *Recall for 10 most common POS types: Log Likelihood*

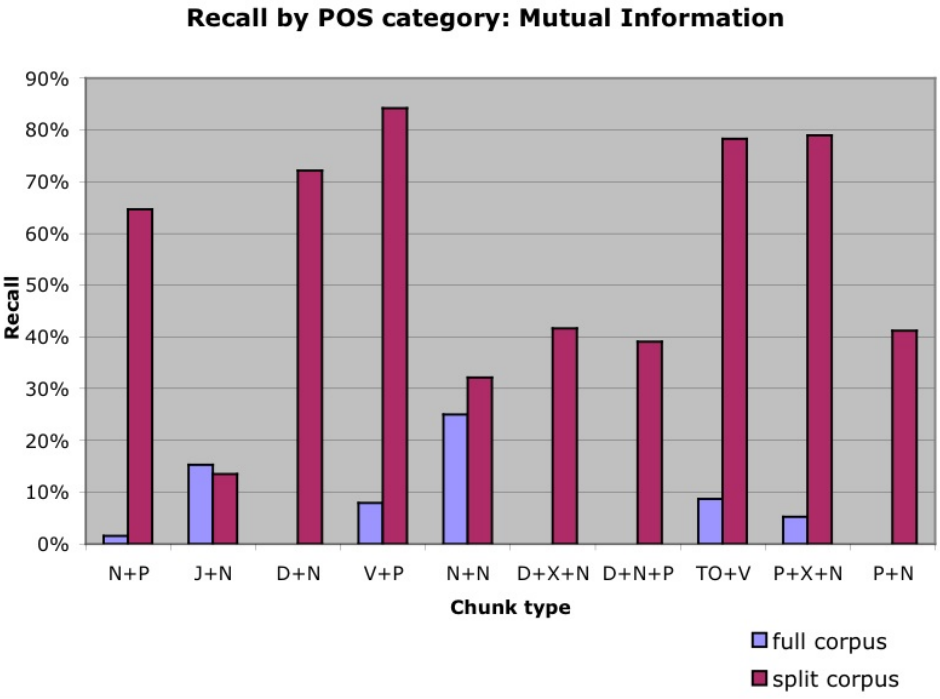

Figure 7.6: *Recall for 10 most common POS types: Mutual Information*

7.2.4 Interpreting the Differences in Chunk Types Found

Why should split corpus versions (except for Raw Frequency) be better at finding chunks containing high-frequency function words? The answer seems

