7-1-2013

Quantitative determinants of prefabs: A corpus-based, experimental study of multiword units in the lexicon

Clayton Beckner

Clayton Beckner
Candidate

Linguistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Jill Morford, Chairperson

Joan Bybee

William Croft

Andrew Wedel

QUANTITATIVE DETERMINANTS OF PREFABS: A CORPUS-BASED, EXPERIMENTAL STUDY OF MULTIWORD UNITS IN THE LEXICON

by

CLAYTON BECKNER

B.A., Physics, Bradley University, 1994
M.S., English, Illinois State University, 1999
M.A., Linguistics, University of New Mexico, 2005

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Linguistics

The University of New Mexico
Albuquerque, New Mexico

July 2013

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of many individuals, both personally and academically.

I thank, first of all, the many participants who generously volunteered their time for my studies, since I could not pursue empirical research without their help.

I am grateful to all of my committee members, Jill Morford, Joan Bybee, Bill Croft, and Andy Wedel, for their mentorship. My research is better due to the challenges they posed, and the assistance they gave me. Special thanks are due to my chair, Jill Morford, for her probing questions, feedback, and support throughout the research process. She was a reliable technical resource, voice of reason, and source of encouragement, and could always help me get half-baked ideas to be more fully-baked. I am also especially grateful to Joan Bybee, whose pivotal work has been indispensible to my psycholinguistic research. I thank her for immersing me in usage-based linguistics, first as a student, then as a research assistant, and finally as a collaborator.

In addition to my committee members, I'm ever grateful to other UNM faculty, past and present, in linguistics and adjoining fields, including George Luger, Caroline Smith, Melissa Axelrod, Christian Koops, Phyllis Wilcox, Catherine Travis, Larry Gorbet, and Vera John-Steiner. They taught me the science of language and empirical research, provided technical help, and provided helpful encouragement.

I thank Nick Ellis, Vsevolod Kapatsinski, Michael Barlow, and Dawn Nordquist for helpful discussions, and for directing me toward useful resources at various stages. I thank Mark Davies for maintaining freely available corpora online, and for answering numerous queries about the workings of his search engines. I am grateful to UNM's Graduate Resource Center for assistance with statistics.

At UNM, I've been fortunate to be part of a thriving community of graduate students in linguistics. I have greatly benefitted from this community in various forms, including moral support and random breeze-shooting. There are too many people to thank everyone individually, but I should specifically acknowledge help at various stages of this project from Keri Holley, Gabe Waters, Jason Timm, Evan Ashworth, Susan Metheney, Logan Sutton, Susan Brumbaugh, Jeannine Kammann, Motomi Kajitani, Iphigenia Kerfoot, Benjamin Sienicki, Laura Hirrel, Sook-Kyung Lee, and Amy Lindstrom.

I'm also lucky to have a wonderful network of friends and neighbors outside of linguistics, including Kristen Fedesco, Drew Sedrel, Meisha Sedrel, Maggie Faber, Andrew Faber, Liz Bowden, Laura Tomedi, Karla Koch, Thondup Saari, Alexa Wheeler, Caleb Wheeler, Richard Frieday, Laura Lance, Kendra Watkins, Holly von Winckel, and Greg von Winckel. Many of these friends have been hugely supportive of my family, and watched our children in times of need. Many also tried out strange experimental tasks at early stages—some of which were destined for the cutting-room floor—and still they remain my friends. Thank you.

Finally, my family has provided me indispensible support and companionship. I am thankful to my children, Saoirse and Roan, for being an endless source of joy and hilarity. Roan, you learned the word 'dissertation' at an age that is surely abnormal, and you said it with alarming frequency. (To this day, though, I'm pleased you taught the word to your Pre-K class.) Now we can get back to reading Tolkien, and flipping in the

living room. To my wife, Danielle, I am indebted in countless ways. Thank you for goading and soothing in the right measure, thank you for your partnership, and your love. On to the next adventure!

QUANTITATIVE DETERMINANTS OF PREFABS: A CORPUS-BASED, EXPERIMENTAL STUDY OF MULTIWORD UNITS IN THE LEXICON

by

CLAYTON BECKNER

B.A., Physics, Bradley University, 1994
M.S., English, Illinois State University, 1999
M.A., Linguistics, University of New Mexico, 2005
Ph.D., Linguistics, University of New Mexico, 2013

ABSTRACT

In recent years many researchers have been rethinking the 'Words and Rules' model of syntax (Pinker 1999), instead arguing that language processing relies on a large number of preassembled multiword units, or 'prefabs' (Bolinger 1976). A usage-based perspective predicts that linguistic units, including prefabs, arise via repeated use, and prefabs should thus be associated with the frequency with which words co-occur (Langacker 1987). Indeed, in several recent experiments, corpus analysis is found to be associated with behavioral measures for multiword sequences (Kapatsinski and Radicke 2009, Ellis and Simpson-Vlach 2009). This dissertation supplements such findings with two new psycholinguistic investigations of prefabs.

Study 1 revisits a dictation experiment by Schmitt et al. (2004), in which participants are asked to listen to stretches of speech and repeat the input verbatim, after performing a distractor task intended to encourage reliance on prefabs. I describe the results of an updated experiment which demonstrates that participants are less likely to interrupt or partially alter high-frequency multiword sequences. Although the original study by Schmitt et al. (2004) reported null findings, the revised methodology suggests that frequency indeed plays a role in the creation of prefabs. Study 2 investigates the

distribution of affix positioning errors (*he go aheads*) which give evidence that some multiword sequences (e.g., *go ahead*) are retrieved from memory as a unit. As part of this study, I describe a novel methodology which elicits the errors of interest in an experimental setting. Errors evincing holistic retrieval are induced more often among multiword sequences that are high in Mutual Dependency, a corpus measure that weighs a sequence's frequency against the frequencies of its component words. Followup analyses indicate that sequence frequency is positively associated with affix errors, but only if component-word frequencies are included as variables in the model.

In sum, the studies in this dissertation provide evidence that prefabricated, multiword units are associated with high frequency of a sequence, in addition to statistical measures that take component words' frequency into account. These findings provide further support for a usage-based model of the lexicon, in which linguistic units are both gradient and changeable with experience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.0 The notions of 'prefab,' and frequency of co-occurrence.

In 1976, Bolinger wrote that in constructing sentences, 'speakers do at least as much remembering as they do putting together' (2). In recent years, a growing number of writers have argued that the word-by-word assembly model of syntax is insufficient, and that speakers rely heavily on formulaic chunks or 'prefabs' during speech comprehension and production (Pawley and Syder 1983, Sinclair 1991, Erman and Warren 2000, Bybee 2006; see Wray 2002 for a broader historical review). The strong version of the foregoing view holds that some multiword sequences are accessed HOLISTICALLY: two or more orthographic words may be retrieved from memory as a prepackaged unit, and the activation of the individual component words is diminished (Bybee 2002, 2003; Kapatsinski and Radicke 2009).

Moreover, it is reasonable to predict that prefabricated units will not be distributed randomly, but will be associated with repeated exposure to particular multiword sequences. A wide range of studies demonstrate that frequency has an effect on linguistic representation in phonology and morphology, and that frequency is crucial to mechanisms of grammaticalization (Bybee 2003, 2006, 2007; Bybee and Hopper 2001; Ellis 2002; Krug 2003). Likewise, from a usage-based standpoint, we would expect that holistic units will have some basis in frequency of use. In the usage-based literature, it is often stated that linguistic units arise out of the 'frequency of co-occurrence' or 'frequency of collocation' of two or more words (see Ellis 2002: 156; Bybee 2002: 317). Such formation of units can be explained intuitively if we imagine that repetition of words gradually strengthens their representation. Langacker (1987) writes:

> Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched to the point of becoming a unit, moreover, units are variably entrenched depending on the frequency of their occurrence (59, emphasis added).

Similarly, Bybee (2002) writes that repetition leads to the formation of syntactic constituents, and that 'items that are used together fuse together' (316). Across various domains, the human mind has a tendency to chunk sequences together when a pattern recurs, and this has the effect of rendering the system more efficient (Graybiel 1998, Bybee and Beckner 2010). In cognition, as well as in human-designed technologies, 'well-designed systems tend to have special representations for the kinds of information they have to process frequently' (Anderson 1978).

Given the foregoing, usage-based theory would predict that frequently co-occurring sequences of words will tend to become accessed as holistic units. Yet as it turns out, there remain some central questions to address regarding multiword sequences. The notion of what 'frequency' actually means is perhaps more complicated than it would seem (Krug 2003, Schmid 2010). There are in fact (at least) two broad mathematical interpretations of what 'frequency of co-occurrence' means with respect to a multiword sequence. The more intuitive interpretation of co-occurrence will be referred to here as token frequency: an absolute frequency measure in which we simply count how often some sequence occurs (i.e., a word sequence, *X Y*) in a corpus.[1] The alternative is to consider a relative frequency interpretation of co-occurrence; in this view, we take note of a word sequence *X Y* relative to all the other instances of the component words (that is,

[1] However, as I discuss at the end of Section 1.2, this is not to suggest that an actual integer tally is necessarily the best way to represent token frequency. Not every past exposure to a linguistic unit carries the same weight, depending on the time elapsed, and the total number of exposures. However, it is often useful to think of the number of exposures as proxy for a more complex representation in cognition.

X in the absence of *Y*, and/or *Y* in the absence of *X*). In this second interpretation, it turns out that even if *X Y* is a relatively rare sequence, we might say that *X* and *Y* 'frequently co-occur'— as long as *X* and *Y* tend to occur together.

In many cases, token frequency and relative frequencies overlap and are interlinked. For instance, as the token frequency of a word sequence soars, its conditional probabilities get a boost; increasing the number of tokens of a sequence also increases internal cohesion. However, it is possible for these measures to veer apart from one another. A given sequence might be characterized by a high relative frequency, but have a relatively low token frequency (e.g., *by dint of*; *scantily clad; vim and vigor*). Other word sequences can have rather high token frequency but relatively low relative frequency because the component words appear frequently and in many different contexts (e.g., *of it*).

In this dissertation, a recurring question is how we are to interpret 'frequency of co-occurrence' with respect to multiword sequences. Which frequency measure is (or which frequency measures are) important in the processing of multiword sequences, and in the formation of linguistic units over time? Is holistic retrieval of word sequences related to token frequency, related to relative frequency, or perhaps related to both? The distinction between token and relative frequencies is of considerable interest because recent experimental studies provide support for both types of measures (e.g., Tremblay, Derwing and Libben 2007, Kapatsinski and Radicke 2009, Ellis and Simpson-Vlach 2009). However, these studies in fact contradict one another in some ways, and the various accounts presented have yet to be reconciled. In Chapter 2, I begin to sort out this literature; I review existing evidence for (and against) both token frequency and relative

frequency measures as determinants of holistic retrieval, and discuss theoretical concerns regarding specific co-occurrence metrics.

A further goal of this dissertation is to supplement the behavioral evidence that certain multiword sequences are retrieved holistically, and thus in Chapters 3 and 4 I report new experimental research. In Chapter 3, I follow up on one of the experimental studies discussed in Chapter 2, which examines the effects of token frequency on verbatim recall in a dictation task. In Chapter 4, I use a new experimental methodology, involving the elicitation of speech errors, to examine the effects of both token frequency and relative frequency on holistic retrieval.

In the remainder of this chapter, I address various preliminaries that are relevant to an empirical investigation of prefabs. I briefly summarize what is assumed (what is not assumed) in a usage-based account of holistic units, in order to inform the predictions of the studies in Chapters 3 and 4.

1.1. **The gradient nature of holistic retrieval**.

In the present study, a prefabricated unit, or 'prefab,' should be understood to mean a multiword sequence which is retrieved from memory as a unit. More precisely, however, we might say that a prefab tends to be accessed as a unit, and gradience in this property is to be expected. Linguistic units of various types, from words to syntactic constituents to constructions are characterized by gradience rather than sharp delineation (Hay and Baayen 2005, Bybee and Scheibman 1999, Bybee and McClelland 2005, Croft 2001), and prefabs are no exception. Identifying a multiword sequence as a prefab makes no claim that it has no internal structure, nor that it can never be assembled word-by-word (Bybee 2010: 35 ff). Bolinger (1976) first introduces the prefab terminology with

the poetic suggestion that ‘our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs’ (1). Taking this imagery a step further, we might note that the availability of prefabs does not mean that other modes of construction are no longer available.

Moreover, prefabricated units may be to varying degrees analyzable, that is, the separate components of the unit may be accessible to some extent with respect to morphosyntax, and/or semantics (Langacker 1987). As one case in point, consider idiomatic sequences (*shoot the breeze; pull strings; kick the habit*), which even in generative models have special status in the lexicon as memorized units (Pinker 1999, Pinker and Ullman 2002). Nunberg, Sag and Wasow (1994) observe that many semantically opaque idioms are analyzable, insofar as they follow regular morphosyntactic patterns, and the syntactic components are interpretable. For instance, it is no coincidence that the idiom *spill the beans* takes the form of a transitive verb phrase (V NP), and English speakers have an understanding of what the component NP refers to. Moreover, although idioms are generally imagined to be fixed entities, it is possible to alter such sequences from their canonical form by drawing upon their analyzable properties. Nunberg, Sag, and Wasow (1994: 500 ff) provide many examples of idioms modified in context (*kick the filthy habit*; *that touched a couple of nerves*) or otherwise exhibiting componential structure (*My goose is cooked, but yours isn’t*).

Similar properties may be found among more semantically transparent sequences which exhibit varying degrees of fixedness: *broach the subject/topic/idea*; *wreak havoc*; *scantily clad*. Such sequences are arguably prefabs, even though they may permit variation in form (e.g., *wreak damage*; *scantily dressed*). Even stronger candidates for

prefabs might be grammaticalized or lexicalized phrases, including English emerging modals (*have to*; *used to*; *want to*) and complex prepositions (*in front of*; *by dint of*; *in spite of*). The interpretation of complex prepositions in particular has been the subject of some debate, based on observations that these sequences do not pass a complete battery of syntactic constituency tests, often based on introspective evidence[2] (Seppänen et al. 1994, Huddleston and Pullum 2002). However, the corpus data indicates that in actual usage, speakers tend to avoid interrupting or altering sequences such as *in spite of* (Hoffmann 2005, Beckner and Bybee 2009).