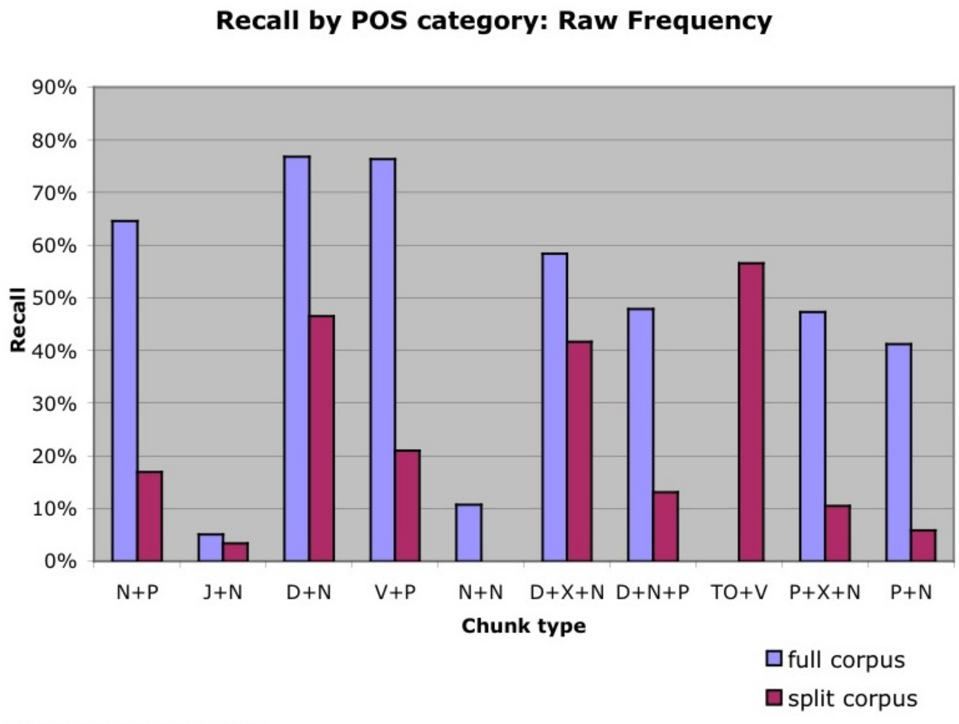

Figure 7.7: *Recall for 10 most common POS types: Raw Frequency*

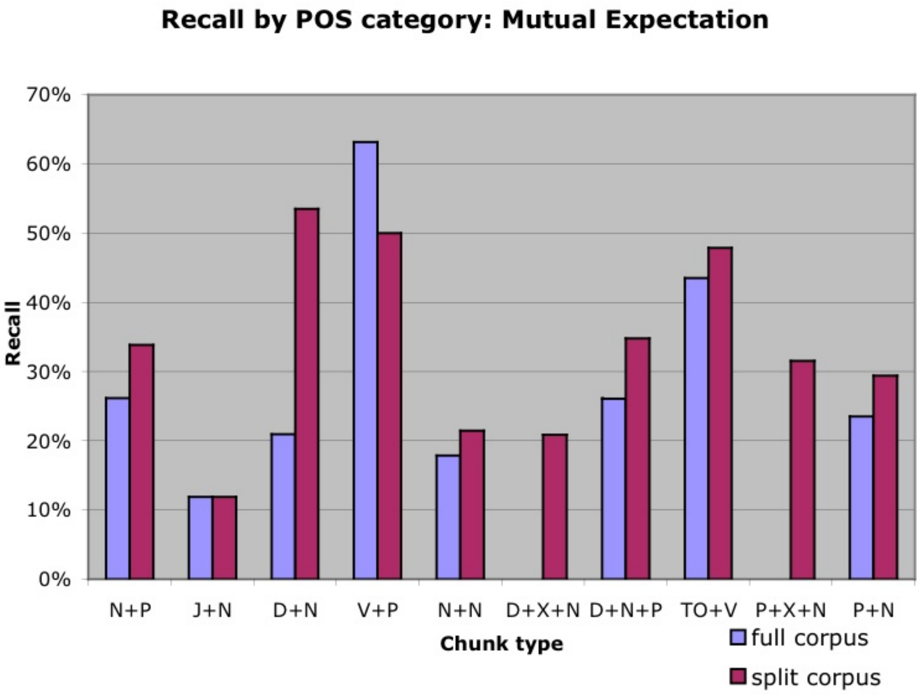

Figure 7.8: *Recall for 10 most common POS types: Mutual Expectation*

likely to be that in the full corpus, common words will have such high frequencies that even if they often appear with certain lower-frequency words, these appearances will not be statistically significant. For Raw Frequency of course,

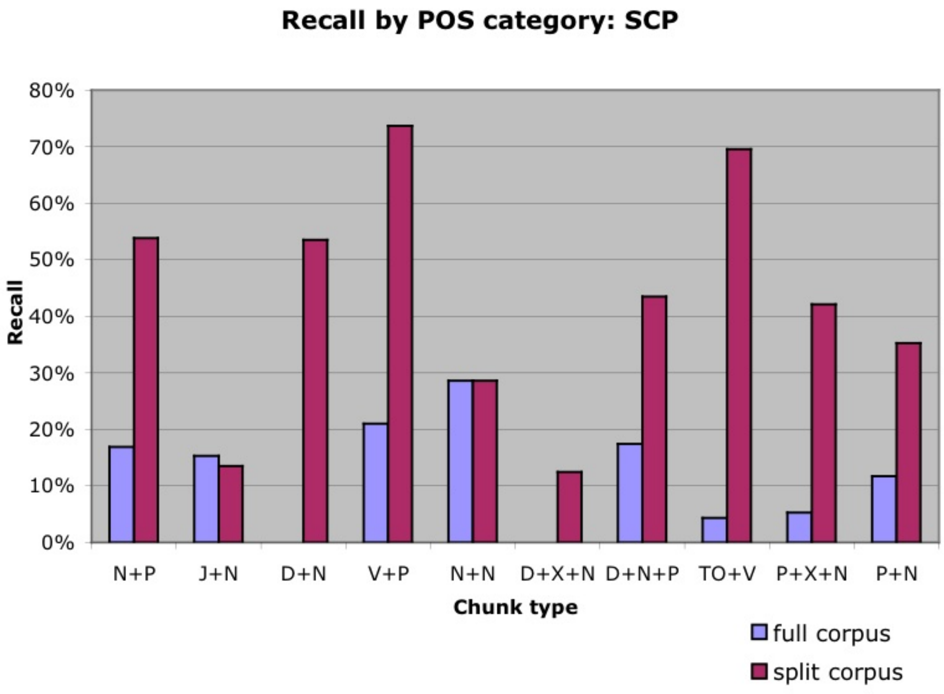

Figure 7.9: *Recall for 10 most common grammatical types: SCP*

this is not an issue, as anything that appears above a certain cutoff in a split corpus will also appear above that cutoff in the full corpus. The problem with Raw Frequency is that it fails to extract chunks involving low-frequency words, and many chunks do contain low-frequency words.

When using a full corpus for training, a trade-off occurs between methods that are good at finding chunks containing high-frequency words (hereafter referred to as HF methods) and methods that are good at finding chunks containing low-frequency words (hereafter referred to as LF methods). By splitting the corpus and thus lowering the relative frequency of high-frequency words, LF methods can also find chunks containing the high-frequency words. In order to improve results from HF methods in a similar manner, the opposite track would need to be taken - the corpus would need to be enlarged. Though some institutions and corporations have access to the large amounts of storage and processing power necessary for dealing with gigantic corpora, such access is by no means widely available. Even if it were, the fact that improved results can be obtained from splitting a corpus into smaller corpora rather than enlarging it means that such expanded use of storage and processing power is unnecessary.

Not only is the use of extremely large corpora unnecessary in this case, but it might also be seen as unrealistic. In a detailed study of child language exposure, Hart and Risley (2003) estimate that by the time child has reached the age of 4, the child will have heard between about 13-45 million words. If in four years, a child hears at most 45 million words and is able to produce and comprehend lexical chunks, then why should an automated system require a corpus on the scale of hundreds of millions or even billions of words to recognize lexical chunks? Of course, the human brain is vastly different from a computer, but at some level one must recognize that if a human brain is able to accomplish tasks based on a certain amount of data, then an automatic system should not require an amount of data that is orders of magnitude greater to accomplish a similar task. If we find that an automatic system is requiring far more information than humans could possibly have access to, we should perhaps look into what humans are doing in greater detail to find out how to improve the efficiency of the system. In the following section I discuss the links between human processing of lexical chunks and the results presented here, and how these insights can lead to improvements in both automatic extraction and second language instruction of lexical chunks.

7.3 Lexical Chunk Processing

7.3.1 Difficulties in L2 Acquisition

Though L1 learners readily acquire and produce the lexical chunks used in their native language, lexical chunks are notoriously difficult for L2 learners to acquire (Bahns and Eldaw 1993) (Brashi 2009) (Nesselhauf 2003) (Ying 2009). Two main hypotheses have been proposed to account for these difficulties: the first is that L2 learners simply have insufficient exposure to the chunks or to vocabulary appearing within them, and the second is that they are blocked, at least in production, from correctly producing chunks because they incorrectly transfer structures and/or lexical item translations from their native language. Regarding the first hypothesis, it may certainly be true in some cases, particularly when chunks involve obscure vocabulary items, but research has found that difficulties in chunk acquisition persist even for advanced L2 learners (Ketko 2009) (Abu-Ssaydeh 2006) and even after extended exposure to the target language (Siyanova and Schmitt 2007).

A further problem for the insufficient exposure hypothesis is that many of the chunks L2 learners have difficulties with contain very common words. For example, Nesselhauf (2003) found that a significant proportion of L2 learners' collocation errors in production were the result of misuse of function words such as prepositions and determiners. Examples she gives of collocation errors made by German learners of English include **fail in one's exams* instead of *fail one's exams*, **raise the question about* instead of *raise the question of*, and **get the permission* instead of *get permission*. These examples were drawn from a corpus of essays written by advanced English learners, mainly 3rd and 4th-year university students studying English. The chances that these students had not encountered the correct versions of the afore-mentioned collocations, none of which use particularly obscure words, are thus quite low.

These data raise the question of why L2 learners who have had extended exposure and/or periods of study of the target language continue to ignore correct input so that they produce and comprehend many lexical chunks incorrectly. The L1 transfer hypothesis can partially account for these data, but why should L2 learners who are otherwise quite proficient and who display advanced knowledge of the vocabulary and syntax of the foreign language continue to have difficulties with chunks? I believe part of the answer to this question resides in the earlier findings, discussed in Chapter 2, relating to lexical chunks and the brain.

7.3.2 Lexical Chunks in the Brain Revisited

As noted previously, substantial evidence exists that lexical chunks, unlike most language, are processed mainly in the right hemisphere of the brain. Data on acquisition also point to an increased role for the right hemisphere in early language acquisition. For example, Waldie and Mosley (2000) found evidence for right hemisphere involvement in early reading which then decreased as children became more advanced readers. Other studies have shown that damage to the right hemisphere is more likely to produce language deficits in infants than in older children (McFie 1961), whereas left hemisphere damage is less likely to produce language deficits in children under age 5 than in people over this age (Bates et al. 2001). Could right hemisphere involvement in early language acquisition somehow be related to right hemisphere involvement in lexical chunk processing? If so, it would help explain why older L2 learners, who no longer use the right hemisphere in language processing to the extent that it is used in early acquisition, have such difficulties with lexical chunks.

The hypothesis that right hemisphere involvement in early acquisition is related to lexical chunk processing is at this point merely a speculation, but further support for the idea comes from other studies that suggest a dichotomy between the right and left hemisphere in the processing of global vs. local phenomena. Namely, evidence from studies of visuospatial processing and memory have offered evidence for a left-hemisphere preference for local processing and a right-hemisphere preference for global processing (Delis et al. 1986). In Gazzaniga et al. (2002), this dichotomy is partially explained as a failure of the right hemisphere to abstract from local information. Indeed, the large body of literature citing heavy left hemisphere involvement in linguistic processing, particularly in processing of vocabulary and syntactic information, suggest that the ability to abstract away from local information is a particular specialty of the left hemisphere.

Abstracting from local information is clearly an important part of linguistic processing. It is particularly important in the operations required to connect a single example of a phonetic or written form of a word with the concept represented by that word and in the operations required to produce and comprehend syntax - precisely the areas where left hemisphere involvement appears to be strongest. At the same time, I suggest that lexical chunks are unique in that they resist such abstraction. Their noted tendencies to include rare or outdated lexical items and syntactic structures along with their oft-cited non-compositionality support this view. A reason that L2 learners find many lexical chunks so difficult may be that they continuously attempt to abstract away from the literal word meanings and/or syntactic structures, both in comprehension and production.

The role of the right hemisphere in both lexical chunk processing and early language acquisition can then be understood better as an inhibitor of the tendency to abstract. Young children, whose right hemisphere plays a larger role in language acquisition, abstract less from the structures they are exposed to. Rather, they first internalize these structures and then, with little explicit instruction in grammar, they begin to abstract away from these structures to create their own individual vocabulary and grammar of the language being learned. As the left hemisphere becomes more involved in language processing, this tendency towards abstraction becomes more developed, until it becomes the main means by which adults produce and understand language.

The problem in second language acquisition, as I suggest above, is that learners want to abstract too much. They are taught a large number of vocabulary items and grammatical rules long before they are taught most lexical chunks, so that by the time they turn their attention towards these chunks, they find it exceedingly difficult to accept the chunks as unqualified wholes. The desire to break up the chunks is too strong. An emphasis on learning individual lexical items and grammar rules may also delay L2 learners from even recognizing lexical chunks when they encounter them in text or speech. L2 learners may mistakenly analyze lexical chunks into their individual components and so fail to realize that the chunks themselves occur more often than should be expected by chance and are thus important structures to be learned in their own right.

7.3.3 From Brain to Computer

How do the difficulties L2 learners experience with lexical chunks relate to the findings presented here on the performance of different methods of lexical chunk extraction? As noted, measures that look at the relative likelihood for chunk words to appear together vs. apart perform relatively well for chunks containing low-frequency content words, but they are not as good at finding chunks containing high-frequency function words. When the corpus is split into several different subcorpora and chunks are found for each subcorpus and then compiled into a single large database, performance on recall of chunks containing high-frequency words substantially improves. I suggest that this improvement can, to a certain extent, be related to the blocking of abstraction tendencies that appears necessary for lexical chunk acquisition by humans.

By splitting a corpus, relative frequencies of function words are decreased, and so their significance in certain lexical chunks such as noun and verb-prepositional phrases becomes more apparent. With a full corpus, the frequency of such words is so high that they are not recognized as contributing parts of a chunk, even when they appear with very low-frequency words. An L2 learner who can immediately map new structures and vocabulary of a new language onto a pre-existing language can be compared to a program using a full corpus of data to discover linguistic information. In both cases, the learner is using a huge amount of data - perhaps too much to notice locally relevant phenomena like lexical chunks. By contrast, children learning language for the first time build a gradually increasing store of linguistic knowledge in concert with a gradually increasing store of world knowledge. They do not abstract away from the language they encounter to the same extent that older learners do, and so they are more easily able to retain things like lexical chunks. Similarly, a program that only uses a small portion of corpus data at a time will find items that would be missed if it were to attempt to look at all the data at once.