The position taken in the present dissertation is that empirical evidence for prefab status — whether in corpus data, or in experimental data —will be probabilistic in nature. While we may talk about a particular complex unit being retrieved 'compositionally' or 'holistically,' these terms actually represent opposite ends of a continuum. In any particular case, the component parts of a unit may be salient to varying degrees (Hay and Baayen 2005). One underlying cause for such gradience is that representation of linguistic units is complex and redundant, and multiple modes of access are in competition with one another. I describe these features of the prefab model in the next subsection.

1.2 **Storage vs. retrieval, frequency, and the maximalist lexicon**.

Often discussions of the mental lexicon pose research questions along the lines of 'are prefabs/formulaic sequences stored as units?' As one example, Schmitt, Grandage,

---

[2] For example, Seppänen et al. (1994) propose the following constructed sentence as evidence that *in spite of* fails the 'coordination' test: *In spite of your objections and **of** the point raised by Dr Andersson, we feel confident that we can proceed with the project.*

and Adolphs (2004) repeatedly say their research examines whether frequent multiword sequences 'are stored holistically or not' (128). However, I will argue that the important empirical questions involve the nature of retrieval from memory, whereas foregrounding storage in an either/or fashion frames the issues in a potentially misleading way. This is apparent if we consider that in a prefab model, the storage of linguistic units is likely to be vast, complex, and redundant. As it turns out, a wide range of sequences (including many that are not especially interesting) are likely to be 'holistically stored' in a sense, and changes in stored representations would need to commence long before holistic retrieval becomes possible.

In this dissertation, I assume the basic architecture of the lexicon to be an exemplar system that permits rich and redundant memory storage (Langacker 1987, Goldinger 1996, Pierrehumbert 2001, Wedel 2006). These exemplars include multiword sequences that are stored whole in memory (Bybee 1998, Bybee 2010, Bod 2006), along with information about frequency and additional factors, such as context of use and semantic-pragmatic inferences. In such an exemplar system, the mental lexicon is dynamic and heteromorphic, including a whole array of units varying in size, fixedness, and generality (Bolinger 1976, Wray 2008). Although this dissertation is principally focused on continuous multiword sequences, the exemplar model may of course be expanded to incorporate more abstract linguistic elements, including constructions of varying degrees of abstraction (Bybee 2010, Croft 2001, Goldberg 2006).

Memory storage in this system is truly 'maximalist' (Langacker 1991) insofar as every multiword sequence experienced leaves a trace in memory, even if the meaning of the sequence is entirely predictable from its component words (Bybee 2006). Speakers

simultaneously track the occurrence of multiword sequences of different lengths, and maintain exemplar categories for each of these sequences. Clearly there are some constraints on this system; we can assume that in processing there is some window size, *n*, which is the maximum number of words that might reasonably be grouped together.[3] Moreover, not every word sequence experienced takes up indefinite residence in memory. In exemplar models, memories decay with time (Pierrehumbert 2001), and word sequences that are not encountered again will fade from memory.

All the same, the proposed exemplar-based lexicon is clearly not constrained by strict parsimony in storage, as would be the case in generative models (Chomsky 1995). With respect to multiword sequences, redundant storage will be common because both parts and wholes will be represented, without any requirement to 'purge' duplicate entries (Langacker 1987). Even if a multiword sequence is stored (and often retrieved) as a unit, this unit will remain embedded in a network of associations, thus maintaining connections with component words elsewhere in the lexicon (Bybee 1998, Bybee 2006). Multiword exemplars compete against these component words for activation during speech comprehension and production (a point to which I return in Chapter 2; see Hay 2001, 2003). If component words are infrequent compared to the multiword sequence, that makes it more likely that the full sequence will be activated as a whole, and the component words will be activated to a lesser degree.

Assuming such a model, it becomes apparent why it is problematic to ask whether or not a particular complex unit is stored holistically. Multiword sequences are

[3] For example, based on working memory restrictions (Miller 1956), seven words (plus or minus two) might approximate an upper bound on the number of words (or chunked items) to be tracked in cognition. Of course, speakers can memorize much longer sequences of words verbatim, but such processes involve long-term memory.

represented as holistic units, and as assemblages of parts. Any particular activation of the sequence involves activating both of these memory representations to varying degrees; retrieval is dependent on the interaction between units in memory, and the nature of retrieval is gradient as a consequence.

Equating prefabs with 'holistically stored sequences' leads to an additional problem, insofar as this account cannot explain how holistic storage might develop for a prefab. The difficulty arises because it is not just highly frequent sequences that are tracked and stored in memory. If repetition plays a role in the creation of units, in fact it is necessary for all multiword sequences (that is, all multiword sequences, delimited by constraints of size and memory decay) to be tracked in memory. An argument to this effect has previously been presented by Bybee (2010),[4] as follows. If multiword sequences ever develop special representational status on the basis of frequency, then they must be stored in memory from the very first instance. If this were not the case, there would be no way for the multiword sequence to accrue frequency information at all. The logical problem is that usage cannot gradually cause a unit to be registered in cognition as 'frequent,' unless (a.) there is some representation for the unit in memory, and (b.) usage of this unit is tracked from the very beginning. Suppose that no frequency information is recorded until the one millionth exposure to a linguistic unit. How could this one millionth exposure ever be detected?

[4] Bybee's (2010) discussion is based on an argument included in a preprint version of Gurevich, Johnson, and Goldberg (2010), which in turn was partly based on observations in Bybee (2006). However, the relevant argument was omitted from the published version of Gurevich et al. (2010), and is currently not presented elsewhere by these authors (Adele Goldberg, p.c.). A similar argument regarding storage of multimorphemic words is presented in de Vaan, Schreuder and Baayen (2007), along with experimental evidence that a single exposure to a novel, complex word leaves a trace in memory.

These observations point to the necessity for a vast lexicon that contains at least a minimal entry for a very large number of n-grams, representing their associated frequencies. In a certain sense, it could be said that even very low-frequency sequences are 'stored' in memory, though this storage may be ephemeral, and the storage is not in itself interesting.[5] What truly distinguishes prefabs from other sequences, then, is not storage. The important difference has to do with retrieval—whether the sequence is primarily accessed as a whole unit, or whether it is primarily assembled from parts.

One final clarification is in order, regarding the representation of frequencies in the maximalist lexicon. In saying that an immense number of n-grams have their frequencies 'stored' in memory, I am not claiming that every exposure is remembered, nor that the information stored is a literal integer count of these exposures. First, as previously noted, we expect older experiences with linguistic units to diminish with time (Pierrehumbert 2001, 2002; Wedel 2006). This means that over time, rarely-encountered sequences will fade, and their stored representations may disappear altogether. Moreover, the distribution of experience over time plays an additional role via the 'power law of practice': in many cognitive domains, early exposures to some item or skill have the greatest impact, since the amount of learning with additional exposure levels off as practice accrues (Anderson 1982, Ellis 2002). Similarly, small frequency differences are

[5] The model I have sketched here proposes that whenever two linguistic units, X and Y, occur in sequence, this updates the frequency information for a separate unit in memory (XY). Such a 'localist' representation may give rise to concerns that memory demands in the lexicon would become intractable (Gluck and Myers 2001, Baayen and Hendrix 2011), but alternate models are certainly possible. It may simply be that the frequency of the transition between X and Y is tracked in cognition — whether as a numerical representation of frequency, as a probability, or as a connection strength in a Simple Recurrent Network (Elman 1990)— without creating a separate stored unit for XY until some threshold is reached. Similarly, Baayen and Hendrix (2011) propose that complex units are represented indirectly via inheritance from simple units, and frequencies are recorded in a co-occurrence matrix. Nevertheless, these alternate models still require devoting resources to track the occurrence of X and Y in sequence, from the very first co-occurrence, and frequency information is simply stored in a different form. In any of these cases, the central empirical question is not whether X and Y are stored as a unit, but whether X and Y are accessed together as a unit.

cognitively salient within a low-frequency range, but these same small differences diminish in importance in higher-frequency ranges (Hay and Baayen 2002). This nonlinear sensitivity to frequency can be described as a logarithmic relationship, and it seems there is a natural inclination for humans (and other primates) to perceive quantities logarithmically (Siegler and Booth 2004, Nieder and Merten 2007). Such findings are of interest in describing the underpinnings of memory representations; however, perhaps more importantly, they suggest certain methodological considerations. Often in behavioral research, it is appropriate to log-transform frequency counts, and I will follow this convention in the statistical analyses presented in this dissertation.

As a precursor to experimental studies of holistic retrieval (Chapters 3 and 4), in the following chapter I survey previous behavioral research in this domain, and delve into the quantitative measures of interest.

CHAPTER 2. QUANTITATIVE MEASURES OF PREFABS: BEHAVIORAL INVESTIGATIONS AND THEORETICAL ISSUES

2.0. Introduction.

One of the earliest insights that we might empirically investigate speakers' linguistic knowledge of co-occurrence patterns comes not from linguistics or psychology, but from information theory. Shannon (1951) observes that 'anyone speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language. Familiarity with the words, idioms, cliches and grammar enables him to fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation' (54). Shannon investigated these ideas rather informally, with a single participant, who he asked to predict the next letter in a series of queries from English text. Shannon's goal was to estimate natural language's entropy— a quantity which represents uncertainty (vis-a-vis predictability) in a message, and which is indirectly related to certain relative frequency measures (Manning and Schütze 1999). Given Shannon's focus on letter-by-letter orthographic representation, clearly his quantitative estimates were influenced by the predictability within words. Nevertheless, Shannon's observations had a broader scope, and he provided an early demonstration that much of English is predictable on the basis of linguistic knowledge.

In the last decade or so, there has been a resurgence in investigations of speakers' knowledge of words in sequence. In the experiment of Shannon (1951), it should be noted that the task essentially uses a participant's behavior to make inferences about the structure of language, that is, without comparing this behavior against patterns of usage (such as might be estimated from a corpus). In current research, of course, we are

interested in comparing corpus measures against observable behaviors in an experimental setting. The present chapter offers a partial review of such experiments as a source of evidence regarding holistic retrieval. Throughout this discussion, I also discuss in detail the quantitative corpus metrics of interest (involving both token frequency and relative frequency), along with their methodological concerns.

2.1. **Evidence that token frequency is associated with holistic retrieval**.

A number of recent experiments provide evidence that sequences that are high in token frequency are easier for speakers to process. Bod (2000) performs a reaction-time study which presents subjects with three-word sentences and asks them to indicate if they are acceptable. The study indeed finds that sentence frequency aids processing, since high-frequency sentences (such as *I love you*) have faster acceptance times than low-frequency sentences (e.g., *I test you*).

Similarly, Reali and Christiansen (2007) investigate the storage of two-word sequences, focusing on the processing of center-embedded constructions. Reali and Christiansen propose that it will be easier to process sentences with relative clauses if the embedded clause is a frequent two-word sequence. Thus, it should be easier to process *The attorney who* [*I met*] *distrusted the detective who sent a letter on Monday night* than *The attorney who* [*I distrusted*] *met the detective who sent a letter on Monday night*, because *I met* is more frequent than *I distrusted*. Using a word-by-word self-paced reading task, Reali and Christiansen find a gradual facilitation in processing over a large range of token frequencies.

Tremblay, Derwing and Libben (2007) investigate self-paced reading times for sentences that contain ‘lexical bundles,’ the most frequent multiword sequences of a

particular length in a corpus (Biber et al. 1999). Tremblay et al. perform a series of self-paced reading tasks using sentences containing either lexical bundles (LBs) or matched non-lexical bundle sequences (NLBs), using frequencies drawn from the full British National Corpus (following Biber et al. 1999). Tremblay et al. designed their NLB sentences by substituting for one 'pivot word' in each case, as in *If workers <u>don't worry about it</u> nothing will happen* (LB sentence) vs. *If workers <u>don't know about it</u> nothing will happen* (NLB sentence). For NLB sentences, the substituted word (i.e., *know*) is chosen so as to be more frequent than the pivot in the LB sentence. The reading times in the experiment give evidence that lexical bundles are processed more quickly than their non-lexical bundle counterparts, as long as the words are presented as multiword sequences or as full sentences (rather than being presented in a word-by-word fashion).

A recent study by Arnon and Snider (2010) investigates subjects' reaction times in an acceptability task. Subjects are presented with four-word sequences out of context, and are asked to indicate whether each item is a possible word sequence in English. The study finds that medium-frequency items are recognized faster than low-frequency items, and that high-frequency items are recognized faster than medium-frequency items. Moreover, Arnon and Snider (2010) pursue further analyses, in which they find that a binary (high/low) categorization for frequency does not provide the best fit to the data. Rather, they find that there is a continuous improvement in performance with relation to token frequency, across the whole range of frequencies considered. Thus, Arnon and Snider (2010) take this result as evidence for a usage-based account of multiword sequences, in which 'every additional occurrence of a sequence strengthens its activation' (76).

The foregoing studies are all consistent with the notion that as a multiword sequence increases in token frequency, it is more likely to be accessed holistically. However, these studies do not provide direct evidence for holistic retrieval[6], because other models of sentence production are able to account for improved performance with increased frequency. As frequency measures increase for a word sequence—in particular, transitional probability—that means that one part of the sequence helps to predict other parts of the sequence. Yet, as Kapatsinski and Radicke (2009: 500) write, 'Sensitivity to predictability does not necessarily imply that the predictor and the predicted fuse into a unit. Rather, co-occurrence may simply make the co-occurring words prime each other.' (See related comments in Tremblay et al. 2007: 19-20.)

Thus, in addition to demonstrating that frequent sequences are easy to process, to support a holistic access model it is necessary to show that the component words in a frequent sequence are relatively difficult to access. If a sequence of words is chunked together into a holistic unit, the component words should have reduced status as separate words, making them less likely to be accessed as individual items with respect to phonology, morphosyntax, and semantics (Hopper 1991, Haiman 1994, Boyland 1997, Bybee and Scheibman 1999, Bybee 2002, Beckner and Bybee 2009). Holistic sequences are, by nature, best retrieved as uninterrupted wholes. Wray (2006) writes:

> Just as a pianist who practices a difficult sequence of notes will, by virtue of that repetition, find it easier to play in [the] future, so it is reasoned that if you become used to producing the articulatory movements that result in a particular routine expression, then this pathway will be strengthened, until it becomes not only fast and reliable but also rather difficult to interrupt, modify, or, if it should go wrong, put back on track without starting from the beginning again (592, emphasis added).

[6] Indeed, Arnon and Snider (2010: 69) specifically acknowledge that their evidence does not address any claims regarding holistic retrieval.

Word monitoring studies provide one methodology that directly investigates holistic retrieval of word sequences, by looking for diminished accessibility of component words. Subjects are asked to monitor for a target word within word sequences having varying frequencies. Vogel Sosa and MacFarlane (2002) measure subjects' reaction times in monitoring for the word *of* in a series of spoken sentences. The stimulus sentences are grouped into four categories with respect to the token frequency of target bigrams, that is, two-word sequences consisting of a variable preceding word plus the word *of*. Vogel Sosa and MacFarlane find that subjects are slower to identify *of* in the most frequent bigram category, providing evidence for the holistic retrieval hypothesis.

Kapatsinski and Radicke (2009) perform a more extensive word monitoring study based on the word *up*. The researchers use token frequency to identify a wide range of Verb + *up* sequences, ranging from ultrahigh-frequency bigrams down to ultralow-frequency, constructed bigrams that are rather unexpected in actual usage. Kapatsinski and Radicke divide these stimuli into seven frequency bins across the spectrum. The reaction time results take the form of a U-shaped curve: subjects are less adept at detecting the particle *up* in extremely improbable sequences, but this ability gradually improves as the bigram frequency increases—indicating that moderate increases in token frequency improve processing in a gradient way, due to increased predictability. However, the ability to detect *up* suddenly declines again in the ultrahigh frequency category, and in this sense the results resemble those of Vogel Sosa and MacFarlane (2002). With respect to this high-frequency end of the spectrum, Kapatsinski and Radicke write that 'the stronger the whole, the weaker the parts' (2009: 518). Following earlier proposals by Alegre and Gordon (1999) regarding a frequency threshold of storage for

multimorphemic words, Kapatsinski and Radicke argue that multiword sequences are retrieved holistically[7] from the lexicon if they are extremely high in token frequency.

If we are to synthesize the various results presented in this section, it is evident that token frequency of a multiword sequence has an effect on retrieval. A sequence becomes increasingly accessible as frequency increases; this is true across the entire spectrum of frequencies, as demonstrated by Reali and Christiansen (2007) and Arnon and Snider (2010). Moreover, above a certain token frequency threshold, the component parts of a sequence become gradually less accessible, as found by Kapatsinski and Radicke (2009), providing direct evidence of holistic retrieval for the sequence.

2.2. **Problems with a purely token frequency-based account.**

Despite the evidence presented in the previous section, there are several complications to be addressed regarding the relationship between high token frequency and holistic retrieval. First, high token frequency is not a necessary condition for unithood of a multiword sequence. Numerous writers have observed that certain word sequences are 'formulaic' and well-known by speakers, even though their token frequency is quite low. For instance, Wray and Perkins (2000) discuss the 'many formulaic sequences whose culturally-based familiarity belies their comparative rarity in real text'—for instance, *That's another fine mess you've gotten me into* (7).

Aside from phrases with cultural significance, speakers are familiar with many mundane expressions that are low in frequency. Hoffmann (2005) argues that English

[7] More precisely, what Kapatsinski and Radicke say is that 'the highest-frequency phrases are stored in memory as lexical unit but... a phrase needs to be extremely frequent to be stored in the lexicon' (2009: 516, emphasis added). However, as I argued in Chapter 1, it is preferable to assume that some type of storage of multiword units commences long before holistic retrieval becomes likely. Indeed, Kapatsinski and Radicke consider this alternate interpretation in a footnote (2009: 516, n. 7).

complex prepositions such as *in front of*, *by dint of*, and *in spite of* are grammaticalized phrases, even though they are relatively rare. The corpus evidence indicates that these complex prepositions are relatively fixed phrases that speakers tend to retrieve without interruption (Hoffmann 2005, Beckner and Bybee 2009). Bybee (2010) says that prefabs are conventional sequences which nonetheless 'do not need to be highly frequent. Just as we can learn a new word with only a few repetitions (sometimes for native speakers only one exposure) so also can we register a prefab after experiencing only one or two tokens' (60). (See also Bybee 2007: 16).

Moreover, it seems that high token frequency is also not sufficient as a determinant of holistic retrieval. Ellis et al. (2009) observe that 'not all high frequency n-grams have clearly identifiable or distinctive functions or meanings; many occur simply by dint of the high frequency of their component words' (64). It is true that the most frequent word sequences from a corpus may not be very intuitive as units, if no other factors are controlled. For instance, consider the ten most frequent word sequences of length 2, 3, 4, and 5 from the Switchboard corpus (2.9 million words, Godfrey, Holliman and McDaniel 1992), presented in Table 2.1[8].

The results from such a purely frequency-based corpus search are rather mixed. There are some nice finds here: discourse-related phrases like *you know*; *I think*; *I mean*; *I don't know*; *as a matter of fact*; and other lexicalized phrases like *a lot of*; and *a little bit*. However, it is clear that frequency alone does not retrieve only clear instances of

[8] I assembled these lists by writing a Java script that tallies n-grams in a corpus, and then sorts by frequency. A few of the oddities in this list (like *I I*) arise here due to the idiosyncrasies of the Switchboard textfiles, which contain no punctuation. Some of these strange results disappear if we use a different corpus such as SBCSAE (Santa Barbara Corpus, DuBois et al. 2000-2005) that marks intonation units. However, counterintuitive sequences persist in the top few result for SBCSAE, including *I don't know if; I don't know what*; *I don't know how*; *and I was just*.

formulaic or conventional phrases from a corpus: consider other very high-frequency n-grams in this set like *you know and*; *I don't know I*; and *I don't know if*. Indeed, Biber (2010) writes that most of the highest-frequency n-grams in a corpus ('lexical bundles') are 'not idiomatic in meaning and not perceptually salient,' and they 'usually do not represent a complete structural unit' (170).

BIGRAMS	**TRIGRAMS**	**4-GRAMS**	**5-GRAMS**
you know (34,487)	*a lot of*	*I don't know I*	*as a matter of fact*
I think (12,830)	*I don't know*	*I don't know if*	*what do you think about*
I don't (11,244)	*you know I*	*a lot of people*	*I think a lot of*
and I (9,907)	*uh you know*	*and things like that*	*one of the things that*
in the (8,791)	*and you know*	*a lot of the*	*I don't I don't know*
and uh (8,455)	*you know and*	*or something like that*	*you know a lot of*
of the (8,401)	*you know the*	*I don't know what*	*I don't know I think*
I I (8,320)	*I don't think*	*uh I don't know*	*I don't know I I*
a lot (8,128)	*I think that*	*and uh you know*	*at the end of the*
I mean (7,256)	*a little bit*	*I don't I don't*	*I don't know if you*

TABLE 2.1. Ten most frequent n-grams in the 2.9-million-word Switchboard corpus, for four different spans. Token frequencies of bigrams are in parentheses, for purposes of comparison with Table 2.2.

Beyond these objections on the basis of speaker intuition, there are behavioral studies which would seem to show that high token frequency is not associated with holistic retrieval of word sequences. Moreover, on first glance these studies seem to contradict the evidence presented in section 2.1, by indicating that there is no processing advantage for high-frequency sequences.

First, I will briefly mention a speech dictation study performed by Schmitt, Grandage and Adolphs (2004), which the researchers interpret as yielding a null result with respect to token frequency of multiword sequences. In this experiment, subjects are asked to listen to stretches of speech and repeat the input verbatim, after performing a math task intended to disrupt short-term memory so as to encourage reliance on prefabs.

Schmitt et al. (2004) find that there is no relationship between token frequency of a word sequence and subjects' accuracy in reproducing that sequence. However, this conclusion is questionable due to several problematic features of the study. Most notably, the coding conventions used to interpret the experimental data are puzzling, and even run counter to the researchers' own predictions about unitary sequences. For the time being, I will defer a longer critique of Schmitt et al. (2004), since the verbatim memory task forms the topic of Chapter 3.

Another line of evidence against token frequency accounts comes from a series of studies by Nick Ellis and colleagues (Ellis, Simpson-Vlach and Maynard 2008, Ellis and Simpson-Vlach 2009). Ellis and Simpson-Vlach (2009) perform four experiments examining processing of word sequences (of length 3, 4, or 5 words) in high-, mid-, and low-frequency categories. For each word sequence in the study, Ellis and Simpson-Vlach (2009) measure reaction time in an acceptability judgment task; measure fluency in reading the sequence aloud; measure priming of the final word using voice onset time in reading aloud; and measure comprehension in context, assessed through reaction time in an accessibility task. (Three of these studies are also described in Ellis et al. 2008). For all four studies, the token frequency of the n-gram was found to have no significant effect for native English speakers. Ellis et al. (2008) offer various explanations for why native speakers seem to be insensitive to token frequency. Native speakers, they argue, have 'reached asymptote' in processing multiword sequences as long as they are of a certain minimum frequency threshold; further increases beyond that basic familiarity do not lead to any boost in processing (2008: 390). Moreover, high-frequency n-grams are of limited usefulness for reasons discussed earlier in this section: they are often incomplete units

(and straddle different syntactic phrases), and they have no unified or idiomatic function (391).

Yet it is strange that the results in the Ellis et al. experiments are so contrary to existing evidence that token frequency improves processing of multiword sequences (as reviewed in Section 2.1). This is all the more striking because the methodologies used by Ellis et al. are quite similar to those used in experiments that have positive results. For instance, Ellis and Simpson-Vlach's (2009) first experiment is almost identical to the acceptability judgment task of Arnon and Snider (2010). Ellis and Simpson-Vlach's fourth experiment, involving comprehension in context, is almost identical to the self-paced reading task in Tremblay et al. (2007).

Thus it is worth considering whether there are important differences in experiment design. There are peculiarities in the design of the Ellis et al. studies which might make us hesitant to draw generalizations about multiword sequences in English. In these experiment, all of the n-grams used were 'academic formulas,' chosen because they are more frequent in academic corpora than in non-academic corpora[9] (see Ellis et al. 2008: 379-38, Ellis and Simpson-Vlach 2009: 64-65). Given this constraint, the sequences labeled as 'high-frequency' in the experiment might be limited to particular contexts, and may not be especially frequent in a speaker's overall experience. Indeed, if we examine the sample stimuli listed in Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), we encounter some cases in which the 'high-frequency' label is surprising, based

[9]Ellis et al. do not specify which corpora they used when they measured token frequency in the final classification of stimuli as high-, medium-, and low-frequency. Their full set of texts used during initial stimulus selection consists of 10 million words, just over half of which are from nonacademic sources (58% nonacademic, 42% academic text overall, including the Switchboard Corpus, FLOB, FROWN, the Michigan Corpus of Academic Spoken English, academic portions of the British National Corpus, and a database of academic journal articles).

on values drawn from the more wide-ranging Corpus of Contemporary American English (COCA, Davies 2008-). As one example, we can compare the 'high-frequency' sequence *the content of* (5.2 per million in COCA) with the 'medium frequency' sequence *and at the* (16.5 per million) and the 'low-frequency' sequence *that the only* (5.69 per million). Thus, in some cases, the sequence categorized as 'high-frequency' in fact does not seem to be higher in frequency than the stimuli categorized as low- or medium-frequency. In other cases, there seem to be no substantial frequency differences between the frequency categories (again, assuming we are consulting a corpus that contains comparatively little academic language, such as COCA). Given such complications, we should be wary of using these findings to draw conclusions about the general cognitive importance of token frequency, especially since contrary evidence exists.

In sections 2.3 and 2.4, I discuss further details of the Ellis et al. studies.

2.3. **Experimental support for relative frequency accounts.**

In recent years, more attention has been paid to various relative frequency measures as an alternative to, or supplement to, token frequency accounts. As introduced in Chapter 1, by 'relative frequency,' I typically mean a frequency measure that controls for the frequency of one or more component words in a multiword sequence. More generally, relative frequency can include any measure which reports absolute frequency relative to other frequencies, typically as a ratio between the frequency of a complex form (multiword or multimorphemic) and the frequencies of its component parts. In a number of studies, Hay (2001, 2002, 2003) has argued that high relative frequency best predicts the formation of complex units in morphology. Hay argues that there has been a

'misguided' emphasis on absolute (token) frequency in the usage-based literature and in much of the psycholinguistic literature (2002: 530).

Hay's position is inspired by morphological race models (e.g., Frauenfelder and Schreuder 1992), which hold that that during activation of a complex morphological form, holistic access competes against access of the individual parts. Within a morphological race model, it is reasonable that relative frequency will have an effect on retrieval: which access route 'wins' the race depends on the frequency of the fully-assembled form vis-a-vis the frequencies of the component parts. For Hay (2001), if a derived word is more frequent than its base, then the derived form is likely to be retrieved as a whole, rather than compositionally. For instance, *im+patient* is about twice as frequent as *patient*, and Hay (2001) argues that accessing the former is thus likely to proceed without depending on accessing the latter.

Hay (2001) offers several lines of evidence in support of this account. In one experiment, Hay (2001: 1047-8) asks subjects to assess the complexity of affixed words in a metalinguistic task; in each query, a derived word that is more frequent than its base is pitted against a derived word that is less frequent than its base. Around 65% of the time, subjects describe the derived word that is more frequent than its base as being less morphologically complex, from which we infer that there is diminished activation of the word's component parts. In a second study, Hay (2001) examines dictionary definitions of derived words to assess semantic transparency. Here, it is assumed that derived forms which do not refer to their base in the definition are semantically opaque, and such forms are accessed holistically rather than via assembly of (semantic) components. Hay finds that relative frequency predicts the development of semantic opacity better than absolute

frequency (though see discussion in section 2.5). For instance, with respect to prefixed items, 38% of words in the high relative frequency category are opaque, compared with 21% of words in the high absolute frequency category. Based on chi-squared analyses, Hay claims that 'the absolute frequency of the derived form appears to have absolutely no effect on' semantic opacity of derived words (2001: 1058). Hay further argues that the apparent effects of absolute frequency in prior studies may be secondary to more important effects from relative frequency, since absolute and relative frequencies are not independent of one another. (However, see Section 2.5 for further discussion).