Naturally, these comparisons between human and machine processing are to be taken at a very general level. Human language acquisition and processing are incredibly complex phenomena for which much remains to be discovered. Most automatic processes for discovery of linguistic knowledge are quite far-removed from comparable human processes, and I do not suggest that automatic processes should mimic human processes. However, there is a lot that can be learned from the knowledge we do have on human language acquisition and processing, and this information can and should be used to improve our automatic systems where possible.

Chapter 8

Concluding Remarks

In this thesis, I have presented a method for the automatic extraction of lexical chunks of unrestricted type containing both contiguous and non-contiguous words. The method was run using five different statistical association measures that have previously been used in related work, and results were reported for precision and recall. The evaluation metric I used was novel in that I evaluated chunk-extraction methods by having them find chunks in a sample input text for which a gold standard list of chunks had been compiled through a combination of dictionary lists and human evaluations.

The most important results of my research were that two of the most commonly used measures for extraction of lexical chunks and related phenomena, namely Log Likelihood and Mutual Information, showed substantially improved performance when they were trained on several successively presented subcorpora as opposed to being trained on the entire corpus in one go. Performance gains were analyzed and mainly attributed to increased recall of chunks involving high-frequency function words, such as prepositions and determiners.

The importance of recognizing that function words are parts of chunks and not functioning in a normal, more syntactically analyzable manner was related to the importance in human lexical chunk recognition of *not* abstracting away from lexical and syntactic information and simply accepting the chunk as a whole. This tendency towards abstraction has been strongly associated with left hemisphere brain functions, and findings that the right hemisphere is involved in early language acquisition and in lexical chunk processing suggest that inhibition of the tendency is both easier for young learners and an important part of lexical chunk learning that is unique in language learning overall. The advantages of the split corpus method found here are thus unique to the particular task of lexical chunk extraction, just as processing of lexical chunks appears to be unique in the human brain. At the same time, lexical chunks clearly play an important role in proficient and fluent language use, and as such, the findings related to computational processes for their extraction are of import.

Bibliography

Abu-Ssaydeh, A. F.: 2006, Multi-word units: Can lexicography have a role in their acquisition?, *Babel* **52**(4), 349–371.

Acosta, O. C., Villavicencio, A. and Moreira, V. P.: 2011, Identification and treatment of multiword expressions applied to information retrieval, *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, Association for Computational Linguistics, Portland, OR, pp. 101–109.

Bahns, J. and Eldaw, M.: 1993, Should we teach efl students collocations?, *System* **21**(1), 101–114.

Barker, G. and Sorhus, H.: 1975, The importance of fixed expressions in oral spontaneity; volumes i-iv, *Technical report*, Public Service Commission of Canada, Ottawa.

Bates, E., Reilly, J., Wulfeck, B., Dronkers, N., Opie, M., Fenson, J., Kriz, S., Jeffries, R., Miller, L. and Herbst, K.: 2001, Differential effects of unilateral lesions on language production in children and adults, *Brain and Language* **79**, 223–265.

Becker, J. D.: 1975, The phrasal lexicon, *TINLAP '75 Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, Association for Computational Linguistics, Stroudsburg, PA, pp. 60–63.

Beeman, M. J. and Chiarello, C.: 1998, Complementary right- and left-hemisphere language comprehension, *Current Directions in Psychological Science* **7**(1), 2–8.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E.: 1999, *Longman Grammar of Spoken and Written English*, Pearson Education Limited, Harlow, England.

Bircan, P.: 2010, *Lexical Approach in Teaching Vocabulary to Young Language Learners*, PhD thesis, Anadolu University Institute of Educational Sciences, Eskisehir.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H. and Demecheleer, M.: 2006, Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test, *Language Teaching Research* **10**(3), 245–261.

Brashi, A.: 2009, Collocability as a problem in l2 production, *Reflections on English Language Teaching* **8**(1), 21–34.

Broca, P.: 1861, Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole), *Bulletin de la Société Anatomique* **6**, 330–357.

Broca, P.: 1865, Sur le siège de la faculté du langage articulé, *Bulletins de la Société d'anthropologie de Paris* **1**(6), 377–393.

Brownell, H. H., Michel, D., Powelson, J. and Gardner, H.: 1983, Surprise but not coherence: Sensitivity to verbal humor in right-hemisphere patients, *Brain and Language* **18**(20-27).

Chan, T. P. and Liou, H. C.: 2005, Effects of web-based concordancing instruction on efl students' learning of verb-noun collocations, *Computer Assisted Language Learning* **18**(3), 231–250.

Church, K. W. and Hanks, P.: 1990, Word association norms, mutual information, and lexicography, *Computational Linguistics* **16**(1), 22–29.

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

Constant, M. and Sigogne, A.: 2011, Mwu-aware part-of-speech tagging with a crf model and lexical resources, *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, Association for Computational Linguistics, Portland, OR, pp. 49–56.

da Silva, J., Dias, G., Guilloré, S. and Lopes, J. P.: 1999, Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units, *in* P. Barahona and J. J. Alferes (eds), *EPIA '99 Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, Springer-Verlag, London, pp. 113–132.

de Oliveira Santos, V. D.: 2011, *Automatic essay scoring: Machine learning meets applied linguistics*, Master's thesis, Rijksuniversiteit Groningen and Universität des Saarlandes, Groningen.

de Villiers, J. G. and de Villiers, P. A.: 1978, *Language Acquisition*, Harvard University Press, Cambridge, MA.

Delis, D. C., Robertson, L. C. and Efron, R.: 1986, Hemispheric specialization of memory for visual hierarchical stimuli, *Neuropsychologia* **24**(2), 205–214.

Dias, G., Lopes, J. G. P. and Guilloré, S.: 1999, Multilingual aspects of multiword lexical units, *Proceedings of the Workshop on Language Technologies - Multilingual Aspects*, Ljubljana, Slovenia, pp. 11–21.

Dias, G. and Vintar, Š.: 2005, Unsupervised learning of multiword units from part-of-speech tagged corpora: Does quanitity mean quality?, *12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, Progress in Artificial Intelligence Serie, C. Bento and A. Cardoso and Gaël Dias, Springer LNAI 3008, Covilhã, Portugal, pp. 669–680.

distributed by Oxford University Computing Services on behalf of the BNC Consortium: 2007, The british national corpus, version 3 (bnc xml edition), URL: http://www.natcorp.ox.ac.uk/.