With respect to multiword sequences, a number of writers have argued that relative frequency of some kind may be important in cognition, leading to the creation of multiword chunks. Bybee (2002, 2010) takes an inclusive approach, arguing that relative frequency effects probably play a role alongside token frequency. For instance, Bybee (2002) says that chunking occurs as a result of very high frequency, but 'more subtle effects can also be found in cases of co-occurrence that are less frequent, leading me to hypothesize that chunking and constituency relate directly to frequency of co-occurrence' (317). The quantitative measures associated with these lower-frequency cases would involve relative frequency, that is, frequency of a whole unit that controls for frequencies of the component parts.

Transitional probability is one such measure; for a two-word sequence, the transitional probability is the raw frequency of the sequence, divided by the token frequency of the first word (Gregory et al. 1999). Transitional probability may also be extended to higher-order word sequences, in which case the quantity reports how likely the final word is to appear given that the rest of the sequence has already occurred

(Jurafsky et al. 2001). Beckner and Bybee (2008) show that many complex prepositions, such as *by dint of, by way of,* and *in spite of,* are characterized by astonishingly high transitional probabilities, even though the word sequences themselves are rare (Hoffmann 2005). These high relative frequencies are one indicator that such sequences are chunked units, alongside other evidence such as morphosyntactic fixedness and semantic opacity (Beckner and Bybee 2009, Hoffmann 2005; see Chapter 1).

Another common metric for relative frequency is Mutual Information (MI), a bidirectional likelihood measure over a word sequence.[10] In its simplest form, Mutual Information divides the frequency of a word sequence by the frequencies of both words in the sequence (often log-transformed). The Mutual Information[11] for a two-word sequence w_1w_2 would then be given by Equation 2.1, where f(x) is the token frequency of a word (or word sequence) (Fano 1961, Church and Hanks 1989, Oakes 1998, Gregory et al. 1999, Manning and Schütze 1999).

(Equation 2.1) $$\text{MI}(w_1w_2) = \log_2 \left[\frac{f(w_1w_2)}{f(w_1)*f(w_2)} \right]$$

It is sometimes said that a higher Mutual Information value indicates a 'stronger cohesion' among words (Gregory et al 1999: 9), or that it is a 'measure of how "tightly" linked two words are' (Davies 2008). More specifically, we may note that this measure tells us how much each word in the pair predicts the other. The ratio in Equation 2.1

[10] One way of viewing Mutual Information for a sequence XY is that it combines the metrics of Transitional Probability (how predictive X is of Y) and Backward Transitional Probability (how predictive Y is of X) (Pelucchi, Hay, and Saffran 2009). Since MI combines two directional measures (Forward and Backward Transitional Probability), it is thus 'bidirectional.'

[11] The quantity described here is also known as the pointwise mutual information. In Information Theory, more sophisticated (and less intuitive) measures exist that are also known as Mutual Information (see Manning and Schütze 1999: 182). I will continue to use the term 'Mutual Information' to refer to pointwise mutual information, following the convention set by Ellis et al. (2008).

quantifies how often the words appear together, in contrast with how often they occur separately. Certainly, such a relative frequency measure (following Equation 2.1, or some variant) may be given a psychological interpretation in a syntactic competition model: the frequency of a complex form in the numerator competes against component frequencies in the denominator.

There are various ways that Mutual Information may be generalized to word sequences longer than two words, but typically the measure will include the frequency of the entire multiword sequence, divided by the product of the individual word frequencies. The collocational analysis program Collocate (Barlow 2004) makes use of the following general definition for Mutual Information of an n-gram, where N is the corpus size (Barlow, p.c. 2010).

$$\text{(Equation 2.2)}\quad MI(w_1w_2w_3...w_n) = \log_2 \left[\frac{N^{n-1} * f(w_1w_2w_3...w_n)}{f(w_1)*f(w_2)*... f(w_n)} \right]$$

There are several reasons Equation 2.2 includes a term for the number of words in the corpus (N). Including the corpus size allows us to make some broad comparisons between MI values drawn from different corpora: a high whole/part ratio observed in a corpus of 100 million words should be given more weight than the same ratio in a corpus of 1 million words. (Nevertheless, MI scores should always be treated with caution, as discussed in the following section.) Another (perhaps more practical) reason to include N (raised to the n-1 power) in the equation is that it makes the resulting MI scores more accessible to human readers. Omitting the size-of-corpus term results in negative MI values, since in Equation 2.1 the argument to the logarithm will almost always be less

than 1. However, including the N^{n-1} term as in Equation 2.2 yields more readily comparable, positive numbers. Consequently, in this dissertation I will generally report the easier-to-read MI values yielded by Equation 2.2.

For comparison with the token frequency results given above in Table 2.1, in Table 2.2 I present two-word sequences with very high Mutual Information, based on an automated Java search of the Switchboard corpus using Equation 2.1. The items listed here are the twenty bigrams with the highest Mutual Information (in descending order), restricting the search to bigrams that occur with a frequency of at least 10 per million. The token frequency of each sequence is given in parentheses.

1. *Los Angeles* (55)	11. *per se* (39)
2. *et cetera* (58)	12. *Super Bowl* (39)
3. *Saint Louis* (35)	13. *science fiction* (33)
4. *Rhode Island* (46)	14. *current events* (48)
5. *Star Trek* (57)	15. *word processing* (42)
6. *Peace Corps* (88)	16. *General Motors* (34)
7. *Soviet Union* (76)	17. *South Dakota* (43)
8. *San Diego* (30)	18. *checking account* (36)
9. *San Francisco* (91)	19. *Washington D.C.* (45)
10.*San Antonio* (91)	20. *square feet* (30)

TABLE 2.2. Twenty Switchboard bigrams with the highest Mutual Information (minimum token frequency of 10 per million). Token frequencies are in parentheses.

The results presented in Table 2.2 do in fact represent rather intuitive multiword units, with a strong tendency toward proper nouns. It is noteworthy that in each case, at least one of the words in the bigram has a restricted distribution (such as *Trek* in *Star Trek*), and this contributes to the especially high Mutual Information values in this set.

In addition to yielding intuitive word sequences, high Mutual Information values also prove useful in accounting for certain patterns in linguistic behavior. For instance,

the likelihood of tapping word-final /t/ or /d/ is best predicted by the Mutual Information between the stop-final first word and the second word (Gregory et al. 1999).

In Section 2.2, I reviewed the null experimental results of Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), who did not find evidence that token frequency is associated with ease of processing (and again, who used a stimulus set that focuses on academic English). However, these studies find evidence that high Mutual Information improves speakers' processing for 3, 4, and 5-word sequences, where Mutual Information is defined as in Equation 2.2. Ellis and Simpson-Vlach (2009) found significant boosts in processing as a result of Mutual Information, in all four tasks: reaction time in a grammaticality judgment task; voice onset time when reading aloud; reaction time for recognizing a sequence's final word; and reaction time for comprehension in context. It should be noted that all four experiments address ease of processing rather than giving direct evidence of holistic retrieval (as discussed in Section 2.1). Indeed, Ellis and Simpson-Vlach (2009) make no claims regarding holistic retrieval, but do argue that they are investigating formulaic sequences that are characteristic of fluent, native speech.

Regarding the difference between token frequency and relative frequency, Ellis et al. (2008) conclude that relative frequency of co-occurrence is more important than raw frequency of occurrence: 'tuning the system according to frequency of occurrence alone is not enough for nativelike accuracy and efficiency. What is additionally required is tuning the system for coherence – for co-occurrence greater than chance' (2008: 391). Ellis et al. (2008) interpret Mutual Information as described here, as 'the degree to which the words in a phrase occur together more often than would be expected by chance' (380). Strictly speaking, this interpretation is incorrect; there is no sense in which a

particular MI value reflects ‘pure chance’ and a maximum MI value would reflect ‘pure correlation.’ Different MI values must be assessed carefully on the basis of the token frequencies of the word sequences involved (Manning and Schütze 1999: 180-182), as I discuss below.

2.4. **Complications with Mutual Information, and Mutual Dependency as one alternative.**

Based on the findings of Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), it seems that Mutual Information can provide a useful indicator of how strongly words in a sequence are associated with one another. Moreover, MI seems to be increasing in popularity as a tool used in corpus linguistics research, including Barlow’s Collocate software (Barlow 2004) and the set of seven online corpora at corpus.byu.edu, including the 450-million-word Corpus of Contemporary English (Davies 2008). Yet it is important to note that MI must be used with considerable caution. In the Natural Language Processing literature, MI is a measure with a troubled reputation, and perhaps with good cause.

In short, MI must be integrated with token frequency in order to give meaningful results. Of course, MI already includes token frequency of a sequence as part of its definition (see the f(w_1w_2) term in Equation 2.1). But MI should also incorporate additional constraints from token frequency in order to avoid some troublesome results. As noted above, the two-word sequences in Table 2.2 were retrieved by restricting the search to items that occur at least 30 times in the Switchboard corpus (10 times per million), then sorting all items by Mutual Information. This particular frequency cutoff is quite arbitrary, and the value chosen influences the results. If we choose a different

minimum threshold (say, 3.33 per million), the search retrieves an entirely different set of bigrams at the top of the list (for instance, *Fatal Attraction* (12), Julia Roberts (10), *Knots Landing* (10), JC Penney (10)).

Continuing this experiment further, we find that if no minimum frequency is imposed, a Mutual Information search yields almost worthless results, yielding many sequences that occur just once in the corpus. Without any frequency filtering, the highest-ranked MI items includes rare words (or misspelled nonwords) that happened to be juxtaposed in this corpus just once (*grooves slaps, terming emerging, automa tic*). It is, however, true that some of the top-ranked sequences are, by chance, units of some kind (*Davy Crockett, varicose veins, topsy turvy*). Even in these cases, MI should be interpreted cautiously, because the measure overestimates the degree of word association. For instance, in the present example, MI indicates that *Davy* and *Crockett* are perfectly dependent, although searching a larger corpus would reveal that each of these words has additional uses.

Manning and Schütze (1999: 181) observe that Mutual Information is, among collocational measures, especially sensitive to problems of 'data sparseness,' that is, the limits imposed by rare occurrences in small corpora. We may partially mitigate such problems by using larger corpora, or by filtering out low-frequency sequences altogether. Evert and Krenn (2001) find that Mutual Information retrieves many useless word sequences for low-frequency items, but performs much better in the upper ranges of frequency. This finding helps to account for the intuitive sequences in Table 2.2, and moreover, helps to account for the successful experimental results in Ellis et al. (2008) and Ellis and Simpson-Vlach (2009). It seems that Ellis et al. may have avoided

difficulties with Mutual Information by winnowing out low-frequency sequences from the set of possible candidates. Frequency counts are not available for individual stimuli in the Ellis et al. studies, but their frequency categories are designed with the following values: the low frequency mean is 10.9 per million; the medium frequency mean is 15.0 per million, and the high frequency mean is 43.6 per million (2008: 380-381). It is worth noting that even their 'low frequency' category threshold is rather high in frequency. For comparison, in the experiment of Tremblay et al. (2007), a low-frequency (NLB) sequence has a frequency that is less than 10 per million for 4-word sequences, or less than 5 per million for a 5-word sequence. Thus, avoiding low-frequency sequences in this way could help circumvent some of the problems that are known to plague Mutual Information.

However, we should consider further theoretical and practical concerns regarding MI as a measure. Manning and Schütze (1999) further observe that Mutual Information systematically ranks items in a counterintuitive way. Once again, the measure has a bias that is subject to undue influence from low-frequency events. Consider a hypothetical example that draws out a point from Manning and Schütze (1999: 181). Suppose there are two different sequences that contain perfectly-dependent word pairs (that is, words that always appear together): *ipso facto* and *scantily clad*. (Neither of these word pairs is perfectly dependent in real corpora, but idealizing the data helps to illustrate the point.)

Suppose that *ipso facto* has a frequency of 100 in a corpus, and *scantily clad* has a frequency of 200. (Indeed, these numbers are comparable to real values in the 450-million word COCA Corpus). Then the MI for *ipso facto* would be given by (for now,

dispensing with the corpus size and log-transformation to focus on proportional relationships):

$$\text{(Ex. 1)}\quad \text{MI}(\textit{ipso facto}) \sim \frac{f(\textit{ipso facto})}{f(\textit{ipso})*f(\textit{facto})} = \frac{100}{100 * 100} = \frac{1}{100} = 0.01$$

Compare this to the more frequent sequence *scantily clad*:

$$\text{(Ex. 2)}\quad \text{MI}(\textit{scantily clad}) \sim \frac{f(\textit{scantily clad})}{f(\textit{scantily})*f(\textit{clad})} = \frac{200}{200 * 200} = \frac{1}{200} = 0.005$$

Thus, although the two sequences have the same amount of dependence (that is, perfect), MI would tell us that the more frequent one should be ranked far lower! That makes little sense; we would like the measure to at least rank the two sequences the same. This strange result is not just a borderline case, either; comparable problems arise when there is less than perfect dependence (Manning and Schütze 1999).

One sensible countermeasure would be to multiply MI by the frequency of the word sequence, that is, to provide an additional contribution from the frequency of the multiword sequence. That is, we may define a modified Mutual Information score for bigrams as in Equation 2.3.

$$\text{(Equation 2.3)}\quad \text{MD}(w_1w_2) = \log_2 \left[\frac{f(w_1w_2)^2}{f(w_1)*f(w_2)} \right]$$

In the previous example, this modified measure would provide the same score to *ipso facto* and *scantily clad*, reflecting the fact that the bigrams exhibit the same amount of dependence.