Dunning, T.: 1993, Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics* **19**(1), 61–74.

edited by Lou Burnard: 2007, Reference guide for the british national corpus (xml edition), URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Evert, S. and Krenn, B.: 2001, Methods for the qualitative evaluation of lexical association measures, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 188–195.

Fahim, M. and Vaezi, R.: 2011, Investigating the effect of visually-enhanced input on the acquisition of lexical collocations by iranian intermediate efl learners: A case of verb-noun lexical collocations, *Journal of Language Teaching and Research* **2**(3), 552–560.

Farlex: 2011, The free dictionary - idioms and phrases.
URL: *http://idioms.thefreedictionary.com/*

Fillmore, L. W.: 1976, *The Second Time Around: Cognitive and Social Strategies in Language Acquisition*, PhD thesis, Stanford University.

Finlayson, M. A. and Kulkarni, N.: 2011, Detecting multi-word expressions improves word sense disambiguation, *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, Association for Computational Linguistics, Portland, OR, pp. 20–24.

Firth, J. R.: 1968, A synopsis of linguistic theory 1930-1955, *in* F. R. Palmer (ed.), *Selected Papers of J. R. Firth 1952-1959*, Indiana University Press.

Gazzaniga, M.: 1967, The split brain in man, *Scientific American* **217**, 24–29.

Gazzaniga, M. S., Ivry, R. B. and Mangun, G. R.: 2002, *Cognitive Neuroscience*, W. W. Norton, New York, NY.

Hakuta, K.: 1974, Prefabricated patterns and the emergence of structure in second language acquisition, *Language Learning* **24**(2), 287–297.

Hart, B. and Risley, T. R.: 2003, The early catastrophe: The 30 million word gap by age 3, *American Educator* **22**, 4–9.

Hsu, J. Y.: 2007, Multiword lexical units and their relationship to impromptu speech.

Hudson, J.: 1998, *Perspectives on fixedness: applied and theoretical*, PhD thesis, Lund University.

Jespersen, O.: 1924, *The Philosophy of Grammar*, The University of Chicago Press, Chicago.

Jiang, N. and Nekrasova, T. M.: 2007, The processing of formulaic sequences by second language speakers, *The Modern Language Journal* **91**(3), 433–445.

Jin, R. and Si, L.: 2004, A study of methods for normalizing user ratings in collaborative filtering, *Proceedings of the Twenty Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, ACM, Sheffield, UK.

Ketko, H.: 2009, Importance of "multiword chunks" in facilitating communicative competence and its pedagogic implications, *The Language Teacher* **24**(12), 5–11.

Klein, D. and Manning, C. D.: 2003, Accurate unlexicalized parsing, *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430.

Kuiper, K. and Haggo, D.: 1984, Livestock auctions, oral poetry, and ordinary language, *Language in Society* **13**(2), 205–234.

Lewis, M.: 1997, Pedagogical implications of the lexical approach, *in* J. Coady and T. Huckin (eds), *Second Language Vocabulary Acquisition*, Cambridge University Press, Cambridge, U.K., pp. 255–270.

Lichtblau, E., Rohde, D. and Risen, J.: 2011, Shady dealings helped qaddafi build fortune and regime.

Lieven, E., Salomo, D. and Tomasello, M.: 2009, Two-year-old children's production of multiword utterances: A usage-based analysis, *Cognitive Linguistics* **20**(3), 481–507.

Lin, D.: 1999, Automatic identification of non-compositional phrases, *Proceedings of ACL-99*, Association for Computational Linguistics, pp. 317–324.

Lindell, A. K.: 2006, In your right mind: Right hemisphere contributions to language processing and production, *Neuropsychology Review* **16**(3), 131–148.

Makkai, A., Boatner, M. T. and Gates, J. E. (eds): 2004, *Dictionary of American Idioms*, Barron's, Hauppauge, NY.

Manning, C. and Schütze, H.: 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

McFie, J.: 1961, The effects of hemispherectomy on intellectual functioning in cases of infantile hemispherectomy, *Journal of Neurology, Neurosurgery, and Psychiatry* **24**(3), 240–249.

Michelbacher, L., Kothari, A., Forst, M., Lioma, C. and Schütze, H.: 2011, A cascaded classification approach to semantic head recognition, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK, pp. 793–803.

Millar, N.: 2011, The processing of malformed formulaic language, *Applied Linguistics* **32**(2), 129–148.

Molloy, R., Brownell, H. H. and Gardner, H.: 1990, *Discourse Ability and Brain Damage*, Springer-Verlag, New York, NY, chapter Discourse Comprehension by Right-Hemisphere Stroke Patients: Deficits of Prediction and Revision.

Moon, R.: 1998, *Fixed Expressions and Idioms in English*, Oxford University Press, Oxford.

Myers, P. S. and Linebaugh, C. W.: 1981, Comprehension of idiomatic expressions by right-hemisphere-damaged adults, *in* R. H. Brookshire (ed.), *Clinical aphasiology: Proceedings of the conference*, BRK Publishers, Minneapolis, MN, pp. 254–261.

Myles, F., Hooper, J. and Mitchell, R.: 1998, Rote or rule? exploring the role of formulaic language in classroom foreign language learning, *Language Learning* **48**(3), 323–363.

Nattinger, J. R.: 1980, A lexical phrase grammar for esl, *TESOL Quarterly* **14**(3), 337–344.

Nattinger, J. R. and DeCarrico, J. S.: 1992, *Lexical Phrases and Language Teaching*, Oxford University Press, Oxford.

Nesselhauf, N.: 2003, The use of collocations by advanced learners of english and some implications for teaching, *Applied Linguistics* **24**(2), 223–242.

Nesselhauf, N.: 2005, *Collocations in a Learner Corpus*, Vol. 14 of *Studies in Corpus Linguistics*, John Benjamins Publishing Company, Amsterdam.

Ngomo, A.-C. N.: 2008, Knowledge-free discovery of domain-specific multiword units, *in* R. L. Wainwright and H. Haddad (eds), *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)*, Fortaleza, Ceara, Brazil, pp. 1561–1565.