I have labeled the quantity in Equation 2.3 as 'MD' to stand for 'Mutual Dependency,' a term coined by Thanapoulos et al.(2002). A number of researchers (Fontenelle et al. 1994, Thanapoulos et al. 2002, Bouma 2009) have independently suggested taking this type of approach, which compensates 'for the bias of the original [Mutual Information] definition in favor of low-frequency events' (Manning and Schütze 1999: 182). Bouma (2009) observes that additionally, a measure such as Equation 2.3 provides a normalized variant of Mutual Information. That is, Mutual Information as defined in Equation 2.1 is an unbounded quantity, but (the argument inside the logarithm of) Mutual Dependency as defined in Equation 2.3 is a probability between 0 and 1, and the measure thus has a more straightforward, probabilistic interpretation.

Taking this notion of normalization a step further, we can extend the definition of Mutual Dependency to allow for sequences longer than two words. More generally, we would raises the frequency in the numerator to the power of *n*, where *n* is the number of words in the sequence. Combining the principles of Equations 2.2 and 2.3, we generalize the definition of Mutual Dependency as in Equation 2.4.

$$\text{(Equation 2.4) } MD(w_1w_2w_3...w_n) = \log_2\left[\frac{N^{n-1} * f(w_1w_2w_3...w_n)^n}{(f(w_1)*f(w_2)*...f(w_n)}\right]$$

As in Equation 2.3, we also include an appropriate contribution from the corpus size, N, which produces easier-to-read, positive MD values.

Mutual Dependency (MD) is still in the general family of Mutual Information-type measures[12], which represent in a direct way a tension between the total frequency of a sequence, and the frequency of the component parts. In fact, the top-ranking results of a

[12] Indeed, Manning and Schütze (1999) really consider it to be a slightly different Mutual Information measure. I use the distinct name coined by Thanopoulos et al. (2002) largely for ease of reference.

search according to Mutual Dependency can be strikingly similar to those using Mutual Information. For illustration, Table 2.3 presents the twenty two-word sequences with highest Mutual Dependency in the Switchboard Corpus.

1. *et cetera* (58)	11. *you know* (34487)
2. *Los Angeles* (55)	12. *death penalty* (180)
3. *capital punishment* (264)	13. *Soviet Union* (76)
4. *United States* (248)	14. *North Carolina* (137)
5. *Star Trek* (57)	15. *credit cards* (274)
6. *Peace Corps* (88)	16. *little bit* (1473)
7. *Rhode Island* (46)	17. *credit card* (275)
8. *Saint Louis* (35)	18. *New York* (356)
9. *San Francisco* (91)	19. *Social Security* (99)
10. *San Antonio* (91)	20. *per se* (39)

Table 2.3: Twenty Switchboard bigrams with the highest Mutual Dependency (minimum token frequency of 10 per million). Token frequencies are in parentheses.

There is a noticeable amount of overlap between the Mutual Dependency items in Table 2.3 and the Mutual Information items in Table 2.2. Perhaps the most telling difference between the lists involves the presence of two high-frequency items in Table 2.3: *you know* and *little bit*. Note that *you know* is an especially striking item; this sequence could never be a top-ranked item using Mutual Information as a measure, due to the very high-frequency component words *you* and *know*.

Indeed, the most important differences between MD and MI are due to the effects of ultra-high-frequency component words. For purposes of illustration, let us go beyond the top-ranked items, and consider examples of how particular word sequences are ranked by each measure. For instance, consider the high-frequency sequence *have to*, which would seem to be a good candidate for a prefab in current English (as attested by

its reduction to *hafta* in casual conversation). Following Equation 2.2, *have to* receives a Mutual Information score as follows, using counts from COCA[13] (Davies 2008):

$$\text{(Ex. 3) MI(}\textit{have to}\text{)} = \log_2 \left[\frac{464{,}020{,}256 * 241{,}484}{2{,}097{,}432 * 11{,}737{,}803} \right] = 2.19$$

The sequence *have to* in fact has a rather low MI score. Compared with all two-word sequences in COCA with the pattern *have* _____, *have to* is ranked #1397 for Mutual Information. Thus, as a random contrastive example, *have to* has a far lower MI score than *have coped*[14]:

$$\text{(Ex. 4) MI(}\textit{have coped}\text{)} = \log_2 \left[\frac{464{,}020{,}256 * 37}{2{,}097{,}432 * 340} \right] = 4.59$$

This ranking certainly runs counter to our sense that *have to* is a multiword unit of some kind in English. The low MI score for *have to* arises because the words *have* and *to* are of very high frequency. In the following section, I argue more generally that quantitative measures should allow for cases in which prefabs contain highly frequent words. For now, let us take the foregoing example as an intuitive indicator that Mutual Information seems to impose excessive 'penalties' on sequences that contain high-frequency words.

A number of variations could be used to overcome this particular limitation, but again let us consider Mutual Dependency as one alternative. Mutual Dependency

---

[13] In this equation: 464,020,256 is the corpus size for COCA; 241,484 is the frequency of *have to*; 2,097,432 is the frequency of *have*, and 11,737,803 is the frequency of *to*.

[14] Readers might object that the low MI value for *have to* is based on inflated counts for *to*, since this word has prepositional uses in addition to the infinitival use apparent in *have to VERB*. However, restricting the frequency of *to* to infinitival uses yields a modified COCA MI score for *have to* of 2.89. This still ranks *have to* far below *have coped*, along with hundreds of other bigrams (for instance: *have opposable* (MI = 4.60), *have preliminarily* (MI = 5.44), or *have sinned* (MI = 6.48).

provides an additional boost to multiword sequences which are more frequent, which helps to counteract the effects of high-frequency terms in the denominator. Following Equation 2.4, we find that Mutual Dependency ranks *have to* and *have coped* in a more intuitive way.

$$\text{(Ex. 5) MD(}\textit{have to}\text{)} = \log_2 \left[\frac{464{,}020{,}256 * (241{,}484)^2}{2097432 * 11737803} \right] = 20.07$$

$$\text{(Ex. 6) MD(}\textit{have coped}\text{)} = \log_2 \left[\frac{464{,}020{,}256 * (37)^2}{2{,}097{,}432 * 340} \right] = 9.80$$

In reviewing Examples (6) and (7), keep in mind that Mutual Dependency scores should only be compared with other Mutual Dependency scores (never with Mutual Information scores). In any case, we find that *have to* receives a higher Mutual Dependency score than *have coped*, as expected from intuition.

A review of the NLP literature on collocation extraction hints that this particular example is indicative of a larger pattern. In a systematic study of bigrams in WordNet and a database of 'named entities' in a journalistic database, Mutual Dependency represents a considerable improvement over Mutual Information (Thanapoulos et al. 2002). A wide array of collocation evaluation metrics are available (Evert and Krenn 2001, Manning and Schütze 1999), but Mutual Dependency offers one rather straightforward quantitative representation of relative frequency, while also addressing some of the shortcomings of Mutual Information. Mutual Dependency will be central to the experiment design in Chapter 4 in this dissertation.

2.5. **The need for absolute frequency alongside relative frequency.**

Above, I have reviewed experimental evidence that both token frequency and relative frequency have an effect on the representation of multiword sequences. The view I will pursue in this dissertation (following Bybee 2010) is that we should include both types of measures in models of usage. Although there have been some contradictory results, thus far there is no convincing evidence that we should ignore either absolute frequency or relative frequency effects in morphosyntax.

The strongest statements to the contrary come from Hay, who claims to have 'demonstrated that relative frequency matters more than absolute frequency' (2001: 1066). It is certainly true that the results in Hay (2001) provide evidence in support of relative frequency. Consider Hay's metalinguistic task, in which subjects decide which of two words is more complex. Around 65% of the time, subjects describe the word that is high in relative frequency as less complex. However, this finding on its own is not sufficient to show that absolute frequency is unimportant, and Hay (2001, 2003) does not report statistics that specifically address this point. Importantly, the high relative frequency category is matched for token frequency on average with the low relative frequency category (rather than being matched by item). Thus, approximately 50% of the high relative frequency stimuli are also higher in token frequency than their opponent words. In some pairings there is a quite large difference in token frequency (compare the high relative frequency word *impatient*, which has a CELEX frequency of 227, with *imperfect*, which has a CELEX frequency of 50). It seems plausible that token frequency could influence subjects' judgments, and further investigation is needed to determine

whether high token frequency plays a role in the 35% of responses that in fact favor the low relative frequency item.

With respect to semantic opacity, it is not clear that Hay uses consistent criteria when comparing absolute frequency with relative frequency. In Hay's analysis of relative frequency, the 'high' frequency category consists of derived words that are more frequent than their bases. This is a rather elite group of derived words; for prefixes, it represents the top 20.8% of words, and for suffixes, it is the top 14.8% (values here are calculated from Tables 5 and 6, Hay 2001: 1053-4). However, for absolute frequency, Hay defines 'high' frequency as above 'average', that is, in the top 50%. Such an uneven choice for high-frequency thresholds hardly seems fair, since special retrieval mechanisms may be apparent only for items that are highest in frequency (see Alegre and Gordon 1999). Moreover, Timm (2012) reanalyzes Hay's (2001, 2003) data, and finds that relative frequency actually accounts for a very small percentage of items which are semantically opaque, leaving the phenomenon essentially unexplained.

In fact, there are many reasons to believe that a relative frequency account would, on its own, be insufficient for describing patterns of processing and change for multiword sequences. First of all, relative frequency measures cannot account for many cases in which a multiword unit is known to have developed. In other words, high relative frequency is not a necessary condition for the development of multiword units.[15] Relative frequency accounts would predict that complex units should be unlikely to form when component words are high in frequency, since wholes and parts are said to compete.

[15] The discussion of statistical matters above already demonstrated that high relative frequency is not sufficient for the formation of multiword units. Relative frequency measures must be used carefully, because left unchecked they can retrieve word sequences of little interest.

In a system driven solely by influences from relative frequency, we would expect to see multiword units typically arising out of low-frequency components. However, this is clearly not the case. For instance, Bybee (2010:47) shows that the English sequence *have to* has developed into a separate unit with a meaning of obligation, even though the verb *have* is extremely frequent (around 10 times as frequent as *have to*). Similarly, the English sequence *going to* has developed a future meaning, arising out of a context which was – relatively speaking – quite rare. In Shakespeare's comedies, *go* appears in a purpose clause only 10% of the time, but still developed into a future marker (Bybee 2006). Moreover, even though *go* is highly frequent as a verb of intransitive motion, it has developed a whole range of other distinct uses in prefabs, constructions, and idioms: *go ahead and* VERB, *go* + VERB, *go it alone, go to hell, how goes it, go with one's instinct*.

In fact, what we find is that new grammatical units (including some elements that are multiword sequences) generally emerge out of highly frequent components. These component words are used in a wide range of contexts by virtue of their semantic generality, and in some of these contexts they develop new, particular meanings (Goldberg 2006, Bybee and Torres Cacoullos 2009, Bybee 2010). It is possible for such changes to occur because lexical categories are not fixed, monolithic entities. Rather, as a result of usage, an item may split off from its erstwhile category and become autonomous in a particular construction (Bybee and Brewer 1980, Bybee 2003). Autonomy of an item may be evident in new morphosyntactic patterns or in new semantic extensions. For instance, English speakers may say things like *The tree is going to lose its leaves*, or *I'm going to go there* – statements that would be nonsensical if all uses of the word *go*

represented the same lexical category (Bybee 2003: 339). Yet relative frequency measures are essentially incommensurable with the fact that in particular diachronic situations, lexical categories can split. If we are analyzing the statistical attributes of a sequence such as *BE going to*, a relative frequency measure (such as Mutual Information) would forever 'penalize' the sequence due to the high token frequencies of *go(ing)* and *to*. As change proceeds, it gradually becomes less and less appropriate to classify *go* in *BE going to* as the same item as intransitive motion verb *go*. Here, token frequency is clearly the superior measure to use, because it is 'self-correcting' with respect to increasing autonomy. Items that are high in absolute frequency are more likely to be autonomous, and vice-versa, with no permanent penalty imposed because a unit happened to originate from a high-frequency item.

A point related to the previous one is that relative frequency cannot, on its own, form the foundation for a theory of language change. To see why, let us assume that multiword units arise out of the productive, compositional use of words. Initially, there is nothing especially fixed about the words used; instead of *in spite of*, for instance, one could just as well say *in defiance of*, or *with spite toward*. This means that the sequence has a relative frequency very close to zero, and high relative frequency thus cannot provide any motivation for change. When relative frequency is high (such as might occur when a complex form is more frequent than its component parts), this represents a rather advanced stage in the formation of a multiword sequence, and we should not be surprised if such sequences have special mental representations. High relative frequency is a sign that a change has already occurred—not the impetus for the change itself. This means

that the mind must track other factors besides relative frequency, and these factors are important in the development of multiword or multimorphemic units.

One such factor would need to be token frequency, and we know that token frequency information is in fact retained regarding multiword sequences. If relative frequency is important, then it immediately follows that token frequency is important, because relative frequency depends on token frequency values. Any model of relative frequencies—whether represented mathematically as in Equations 2.1-2.4, as the deciding factor in a dual route model as in Hay (2001, 2003), or as exemplars of varying strengths—already presumes some mental representation for token frequency. Although Hay (2001) argues that absolute frequency is not independent of relative frequency, this criticism cuts both ways; indeed, one could make the argument that absolute frequency is more important than relative frequency, because the latter depends by definition on the former.

In sum, it seems we need not pit token frequency and relative frequency against one another as theoretical adversaries (for another expression of this view, see Krug 2003). There is no reason to assume that our minds track only one statistical measure, and indeed, it seems we track multiple patterns in language simultaneously (Klein and Yu 2009). In various domains, experimental evidence shows that processing of input is influenced by multiple factors at once—for instance, similarity of items to previously encountered items, and frequency of those items (Nosofsky 1988). A number of studies (including Saffran et al. 1996, Saffran and Wilson 2003, Marcus et al. 1999, Perruchet and Desaulty 2008, Pelucchi, Hay and Saffran 2009) would indicate that the mind tracks a variety of statistical patterns, simultaneously and unconsciously. Language processing

and language change are likely to emerge out of an assortment of mechanisms that interact (Hopper 1987, Beckner et al. 2009, Beckner and Bybee 2009). It is reasonable to expect that the mind is capable of tracking both relative frequency and token frequency patterns, and that such factors make independent contributions to the formation of units.

2.6. **Toward an integrated model.**

In this dissertation, following Bybee (2010:46-7), I propose that both relative frequency and absolute frequency patterns are tracked in cognition, and are associated with language change. With respect to the formation of multiword units, more than one mechanism may lead to holistic retrieval. As a multiword sequence becomes more frequent in comparison to its components, then relative frequency increases, and holistic retrieval becomes more likely. Alternately, high token frequency may also independently encourage the holistic retrieval of a multiword sequence. Bybee (2010: 46) concedes that Hay (2001) is correct that relative frequency is important, but also suggests 'that at extremely high token frequencies, loss of analyzability and transparency will occur independently of relative frequency.' It is reasonable to believe that if a complex unit is frequent enough, it will tend to be retrieved holistically on the basis of its strong representation in memory—no matter how frequent the component parts are.

Given the foregoing, the remaining chapters in this dissertation will seek evidence in support of <u>both</u> token frequency and relative frequency. In Chapter 3, I seek to rectify a null experimental result for token frequency, namely, the verbatim memory investigation of Schmitt et al. (2004). In Chapter 4, I present a study of syntagmatic speech errors, which is more ambitious insofar as there are controls for both token frequency and relative frequency (specifically, Mutual Dependency) as independent variables.

CHAPTER 3. PREFABS AND VERBATIM MEMORY: A DICTATION METHODOLOGY RECONSIDERED

3.0. Introduction to the dictation methodology.

In Chapter 2, I alluded to a speech dictation task performed by Schmitt, Grandage and Adolphs (2004), which failed to find a significant effect of token frequency on participants' memory for multiword sequences. Schmitt et al. argue that their study 'suggests that corpus data on its own is a poor indicator of whether [multi-word] clusters are actually stored in the mind as wholes' (2004: 147).[16] In the present chapter, I will reexamine the Schmitt et al. dictation methodology as a source of evidence, based on existing data as well as newly-gathered data. In the remainder of this section, I describe the rationale behind using verbatim memory to provide insights into multiword units. In Section 3.1, I provide a critique of the dictation study as implemented by Schmitt et al. (2004), and I reanalyze the existing Schmitt et al. data in light of various theoretical considerations. Contrary to the researchers' claims, their results provide some evidence that token frequency has an influence on performance in the dictation task. In Section 3.2, I follow up on these critiques by describing my own dictation experiment, which provides further evidence that token frequency does indeed play a role in the creation of prefabs.

Schmitt et al. (2004: 130) proposed a verbatim memory dictation task, initially inspired by measures used in the field of second language assessment. The use of dictation measures offers a potentially rich source of data regarding prefabricated units, since speakers' memory for the verbatim content of text is found to be ephemeral in certain methodologies (Sachs 1967, though see also Gurevich, Johnson, and Goldberg

[16] As discussed in Chapter 1, I would prefer to reframe Schmitt et al's (2004) question here as involving whether certain multiword clusters are retrieved from memory as wholes.

2010). The contexts in which verbatim memory breaks down—and the contexts in which it tends to be preserved — could reveal patterns about the processing units in language. Bolinger (1976) writes that prefabs 'have the magical property of persisting even when we knock some of them apart and put them together in unpredictable ways' (2). Along these lines, we might think of the verbatim memory task as 'knocking apart' language in an experimental setting, to see which pieces remain standing. The basic methodology is to (i.) have participants memorize a stretch of words in sequence, containing target sequences of interest, then (ii.) disrupt the memory of the memorized text with a distractor task, and finally (iii.) ask participants to reconstruct the original text, looking for regularities in which target sequences tend to be reproduced intact. Schmitt et al. predict that frequently-encountered 'recurrent clusters' of words should be retrieved and produced quite readily as wholes, and thus the participants' responses are expected to contain all (and only) the words of the original stimulus.

In their experiment design, Schmitt et al. selected candidate stimuli from a variety of grammars and reference guides, including the list of lexical bundles in Biber et al. (1999). The researchers chose 25 'recurrent clusters, varying from relatively frequent to relatively infrequent' (2004: 129). These target sequences varied in length between two words (*go away*; *you know*) to six words (*to make a long story short*, *I don't know what to do*). A full listing of the stimulus sequences appears in Table 3.1, below in Section 3.1.

It is worth reiterating that <u>all</u> of the sequences used in the experiment were, in the researchers' estimation, 'recurrent,' and thus there are no matched low-frequency sequences for comparison. As the experimenters point out, there is indeed a range of frequencies represented among the stimuli; the most frequent sequence, *you know*,

appears more than 42,000 times in the British National Corpus, but the least frequent sequence, *to make a long story short*, appears only twice. However, large differences in frequency are to be expected across any set of n-grams if the number of words (n) is not held constant. By way of illustration, we can consider the most frequent n-grams of varying lengths in the Switchboard corpus. The most frequent two-word sequence (*you know*) occurs more than 34,000 times. The most frequent six-word sequence (*it was nice talking to you*, reflecting the rather proscribed telephone context for this corpus) occurs only 85 times. Despite the vast differences in corpus frequencies, it would be ill-advised to say *you know* represents 'high frequency' and *it was nice talking to you* represents 'low frequency,' since they each represent the highest frequencies of their respective n-gram types. The conflation of n-gram length with n-gram frequency thus raises certain concerns about the Schmitt et al. (2004) experiment design.

Schmitt et al. embedded the 25 target sequences into sentence contexts; each 'burst' to be memorized was between 20 and 24 words long. The sentences were constructed so as to fit into a narrative (a story about picking up a garrulous hitchhiker), consisting of the 25 bursts to be tested, plus an additional 14 bursts included for story continuity. In the experiment, participants heard a sentence burst to commit to memory, and immediately afterward were presented visually with two numbers to be added together. The distractor math task was intended to overload cognitive resources, so that subjects' responses would not simply be based on a recitation from memory. Participants in the experiment thus needed to provide a spoken answer to the math question first (e.g., 52 + 29 =?), after which they attempted to repeat aloud the original stimulus sentence word-for-word. Schmitt et al. (2004: 130, 132) reason that the intervening distractor task

forces participants to rely on linguistic knowledge (including formulaic sequences) to reconstruct the original sentence, rather than merely relying on working memory.

Schmitt et al. (2004) gathered data from 30 native English speakers at the University of Nottingham.[17] The researchers coded responses as belonging to one of the following three categories: 'produced correctly'; 'partially incorrect'; and 'not produced.' In the 'partially incorrect' group, Schmitt et al. included any response which was spoken with a discontinuous or disfluent intonation contour. The researchers do not comment on the 'not produced' category, but it presumably includes responses which are insufficient (such as those in which a participant cannot remember most of the sentence), in addition to those in which the target sequence is replaced by an altogether different (set of) word(s). The three coding categories were used to compute a composite performance score for each item, as follows: correct responses were assigned 2 points; partially incorrect/disfluent responses were assigned 1 point, and responses in which the target sequence was fully absent received 0 points (Schmitt et al. 2004: 134). Averaging across the 30 participants yielded a mean performance score for each item (ranging between 0 and 2).

Based on the foregoing coding conventions, Schmitt et al. conclude that, contrary to expectations, there is 'no reliable relationship' between frequency of occurrence and mean performance in the dictation task (2004: 139). A Pearson correlation test between target sequence frequency in the British National Corpus and mean performance is not significant ($p = 0.315$). Similarly, the correlation is not significant if target sequence frequencies are drawn from the CANCODE corpus ($p = 0.961$). Schmitt et al. (2004)

[17] The study also included a comparison with 45 second-language English speakers, although that comparison is not of immediate interest here given the null findings for native speakers.

acknowledge certain limitations to their study, but argue that their 'methodology has successfully questioned whether recurrent clusters are holistically stored' (146).

Despite the conclusions reached by Schmitt et al. (2004), I argue in Section 3.1 that caution is appropriate in interpreting these apparent null findings.

3.1. Critique and reanalysis of Schmitt et al. (2004).

In the Schmitt et al. data, it is true that participants were strikingly inaccurate at recalling several of the recurrent sequences. For instance, only three (10%) of the subjects were able to reproduce *I see what you* accurately, most often giving a 'partially incorrect' replacement (25 out of the 30 participants). Likewise, only three of the subjects accurately repeated *in the same way as*, more often giving a partially incorrect response (11 subjects), or, even more often, omitting the sequence (16 subjects). Other especially low-scoring items were *as shown in figure* (3 correct responses out of 30), and *aim of this study* (2 correct responses out of 30). The full tally of errors coded by Schmitt et al. is given below in Table 3.1.

Several points are in order regarding lapses in verbatim memory in the experiment. First, we must be careful to look for frequency-based differences in memory performance, rather than comparing against what we imagine is a reasonable threshold for memory accuracy. As noted in Section 3.0, there are potential pitfalls in looking for statistical differences among the Schmitt et al. stimuli, since all the n-grams are recurrent, and items are not matched on the basis of frequency and word length. Nevertheless, we will see below that some meaningful statistical differences are still observable.

Moreover, it is worth examining the contexts in which the recurrent clusters appeared in the experimental materials. The Schmitt et al (2004) experiment embedded

target sequences in a long narrative, in an attempt to situate these items into a naturalistic context. While the reasoning behind this feature is understandable, the effects of the longer narrative structure were perhaps other than what was intended. In fact, a central topic of the narrative was the rambling nature of a hitchhiker's speech, and the hitchhiker in the story exhibits sudden shifts in topic and register. Several of the target sequences chosen are typically limited to written, academic contexts, but are used in the midst of an otherwise nonacademic narrative. Two example 'bursts' from the experiment are given in (1) and (2).

(1) *'Would you pay that? Look. This one, as shown in Figure 1 opposite.' I glanced over at the page he was holding up*.

(2) *'It says the aim of this study was to test human endurance.' The hitchhiker was testing mine as he jumped from topic to topic*.

As noted above, the responses for these items tended not to be classified as 'produced correctly.' One concern may be that participants could find it difficult to reproduce academic sequences fluently amid a more casual conversational context.

More importantly, some of the target sequences in the experiment were used in such awkward contexts that rewording would actually be encouraged, especially given the time pressure imposed. We cannot assume that speakers will recall multiword sequences verbatim regardless of the context in which those sequences are encountered, and the particular ways in which participants' responses deviate from the target may be revealing. Consider the three examples below.

(3) *I see what you would want a dam for though, so maybe they could just build a smaller one in its place*.

(4) *He started looking through my Cosmopolitan magazine and said, 'It's not too bad, this one, although I don't usually read women's magazines, you understand*.

(5) *I didn't answer, letting his voice drift over me in the same way as the snow drifted over the hills in the distance*.

In example (3), *I see what you* appears in the context *I see what you would want a dam for*, rather than the conversational contexts that actually make this sequence rather frequent (i.e., *I see what you mean*). Presumably, in this case, participants often substituted the more natural-sounding variant *I see why you would want a dam*.[18] Similarly, in example (4), it would be understandable for participants to collapse *It's not too bad, this one* into *This one's not too bad*. Moreover, with respect to frequent omissions of *in the same way as* for sentence (5), it seems likely that subjects would often substitute the more economical variant *like*. The availability of single-word substitutions is, in fact, commonly used as a diagnostic for syntactic constituency (Quirk and Mulholland 1964, Fabb 2012). Such a substitution is hardly evidence that the sequence is non-formulaic, even though Schmitt et al. would code it as counter-evidence.

Thus, a further critique of the Schmitt et al. study is that it is not attentive to the subtleties of subject responses. As described in Section 3.0, the coding system in their study considers exact verbatim responses to be evidence of formulaicity (assigning 2 points), and assigns 'partial credit' (1 point) for partially correct (which is to say, partially incorrect) responses. However, a high proportion of partially (in)correct responses on an item actually indicates that participants are not processing that item as a holistic unit.

[18] Indeed, 25 out of 30 subject responses (83.3%) were classified as 'partially incorrect' (Schmitt et al. 2004: 136), which would be consistent with this pattern. However, details are not available regarding participants' responses.

Indeed, the scoring system used in Schmitt et al.'s quantitative analysis is problematic, and is at odds with the researchers' own observations. They point out that 'the "Partially Incorrect" category [of responses] is probably the most telling in this study' (2004: 135). More specifically, they observe that a propensity toward 'partially incorrect' responses is most indicative of non-holistic retrieval:

> [I]f clusters were not produced intact, when the dictation task was to reproduce them exactly, this indicates that they were not readily available, which would argue against their being stored in the lexicon... [C]lusters which were attempted, but not reproduced intact, give the clearest indication that those clusters were somehow not prominent in the mind... [W]e know that the participant was producing word strings similar to the cluster, and with the same semantic content, but not actually reproducing the cluster in the dictation. (Schmitt et al. 2004: 137, emphasis added)

Given these observations, it is mystifying that Schmitt et al. chose to score responses such that partially incorrect answers received partial credit, rather than assigning a penalty. A reanalysis of the Schmitt et al. data seems to be in order, to investigate whether the quantitative results are indeed null. Although certain details of participant responses are not available (for instance, specific errors for each item), we have available the raw numbers coded into each response category (numbers 'produced correctly,' 'partially incorrect,' and 'not produced.'). Thus for the remainder of this subsection, I will present data reanalyses based on these raw numbers.

As a technical detail in the present reanalyses, I will use log-transformed frequency counts before performing statistical tests. The corpus frequencies of the target n-grams in British English are available from the British National Corpus; these are reported in Table 3.1 based on searches with BYU-BNC (Davies 2004-). In their correlation analyses, Schmitt et al. (2004) apparently relied on raw corpus frequency

counts. This can be verified by computing a Pearson correlation between the raw BNC counts and the values labeled 'Schmitt mean performance score.' This test yields a p-value of 0.321 (r = 0.098), which is quite close to the null p-value of 0.315 reported by Schmitt et al.[19] However, as discussed in Chapters 1 and 2, it is preferable to log-transform frequencies as a precursor to statistical tests, since logged values seem to more closely correspond to mental representations of frequencies. Thus, although I list raw corpus counts in Table 3.1, all reported results are based on log-transformed (base 2) frequency counts.