Nivre, J. and Nilsson, J.: 2004, Multiword units in syntactic parsing, *MEMURA 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications, Workshop at LREC 2004*, Lisbon, Portugal, pp. 39–46.

Osman, N. B.: 2009, Routinizing lexical phrases on spoken discourse, *International Education Studies* **2**(2), 188–191.

Oxford Collocations dictionary for students of English: 2011.
URL: *http://www.lixiaolai.com/ocd/index.html*

Pawley, A. and Syder, F. H.: 1983, Two puzzles for linguistic theory: nativelike selection and nativelike fluency, *in* J. C. Richards and R. W. Schmidt (eds), *Language and Communication*, Longman Group Limited, London, chapter 7, pp. 191–226.

Pecina, P.: 2005, An extensive empirical study of collocation extraction methods, *Proceedings of the ACL Student Research Workshop*, Association for Computational Linguistics, Ann Arbor, MI, pp. 13–18.

Perera, N. S.: 2001, The role of prefabricated language in young children's second language acquisition, *Bilingual Research Journal* **25**(3), 327–356.

Peters, A.: 1983, *The units of language acquisition*, Cambridge University Press, New York, NY.

Pienemann, M.: 1998, *Language Processing and Second Language Development*, Vol. 15 of *Studies in Bilingualism (SiBil)*, John Benjamins Publishing Company, Amsterdam.

Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y.: 2009, Improving statistical machine translation using domain bilingual multiword expressions, *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, ACL and AFNLP, Suntec, Singapore, pp. 47–54.

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

Sinclair, J.: 2004, *Trust the Text*, Routledge, London.

Siyanova, A. and Schmitt, N.: 2007, Native and nonnative use of multi-word vs. one-word verbs, *IRAL* **45**, 119–139.

Smadja, F.: 1993, Retrieving collocations from text: Xtract, *Computational Linguistics* **19**(1), 143–177.

Stubbs, M.: 2007, Quantitative data on multi-word sequences in english: the case of the word world, *in* M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds), *Text, Discourse and Corpora*, Continuum, London, pp. 163–189.

SurveyMonkey.com: n.d., Surveymonkey.
URL: *http://www.surveymonkey.com*

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

van Lancker, D. R. and Kempler, D.: 1987, Comprehension of familiar phrases by left-but not by right-hemisphere damaged patients, *Brain and Language* **32**, 265–277.

van Lancker-Sidtis, D.: 2009, *Formulaic Language*, Vol. 2, John Benjamins Publishing Company, chapter Formulaic and novel language in a 'dual process' model of language competence, pp. 445–472.

van Lancker-Sidtis, D. and Postman, W. A.: 2006, Formulaic expressions in spontaneous speech of left- and right- hemisphere-damaged subjects, *Aphasiology* **20**(5), 411–426.

Vechtomova, O. and Karamuftuoglu, M.: 2004, Approaches to high accuracy retrieval, *in* E. Voorhees and L. Buckland (eds), *Proceedings of the 13th Text Retrieval Conference*, Gaithersburg, MD.

Waldie, K. E. and Mosley, J. L.: 2000, Developmental trends in right hemispheric participation in reading, *Neuropsychologia* **38**, 462–474.

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

Wible, D., Kuo, C.-H., Chen, M.-C., Tsao, N.-L. and Hung, T.-F.: 2006, A computational approach to the discovery and representation of lexical chunks, *TALN 2006 Workshop on NLP and CALL*, University of Leuven, Belgium.

Winner, E. and Gardner, H.: 1977, The comprehension of metaphor in brain-damaged patients, *Brain* **100**, 717–729.

Wood, D.: 2002, Formulaic language in acquisition and production: Implications for teaching, *TESL Canada Journal* **20**(1), 1–15.

Wood, D.: 2009, Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study, *Canadian Journal of Applied Linguistics* **12**(1), 39–57.

Wray, A.: 2002, *Formulaic Language and the Lexicon*, Cambridge University Press, Cambridge, U.K.

Ying, Z.: 2009, *An empirical study of L2 learners' use of lexical chunks and language production*, PhD thesis, Kristianstad University College, School of Teacher Education.

Yorio, C. A.: 1980, Conventionalized language forms and the development of communicative competence, *TESOL Quarterly* **14**(4), 433–442.

Appendix A

Rater Instructions

Lexical chunks are groups of words that, like individual words, are stored and retrieved as wholes in the memory of language users. Though they are formed from multiple words, they are not necessarily created from scratch by syntactic processes each time they are used.

Examples of lexical chunks in English are phrases like “How do you do?” and “Nice to meet you”. Lexical chunks also include idioms, like “dead as a doornail” and collocations - words that are commonly used together, as in “strong tea” as opposed to “powerful tea”, which one would not expect to hear from a native speaker of English.

Another feature of lexical chunks is that they can contain slots into which a certain set of words can go. For example “It’s a quarter to X”, “as X as”, and “Please pass the X”.

The following task asks you to look at different groups of English words, some with slots and some without, and decide whether these groups make up a lexical chunk. You should use your intuition as a native English speaker to rate each chunk as falling into one of the following categories:

Definitely not a chunk
Probably not a chunk
Maybe not a chunk
Could go either way
Maybe a chunk
Probably a chunk
Definitely a chunk

Please try to be as honest as possible in answering these questions. There are a lot of chunks to look at, so if you start to get tired, take a break and come back to the task. The chunks are split across 5 different surveys, so feel free to do one survey, take a break, and then do another one. While you do not have to do all the surveys in order, please do the fifth one last.

In the chunks you will see, the symbol ‘_X_’ is used to represent a gap of exactly one word. If ‘_X_X_’ appears, this means that two words are missing.

On each page, a box is provided for comments. If you would like to explain any of your answers or mention anything else about the chunks you see, please use this box to do so.

If you have understood all the instructions and are ready to proceed to the first question, please click on the ‘Next’ button to begin.

Appendix B

Chunks Found by Different Methods: A Sample

Below is the first sentence from the New York Times article, followed by lists of chunks that the five different methods found in the sentence[1]. Chunk lists are given for full and split versions of each method.