Based on log-transformed BNC counts, then, we can obtain Pearson correlations as follows. First, consider the number of fully correct responses as a possible indicator of holistic retrieval (see counts in the column labeled 'produced correctly' in Table 3.1). This variable (not analyzed separately in Schmitt et al. 2004) turns out not to be significant in the present reanalysis: p = 0.245, r = 0.12. However, if we examine the number of 'partially incorrect' responses, the correlation is significant (p = 0.045). The coefficient is negative, indicating that higher-frequency sequences are less likely to prompt partially-correct responses, although the correlation is in the weak-to-moderate range (r = -0.35).

Finally, it is worthwhile to recompute a 'mean performance' score based on the observation that partially correct responses should be assigned negative points in an overall assessment. I thus compute a revised composite score as follows: add 1 point for a

[19] Simply switching from raw frequencies to logged frequencies produces a small improvement in Schmitt et al.'s correlation result, although the result is still not significant. A Pearson correlation test between logged BNC frequency and mean performance score (as calculated by Schmitt et al. 2004) yields p = 0.24 , r = 0.15.

correctly produced response, and deduct 1 point for a partially incorrect response.[20] Zero points are assigned in either direction for target sequences that are 'not produced,' since this is a potentially heterogeneous category—including sequences which are fully substituted with other items (such as *in the same way as* > *like*), as well as those which are omitted for other reasons. The total number of points assigned is divided by the number of participants (30), resulting in average scores reported in the column labeled 'Reanalysis: Mean performance' in Table 3.1. In order to emphasize the differences from the Schmitt et al. (2004) mean performance score (and for ease of comparison with values reported in Section 3.2), the figures reported here are further transformed into values along a [-100, 100] scale. On this scale, an item which is always recalled fully accurately would receive a score of 100; an item which is always recalled in a partially incorrect way would receive a score of -100. A Pearson correlation between (log) BNC frequency and the reanalyzed mean performance falls short of significance: $p = 0.069$, $r = 0.30$.

One concern about the foregoing analyses might be that the frequency values are based on a predominantly written corpus; the British National Corpus consists of only 10% spoken English. As such, it is possible that the corpus fails to accurately represent the frequencies of sequences that are actually familiar to speakers of British English in conversation. Consider, for instance, the sequence *go away*, which occurs with a frequency of 12.44 per million words in the full BNC, but 35 per million words in the

[20] Clearly, an alternate approach would be to assign 2 points for each correctly produced response, 1 point for items which are 'not produced' in the response, and 0 points for 'partially incorrect' responses. For purposes of statistical analysis, such an approach would indeed be mathematically equivalent to the measure described above, and would more closely parallel the [0, 2] scale used by Schmitt et al. However, I prefer to use a [-1, 1] scale, since it more intuitively represents the fact that 'partially incorrect' responses should incur a penalty on the overall score for an item.

target cluster	BNC Frequency	Produced correctly	Partially incorrect	Not produced	Schmitt et al.: Mean performance	Reanalysis: Mean performance
to make a long story short	2	23	3	4	1.633	66.67
I don't know what to do	87	27	2	1	1.867	83.33
to give you an example	11	8	10	12	0.867	-6.67
as a matter of fact	377	21	4	5	1.533	56.67
from the point of view	520	19	5	6	1.433	46.67
in the same way as	657	3	11	16	0.567	-26.67
is one of the most	660	27	2	1	1.867	83.33
in the middle of the	1513	17	2	11	1.200	50.00
aim of this study	56	2	16	12	0.667	-46.67
it's not too bad	58	16	11	3	1.433	16.67
I see what you	105	3	25	2	1.033	-73.33
you've got to have	191	16	10	4	1.400	20.00
as shown in figure	191	3	17	10	0.767	-46.67
what I want to	270	21	6	3	1.600	50.00
it was going to	374	21	6	3	1.600	50.00
as a consequence of	427	13	6	11	1.067	23.33
in a variety of	732	15	11	4	1.367	13.33
in the number of	1019	18	9	3	1.500	30.00
in addition to the	1191	18	10	2	1.533	26.67
night and day	109	16	1	13	1.100	50.00
on and off	468	25	0	5	1.667	83.33
something like that	1245	16	5	9	1.233	36.67
go away	1244	28	0	2	1.867	93.33
for example	23531	18	0	12	1.200	60.00
you know	42317	24	0	6	1.600	80.00

TABLE 3.1. Listing of data for the 26 multiword sequence stimuli used in Schmitt et al. (2004).

spoken portion. Similarly, *you know* occurs only 423 times for every million words in the full BNC, but 30,814 per million words in the spoken portion. It seems that the full BNC may under-represent the frequencies of certain conversational sequences, which participants in the experiment indeed remembered quite accurately (for instance, 80% full accuracy on *you know*, and 93% full accuracy on *go away*).

Thus further analysis based on spoken data would seem to be appropriate. Schmitt et al. (2004) report a second analysis (also null) using spoken English frequencies from CANCODE. The CANCODE corpus is not publicly available, and I thus present a second correlation analysis using log-transformed frequencies from the spoken portion (10 million words) of the British National Corpus (Davies 2004-). The spoken BNC frequency counts for the Schmitt et al. stimuli are listed in Appendix 3.1. Here, the Pearson correlation tests are somewhat more successful than those based on the mostly-written BNC. First, the correlation between (log) spoken BNC frequency and the number of fully correct responses is positive and significant, with $p = 0.02$ and $r = 0.41$. The correlation between (log) spoken BNC frequency and the number of partially incorrect responses is significant, and as expected, negative, with $p = 0.017$ and $r = -0.43$. Finally, the revised mean performance score (assigning a point for fully correct responses, and penalizing a point for partially correct responses) also yields a significant result. The correlation between (log) spoken frequency and the mean performance score is positive ($r = 0.44$) and significant ($p = 0.014$).

In sum, the reanalyses based on log-transformed spoken frequencies yield significant Pearson correlations, albeit correlations that are in the weak-to-moderate range. However, these correlations are somewhat improved if our analyses exclude one particularly questionable stimulus from the experiment. As noted above, the context for the sequence *I see what you* was rather anomalous in the experiment (*I see what you would want a dam for though*). Visual inspection of the scatterplot between (log) spoken frequency and performance measures in the Schmitt et al. data indicates that this item is an outlier. Outlier status is confirmed by examining the Pearson residuals for analyses of

'produced correctly,' 'partially incorrect,' and revised mean performance values. In all three cases, *I see what you* is the only item with residuals that are two standard deviations from the mean, and exclusion of this item is thus justified. Based on the 24 remaining target sequences, the Pearson correlation between (log) spoken BNC frequency and fully correct responses is again positive and significant ($r = 0.46$, $p = 0.01$). The correlation between (log) spoken BNC frequency and partially incorrect responses is moderate and again negative ($r = -0.57$, $p = 0.002$). Finally, the correlation for the modified mean performance score is positive and moderate, with $r = 0.53$ and $p= 0.004$.

It seems, then, that Schmitt et al. (2004) may have been premature in concluding that their experiment provided no evidence of a frequency effect on verbatim memory performance. Higher-frequency sequences were in fact more prone to be recalled accurately, and, more importantly, were less prone to be produced in a partial or disfluent fashion. Moreover, when these indicators are combined into a summary statistic (a revised mean performance score), higher-frequency sequences correlate with higher overall performance. The Pearson correlations are significant, and moderately strong, if the analyses are based on (logged) spoken frequencies, and if we eliminate one particularly troublesome item from the analysis set.

All the same, the reanalyzed data from Schmitt et al. (2004) may remain open to certain criticisms. Several of the target sequences are problematic in the experiment narrative, and reduced accuracy is open to interpretation without having more details about participant responses. Moreover, in the existing data, there are no particular controls for frequency (since all the multiword sequences considered were 'recurrent'), and target sequences are not controlled for the number of words. It is possible to attempt

a post hoc control for n-gram length by separately analyzing subsets of the Schmitt et al. sequences, but this requires considering rather small numbers of data points, and the results are mixed.[21] The verbatim memory methodology is thus in need of further investigation, and I describe the methods and results of a revised experiment in Section 3.2.

3.2. Verbatim dictation revisited: A new experiment

An updated experimental study is described here, which takes into account the various design concerns noted above regarding Schmitt et al. (2004). The present approach makes comparisons between matched multiword sequences in high-frequency and low-frequency categories, and incorporates the various assessments (number of fully correct responses, number of partially correct responses, and revised mean performance) used in the reanalyses of Section 3.1.

3.2.1. **Selection of stimulus sequences**.

The revised experiment is based on a set of 26 target multiword sequences, divided into a set of 13 high-frequency sequences, and a matched set of 13 low-frequency sequences. The target sequences range from 2-5 words, and are matched across categories for number of words. The sequences in each category are also matched for part of speech throughout, which ensures that in a traditional syntactic analysis, the sequences will have

[21] For instance, Schmitt et al.'s set of 25 stimuli includes 10 target sequences that are 4 words long (assuming we exclude the problematic *I see what you*). A separate analysis of this set of 10 items indicates that measures follow the predicted patterns, and the correlations with log frequency are highly significant. For instance, for the revised mean performance score, $r = 0.91$, $p < 0.001$. However, this pattern is not borne out among the (albeit smaller) set of 6 target sequences of length 5. Here, the correlation between log spoken frequency and revised mean performance yields $r = 0.30$ and $p = 0.28$, and the other measures are also not significant.

a similar constituent status. Thus, the high-frequency sequence *in the same way as* (P Det Adj N P) is matched with the low-frequency sequence *to the same time as*.

More specifically, whenever possible, matched items contain all of the same words except for one 'pivot word.' Thus, the high-frequency sequence *on the part of* and the matching low-frequency sequence *on the life of* both follow the template *on the* N *of*. The 'pivot word' approach is inspired by the design of Tremblay et al. (2007), and is intended to ensure that the most salient cross-category frequency differences involve the entire multiword sequence, rather than individual words. The purpose of this precaution is to minimize processing advantages for the high-frequency sequences solely on the basis of individual word frequencies. Thus, pivot words are chosen so as to be similar in frequency between the two categories, or alternately, so as to bias word frequency differences in favor of the low-frequency category whenever possible.

The final set of stimulus sequences, along with the relevant frequency measures, is listed in Table 3.2. Since the participants in the study are to be speakers of American English, the frequency values used are drawn from the 450-million word Corpus of Contemporary American English (COCA, Davies 2008-), of which 20% is spoken English. For each pair of matched items, it was required that the pivot word frequency for the low-frequency sequence needed to be at least 75% of that for the pivot word in the high-frequency counterpart. There is a single exception to this general requirement: the high-frequency sequence *in the middle of* contains a pivot (*middle*) that is considerably more frequent than the pivot (*style*) in the matched low-frequency sequence *in the style of*. However, this exception is mitigated if we take word classes into account. In the sequences of interest, the pivot words (*middle*, *style*) function as nouns, and in nominal

uses, *style* is somewhat more frequent than *middle* (with COCA counts of 37,224 and 35,026, respectively).

On average, the pivot words are comparable in frequency between the high-frequency and low-frequency sequences, with a slight advantage in individual-word frequency among the low-frequency sequences. Based on the frequencies of the entire multiword sequence, the cross-category differences are far more striking: on average, the high-frequency sequences are more than 9 times as common as the low-frequency sequences.

HIGH FREQUENCY SEQUENCES			**LOW FREQUENCY SEQUENCES**		
Target Sequence	**Sequence frequency**	**Pivot word frequency**	**Target Sequence**	**Sequence frequency**	**Pivot word frequency**
*in the same **way** as*	374	475731	*to the same **time** as*	3	735572
*all of a **sudden***	5989	18912	*all of a **boy***	1	73144
*as a **result** of*	11947	68844	*as the **name** of*	55	123041
*for the **sake** of*	3545	9948	*for the **child** of*	9	129974
*in the **middle** of*	18926	81514	*in the **style** of*	413	37345
*on the **part** of*	6036	224529	*on the **life** of*	248	320046
*as **soon** as*	17561	80025	*as **big** as*	2098	208034
*in **spite** of*	7049	7647	*in **fear** of*	369	49395
*in **terms** of*	35474	66617	*in **things** of*	11	255025
*on **top** of*	13298	128709	*on **half** of*	70	109457
***according** to*	96800	97380	***looking** to*	5433	132824
***back** to*	112906	569622	***through** to*	2960	431768
*out **of***	270510	11974008	*out **to***	51709	11734566
AVERAGE	**46,186**	**1,061,807**	**AVERAGE**	**4,875**	**1,103,092**

TABLE 3.2. Listing of matched stimuli in the high-frequency and low-frequency sequence categories. Pivot words are highlighted in bold. All frequency counts are based on the full COCA corpus (450 million words, Davies 2008-).

3.2.2. **Stimulus sentences and presentation**.

As noted previously, some of the target stimuli in the Schmitt et al. (2004) study were situated in awkward contexts, perhaps as a result of constraints imposed by the continuous narrative. Thus, in the present study, I abandon the narrative structure, and

instead attempt to fit each target sequence into individual, natural-sounding sentences. Appropriate contexts were identified by examining a sampling of usages from the COCA corpus, and used as the basis for constructing matched sentences. All sentences in the study were in the range of 19 – 24 words, which is similar to the lengths of bursts (20-24 words) used in the Schmitt et al. study (2004: 132). Sentences were devised with a similar structure; each sentence contained the target sequence close to the middle of the sentence, after an introductory clause. Sample sentences are provided in (6) and (7).

(6) *The two neighbors were talking over the backyard fence, but they were interrupted when* ***all of a sudden*** *their dogs started barking.*

(7) T*he current gallery exhibit seems oddly familiar to me, because the drawings are*
all of a boy *who lived in my neighborhood.*

The set of 26 stimulus sentences was randomized into a set sequence; the full set of sentences is listed (in presentation order) in Appendix 3.2, along with the math distractors used. The stimulus sentences and distractors were recorded digitally in Audacity 1.2.6, by a native speaker of American English (the author).

Participant responses were written rather than spoken, to permit simultaneous data collection as described below. The use of written responses, unfortunately, does not allow for the analysis of disfluencies in participant responses. However, an advantage of written data collection is that the slower nature of responses renders the verbatim memory task somewhat more challenging, and encourages more overall deviation from the target sentences.