> WASHINGTON - In 2009, top aides to Col. Muammar el-Qaddafi called together 15 executives from global energy companies operating in Libyas oil fields and issued an extraordinary demand: Shell out the money for his country's $1.5 billion bill for its role in the downing of Pan Am Flight 103 and other terrorist attacks.

---

[1] In the evaluation, quantities such as "$1.5 billion bill" were written out in full as "one point five billion dollar bill". Further note that, as in training, all punctuation was removed, and capital letters were switched to lowercase

Log Likelihood	**Mutual Information**	**Raw Frequency**	**Mutual Expectation**	**SCP**
oil fields	executives from	in X top	in X ′s	executives from
an extraordinary	oil fields	companies X in	and X and	oil fields
point X billion	its role in	operating in	an extraordinary	the money for
five billion		in X ′s	out the	bill for
its role		oil X and	the X for	its role in
		fields and	money for	role in the X of
		and X an	for his	
		an X demand	his X ′s	
		an extraordinary	country ′s	
		out the X for	′s one	
		out X money	one point five	
		out the	one point X billion	
		the money for	one X five	
		the X for his	one point	
		the X for	point five	
		the money	for its	
		money for	its role in	
		for X country	its role	
		for his	role X the	
		his X ′s	role in	
		country ′s	in the X of	
		s X point	in the	
		′s one	the X of	
		one point five	and other	
		one X five		
		one point		
		point X billion		
		point five		
		bill for		
		for its		
		its X in the		
		its X in		
		its role		
		role in the X of		
		role in the		
		role X the		
		role in		
		in the X of		
		in the		
		the X of		
		and other		

Table B.1: *Chunks found in sample sentence by full versions of statistical methods*

Log Likelihood	Mutual Information	Raw Frequency	Mutual Expectation	SCP
operating in	in X top	in X ′s	in X ′s	in X ′s
oil fields	operating in	out the	an extraordinary	oil fields
and issued	in X ′s	the X for	out X money	an extraordinary
an extraordinary	oil X and	the money	out the	the X for
shell X the	oil fields	for his	the X for	the money
out the	fields and	his X ′s	the money for	money for
the money	and issued	country ′s	the money	country ′s
money for	an extraordinary	one point	money for	one point five
for X country	shell X the	for its	for X country	one point
for his	out the	its X in	for his	point X billion
his X ′s	the X for	role in	his X ′s	point five
country ′s	the money for	in the X of	country ′s	bill for
′s one	the money	in the	′s one	for X role
one point five	money for	the X of	one X five	for its
one point	for X country	and other	one point X billion	role in
point X billion	for his		one point five	in the X of
point five	his X ′s		one point	in the
bill for	country ′s		point X billion	the X of
for X role	s X point		point five	and other
for its	′s one		for X role	
its role	one X five		for its	
role X the	one point X billion		its X in	
role in the X of	one point		role in	
role in	point five		in the X of	
in the X of	bill for		in the	
in the	for X role		the X of	
the X of	for its		and other	
and other	its role			
	role X the			
	role in the			
	role in			
	in the X of			
	in the			
	the X of			
	and other			

Table B.2: *Chunks found in sample sentence by split versions of statistical methods*

Appendix C

Gold Standard Chunks

The following chunks composed the gold standard list used in the evaluation. Chunks containing up to 1 gap are shown in Tables C.1 to C.4 and chunks containing 2 gaps are shown in Table C.6.

30-year agreement
a batch
a batch of
a deal
a deal with
a dozen
a few
a member
a member of
a partner
a partner with
a piece
a piece of
a piece of the action
a player
a settlement
a settlement over
a stash
a stash of
a triumph
a triumph of
a X business
a X cable
a X company
a X consultant
a X culture
a X deal
a X dispute
a X dispute between
a X firm
according to
according to the
accounts in
acts of
acts of terrorism
administration of
adopted X lifestyles
after X restrictions
agreement on
agreements with
aides to
all X worth
allied countries
ambassador to
american ambassador
american businesses
american companies
american corporations
american diplomats
american officials
american X stars
amounted to
an agreement
an agreement on
an exhibit
an exhibit of
an interview
an X committee
an X of
an X on
banks X collected
based in
basis of
batch of
be done
be reached
become X partner
been frozen
believed X that
benefits of
bernard l madoff
bilateral X relations
bill for
billion dollar bill
billions of
billions of dollars
blocked access
blocked by
bonuses for
break X impasse
british virgin islands
british X islands
build fortune
business consultant
business dealings
business decisions
business gains
business partner
business with
by authorities
cement X deal
classified documents
clear that
close friend
close to
close X allies
col muammar el-qaddafi
collected X fees
colonel qaddafi
commerce department
committee set up
communications company
company officials
company spokesman
company X courted
condition of
condition of anonymity
consequences for
consultancy agreements
contracts worth
control of
controlled by
could not
could X be
daniel e karson
deal with
deals with
decision in
declined to
declined to identify

Table C.1: *Gold standard chunks, up to 1 gap (pt. 1)*

a X friend of
a X newspaper
a X of
a X official
a X over
a X son
a X spokesman
a X stake in
a X team
about to
access to
another X from
appeared willing
armed confrontations
around the world
around X world
as a partner
as a player
as part
as part of
as well
at least
by the X government
by X critics
by X diplomats
called together
came under
came under X investigation
caught in
caught in the middle
caught in the middle of
caught in X middle
cellular phone
demanded payment
did not
did X accept
did X comply
diplomatic cable
diplomatic cables
direct stake
discovered that
dispute between
do business
do so
documents obtained
dog of
doing business
dole out
dole out business
dollar bill
dollars in
dollars in cash
don t
don X think
done on
done on X basis
downing of
economic relations
economic sanctions
eldest son
energy companies
enhanced X standing
episode in
equipment for
exhibit of
exhibit of X paintings
extract millions
extraordinary demand
failed to
familiar with
fearful of
few years
fierce dispute
finally reached
financial manager
first X years
five billion
flush with
flush X cash
for acts of
for comment
for months
for the sake
for the sake of
for the X settlement
for X country
for X projects
for X role in
for X sake
for X sake of
foreign company
foreign firms
friend of
from the outset
from X companies
from X outset
frozen by
frozen by authorities
gained footholds
george w bush
give in
give X to
given up
global X companies
go through
going into
going into business
going X business
got caught
government investigation
had given
had made
have been
help of
help X avert
hold on
hollywood film
huge sums
hundred million
idris abdulla abed al-sonosi
in a batch
in a batch of
in an interview
in case
in cash
in february
in may
in retail
in the administration
in the administration of
in the downing of
in the middle
in the middle of
in the names
in the names of
in the years
in the years since
in violation
in violation of
in which the
in X administration
in X administration of
in X batch of
in X downing of
in X middle
in X middle of
in X names
in X names of
in X years
in X years since
included in
including X stash
indicated that
industry executives
infrastructure projects
insisted on
insisted that
international communications
international corporations
international sanctions
international X companies
international X company
intervention by
into business
into business with
into the X market
investment opportunity
investments in
involved in
is familiar
is familiar with
is uncertain
is using
is X with
large bonuses
launder X sums
lavish lifestyles
legal settlement
libyan culture
libyan government
lifting of
lifting of X sanctions
local agent
local business
local X councils
long-term contracts
looking back
loyal to
lucrative deal
lucrative fees
mad dog
made X clear
made X investments
making X decisions
managing partner
many of
may have
may have been
may help
member of
members of
members of X tribes
middle east
middle of
military intervention
millions of
millions of dollars
money for

Table C.2: *Gold standard chunks, up to 1 gap (pt. 2)*

more as
more than
moved into
muammar el-qaddafi
muhammad qaddafi
names of
new york
no investments in
nuclear capabilities
obtained by
obvious risks
occidental petroleum
of a X dispute
of a X plant
of billions
of dollars
of members
of the meeting
of the X intervention
of the X ruler
of X businesses
of X corporations
official in
official in X administration
officials believe
officials granted
officials said
officials warned
oil companies
oil company
oil exploration
oil fields
on condition
on condition of
on the basis
on the basis of
on X basis
on X basis of
on X contracts worth
on X settlement over
one billion
one point five
one point five billion
opened in
operating in
outcome of
over control of
pan am
part of
part of X agreement
partner at
partner with
pay X bonuses
payment for
personally involved
piece of
piece of X action
point five
point five billion
political allies
ponzi scheme
pop stars
posh homes
potential short-term X gains
president george w bush
private parties
provide equipment
provide X for
reached for
reached for comment
reached in
reached X comment
recent years
recently X been
remain loyal
removal from
removal from power
reportedly used
restrictions X lifted
ridiculed by
rife with
rife X corruption
risk-consulting firm
role in
role in the
royal estates
royal family
said that
sake of
sanctions on
say in
saying that
sectors of
serious consequences
serious consequences for
serve as
serve X as
service companies
set up
set up accounts
set X by
settlement over
seven million
seventy billion
several X officials
shady dealings
shell out
shell out X money
short-term X gains
shut down
signing bonus
signing off
signing off on
signing X on
smaller operators
soccer team
soda maker
son of
sovereign fund
speaking X on
speaking X on condition
spelling of
spokesman said

Table C.3: *Gold standard chunks, up to 1 gap (pt. 3)*

sponsored X exhibit
stake in
stash of
state department
state-owned company
state-run X company
stole billions
strong-arm tactics
struck X deals
summary of
sums of
sums of money
sums X money
t think
telecommunications firms
tens of billions
terrorist attacks
that the benefits of
that X benefits of
the administration
the administration of
the basis
the basis of
the benefits
the benefits of
the businesses
the cable
the client
the companies
the company
the country
the deal
the decision in
the downing of
the episode
the family
the government
the help of
the impasse
the lifting of
the lifting of X sanctions
the meeting
the middle of
the military
the money
the names
the names of
the officials
the outcome of
the payments
the plant
the regime
the sake
the sake of
the strongman
the way
the wealth
the west
the work
the world
the year
the years
the years since
the X administration
the X ambassador
the X ambassador to
the X company
the X family
the X for
the X government
the X intervention
the X market
the X of
the X officials
the X ruler
the X settlement
the X since
this is the
to be
to break
to broaden
to cement
to do
to do so
to dole
to enter
to extract
to give
to go
to hide
to identify
to keep
to pay
to provide
to rehabilitate
to renounce
to resume
to run
to serve
to shut
to the X government
to win
to work
top aides
top officials
trade agreement
trade restrictions
trade restrictions X lifted
trade with
trips to
triumph of
two hundred
two hundred million
two X million
uncommon for
under way
under X investigation
united states
united states commerce department
various sectors
violation of
violation of X sanctions
virgin islands

Table C.4: *Gold standard chunks, up to 1 gap (pt. 4)*

warned X officials
was not
was unthinkable
was X uncommon
went through
were hesitant
white house
with an agreement
with an agreement on
with cash
with the help
with the help of
with X agreement on
with X government
with X help
with X help of
with X kickbacks
work out
worked with
worth buying
worth X selling
would be
would have
would serve
years after

Table C.5: *Gold standard chunks, up to 1 gap (pt. 5)*

a X X cable
a X X in
a X X scheme
a X X summary
a X X summary of
according to the X X document
an X X company
as X X agent
blocked X X from
by X X government
came X X investigation
companies X X comply
done X X basis
for X X settlement
has X X stake
in the X X bombing
in the X X islands
in the X X years
in the X X years after
international X X telecommunications
international X X telecommunications firms
issue X X demand
large X X company
lifting X X sanctions
looking X on X decision
made X X in
might X X pay
of X X dispute
of X X dispute between
of X X intervention
of X X plant
one X X billion
outcome of X X intervention
outweighed X X risks
pay X X for
potential X X gains
rife X X kickbacks
sanctions X X imposed
sectors of the X X economy
shell X X money
the X X company
the X X documents
the X X islands
the X X of
the X X years after
warned X X that
with X X stars

Table C.6: *Gold standard chunks, 2 gaps*