3.2.3. **Participants and data collection.**

Data collection was performed in an introductory Psycholinguistics course at the University of New Mexico. The verbatim memory activity was part of the class curriculum, which included sections on the nature of retrieval from the mental lexicon, and experimental methods in psycholinguistics. Students in the course participated in the exercise voluntarily, as a precursor to writing lab reports that analyzed and discussed the group results. Participation in the classroom exercise was not mandatory (indeed, it could not be, since responses were anonymous), although participating in the task did help students better understand the analysis assignment.

The course instructor explained the verbatim memory task, and advised that everyone in the class would be given the opportunity to try out the task firsthand. Participants filled in their responses with paper and pencil. In a preliminary questionnaire, participants indicated on their answer sheets whether they were native English speakers, and whether they had a history of speech or hearing disorders. Once the task was completed, the instructor advised the class that they had a choice whether or not to include their responses in the analysis. Students who wished to have their data analyzed passed their papers forward, and those who preferred not to participate could simply keep their papers. No names were provided on the student papers, so the class instructor could not tell which students chose to participate or not participate. Approximately 70% of the enrolled class chose to participate and have their data analyzed, yielding 43 native English participants. None of these 43 participants reported a history of speech or hearing disorders.

The procedure for written responses was as follows. Subjects heard each target sentence once. As in the Schmitt et al. study, two seconds after the sentence completed, subjects heard two numbers they needed to add together. Participants were advised that they were allowed to calculate the sum on paper if they so wished, or 'in their heads.' The important feature of the addition task was to provide interference with verbal memory, which could be accomplished whether or not the two numbers were written on paper. After completing the addition task, subjects then wrote down the sentence they had previously heard, attempting to write each sentence accurately as possible on a word-for-word basis. Even though responses were written, there was considerable time pressure for subjects to write answers quickly, since only a fixed amount of time was available before the start of the next stimulus.

A short practice round (containing two sentences) familiarized participants with the task, prior to presentation of the 26 trial sentences.

3.2.4. **Results**.

3.2.4.1. **Initial assessment and removal of outliers**.

As an initial assessment of the suitability of participant responses, the response sheets for the 43 native English speakers were coded with respect to 'fully correct' responses and 'insufficient/uncodable' responses. This initial coding step is intended to exclude any participants who may not have been fully engaged in the verbatim memory task.

First, responses were coded as 'fully correct' if they contained all (and only) the words of the target sequence, in the correct order. Across the 43 subjects, the number of fully correct responses (out of 26 stimuli) ranged between 5 and 22. The average number

of fully correct responses is 12.11 (SD= 4.32) out of 26, or 46.60%. The relatively low accuracy perhaps attests to the general difficulty of recalling approximately 20 words verbatim, following a math distractor task.[22] Nevertheless, fully correct responses were the most common type of response (that is more common than any particular category of error). Moreover, no participant scored below two standard deviations in terms of the number of fully correct responses, and thus this particular measure was not used to exclude any participants from the study.

Secondly, as part of the initial assessment, participant answer sheets were examined in order to identify responses that were insufficient or uncodable for various reasons. Examples of insufficient responses would be items that were left blank, or responses that were incomplete to such an extent that it is unclear whether the participant understood the sentence. Other insufficient responses would be those in which the participant misheard a crucial part of the sentence (e.g., *in spite of* > *in light of*), or in which the participant clearly misunderstood the meaning of a sentence (for example, misassigning semantic roles). An example of such a response is provided in sentence (8b), with the original stimulus provided in (8a).

(8a) TARGET: *One of downtown's most memorable landmarks is an elaborate church, which dates* ***to the same time as*** *the famous courthouse*.

(8b) RESPONSE: *One of the town's most famous landmarks is the church whose clock is the same time as the courthouse's*.

Finally, in a few rare cases, responses had to be rejected as insufficient because the participant's handwriting was illegible.

[22] By comparison, in the Schmitt et al. (2004) study, on average the 30 native English participants were fully accurate 55.73% of the time.

Across the 43 participants, the number of insufficient responses varied between 0 and 15, with an average of 3.49 insufficient responses (out of 26) per participant (SD= 3.28). The average percentage of insufficient responses is thus 13.42%. However, it is disconcerting that some participants had up to 57.7% uncodable responses, and a decision was made to exclude any participants whose performance on this measure was two standard deviations below the mean. This criterion led to the exclusion of three participants.

Additional participant filtering was imposed on the basis of mathematical accuracy, to ensure that all participants were fully engaged in the math distractor task in addition to the verbatim memory task. On the whole, participants were quite attentive to the addition problem, since the average accuracy was 92.03% (mean incorrect responses 2.07, SD = 1.67). Almost all errors were very plausible clerical mistakes which might be made in a high-pressure situation (with an answer either off by 1, or off by 10 from the target). However, there were some isolated math responses which seemed less plausible, and a few participants showed performance that was considerably poorer than the average. Analysis indicates that there were two participants whose accuracy on the distractor task was more than two standard deviations below the mean. Exclusion of these participants further reduces the total pool of participants to 38.

Once the five participants described above are excluded (three on the basis of verbatim memory, and two on the basis of the distractor task), the following summary statistics characterize the remaining 38 participants. On average, accuracy on the math distractor task is 93.42%, with a mean number of incorrect responses of 1.71 (SD = 1.21). With respect to the verbatim memory task, among the 38 remaining participants the

number of fully correct responses ranges between 6 (N=2) and 22, with an average of 12.73 (SD= 4.11), or 48.99%. The number of insufficient responses ranges among participants between 0 (N = 6) and 7 (N=3), with an average of 2.68 insufficient responses per participant (SD= 2.09), or 10.32%. All of the 38 participants remaining in the study have a larger number of fully correct responses than insufficient responses.

3.2.4.2. **Quantitative results**.

Based on the 38 participants included in the study, I first present several broad quantitative analyses, to be followed by more in-depth qualitative discussions. For the sake of simplicity in these statistical tests, we can assume a three-way distinction in response types for each target sequence: 'fully correct,' 'partially (in)correct,' 'not produced.' Responses in which the target sequence is 'not produced' are distinct from uncodable or insufficient, since 'not produced' responses constitute valid dictations that express the original meaning of the stimulus sentence, but without using any words from the target sequence. As I discuss in detail below, the 'not produced' category includes a rather heterogeneous range of responses, which may or may not provide any evidence regarding holistic retrieval of the target sequence. For our present purposes, the prefab hypothesis predicts that higher-frequency sequences should be (a) more likely to be produced in a fully correct form, and (b) less likely to be produced in a partially (in)correct form. Additionally, an appropriate composite measure, the (revised) mean performance score as discussed in Section 3.1, should tend to be higher for higher-frequency sequences.

HIGH FREQUENCY SEQUENCES				LOW FREQUENCY SEQUENCES			
Target Sequence	**Percent fully correct**	**Percent partially incorrect**	**Mean performance score**	**Target Sequence**	**Percent fully correct**	**Percent partially incorrect**	**Mean performance score**
*in the same **way** as*	*15.63	*68.75	*-53.13	*to the same **time** as*	*40.63	*59.38	*-18.75
*all of a **sudden***	40.54	27.03	13.51	*all of a **boy***	38.24	61.76	-23.53
*as a **result** of*	*28.00	16.00	*12.00	*as the **name** of*	*62.86	37.14	*25.71
*for the **sake** of*	65.71	11.43	54.29	*for the **child** of*	29.17	62.50	-33.33
*in the **middle** of*	54.05	32.43	21.62	*in the **style** of*	18.75	81.25	-62.50
*on the **part** of*	*34.48	0.00	34.48	*on the **life** of*	*47.22	30.56	16.67
*as **soon** as*	100.00	0.00	100.00	*as **big** as*	94.74	5.26	89.47
*in **spite** of*	*40.54	0.00	40.54	*in **fear** of*	*64.71	32.35	32.35
*in **terms** of*	75.76	9.09	66.67	*in **things** of*	60.53	39.47	21.05
*on **top** of*	92.11	7.89	84.21	*on **half** of*	25.00	65.63	-40.63
***according** to*	*50.00	0.00	50.00	***looking** to*	*69.70	30.30	39.39
***back** to*	83.78	13.51	70.27	***through** to*	8.82	38.24	-29.41
*out **of***	85.71	8.57	77.14	*out **to***	57.14	37.14	20.00
AVERAGE	**58.95**	**14.97**	**43.97**	**AVERAGE**	**47.50**	**44.69**	**2.81**

TABLE 3.3. Quantitative results for three measures in the verbatim memory task. Scores of the same type (for instance, percent fully correct) should be compared for each matched item. Pairs marked with asterisks are those in which the observed scores are contrary to the expected pattern (see section 3.2.4.3 for discussion).

In Table 3.3 above, I list the quantitative results for each target sequence. For all quantitative results discussed here, I report observed response types as a percentage of all codable responses, rather than as a percentage of all responses. In other words, the responses considered 'insufficient' as described in 3.2.4.1 are excluded in order to provide a more accurate assessment of the likelihood of recall the entire sequence or only a part of it. On the whole, there are fewer insufficient responses among the high-frequency sequences than among the low-frequency sequences; across all participants, there are 45 and 57 insufficient responses in the two categories, respectively. However, the distribution of insufficient responses is not significantly different between the two

categories. A one-tailed paired t-test by item yields $p = 0.29$, and by participant yields $p = 0.09$. These null results suggest that it is reasonable to exclude insufficient responses from percentage calculations without influencing the results for one or the other categories.

First, then, we can consider the distribution of fully correct responses with respect to frequency. As discussed in Section 3.2.4.1, such responses are those that contain all the words of the target sequence, without interruption or alteration. Among high-frequency sequences, 58.95% of responses are fully correct (again, as a percentage of all codable responses), compared with 47.50% of low-frequency responses that are fully correct. In a one-tailed t-test paired by item, this difference does not reach significance, with $p = 0.13$. However, in a one-tailed t-test paired within participants, the difference is highly significant, with $p < 0.001$. Likewise, a Pearson correlation between log COCA frequency and the percent of fully correct responses is significant and positive, with $p = 0.018$, although in a weak to moderate range ($r = 0.41$).

Perhaps a more telling indicator of holistic retrieval is the likelihood of 'partially incorrect' recall among target sequences of different frequencies. As observed by Schmitt et al. (2004), if a target sequence is recalled in partial form, or in altered form (such as being interrupted, rearranged, or containing one or more substituted words), this would constitute evidence that the sequence is not being recalled as a holistic unit. Conversely, a disinclination to replace or interrupt any words in a multiword sequence may be considered evidence that the sequence tends to be recalled in a unitary fashion. To assess this variable, participant responses are coded as 'partially incorrect' if the response contains at least one word from the target sequence, but one or more other words within

the target sequence is replaced, inserted, rearranged, or changed morphosyntactically.[23] Responses incorporating any of the foregoing modifications are taken as evidence that the sequence was comprehended and/or recalled as a collection of individual words, rather than as a continuous unit. Examples of such responses from the data are below. Sentence (9a) presents the original stimulus sentence; in the participant's response in (9b), an extraneous word is inserted, in (9c), a word from the target sequence is omitted; and in (9d), a word from the target sequence is replaced with a related word.

(9a) TARGET: *In the garden, we found insect damage **on half of** the plants, so we decided that we might have better luck next year*.

(9b) PARTIALLY INCORRECT RESPONSE, 1: *In the garden, we found insect damage **on over half of** the plants, so we decided that we might have better luck next year*.

(9c) PARTIALLY INCORRECT RESPONSE, 2: *We found insect damage **on half** the plants, so we thought we'd have better luck next year*.

(9d) PARTIALLY INCORRECT RESPONSE, 3: *In the garden we found insect damage **on some of** the plants so we thought we would have better luck next year*.

When coded with respect to partially incorrect responses, there are significant differences between the high-frequency and low-frequency target sequences. As a percentage of all codable responses, for high-frequency sequences, 14.97% of responses are partially incorrect, compared with 44.69% of low-frequency sequences. This difference is statistically significant, as demonstrated by one-tailed, paired t-tests, which yield $p < 0.001$ whether grouped by item or by participant. Moreover, the correlation is significant between log frequency of each sequence and the percentage of partially incorrect responses, with $r = -0.72$ and $p < 0.001$. As we would predict, the correlation is

[23] Alternate operational definitions of 'partially incorrect' responses are certainly possible. I discuss certain definitional complications in Section 3.2.4.3.

negative (and strong), indicating that in the verbatim memory task, higher-frequency sequences are less likely to be produced in a partially incorrect form.

Finally, as a broad quantitative measure, we may consider a composite score which combines the two previous measures regarding fully correct responses, and partially correct responses. As in Section 3.1, I compute a mean performance score by assigning one point for each fully correct response, and deducting one point for each partially incorrect response. For ease of comparison with the other measures discussed in this section, I calculate the score as a percentage of all codable responses. Thus, each mean score varies along the range from -100.0 (which would indicate that responses are partially incorrect 100% of the time) to +100.0 (indicating that responses are fully correct 100% of the time). This analysis yields a mean performance score for high-frequency sequences of 43.97, and for low-frequency sequences of 2.81. These differences are statistically significant: a one-tailed t-test, paired by item yields $p = 0.0043$, and paired by participant yields $p < 0.001$. Across all target sequences, the correlation between log frequency and mean performance is also significant; $p < 0.001$, and the r-value is positive and moderate ($r = 0.62$).

As an addendum to these quantitative results, I should note that certain of the participant responses presented difficulties in my coding choices. As discussed above, I coded a response as ‘partially incorrect’ if a word from the target sequence appeared in the response, but one or more other words was altered in some way. This approach led to rejecting certain responses as being ‘partially incorrect,’ although a case could be made for classifying them as such. Most notably, with respect to the high-frequency sequence *in spite of*, 20 participants used *despite* in its place. This is a striking and noteworthy

pattern, which I discuss in greater depth in Section 3.2.4.3. In the data presented in Table 3.3, I coded such responses as 'replacements' rather than 'partially incorrect.' (See Section 3.2.4.3 below for a discussion of sequence replacements.) Indeed, note that *despite* does not contain any analyzable morphemes, and thus it is reasonable to claim that its substitution for *in spite of* evinces no activation of *spite* as a separate word. However, it could also be argued that *despite* should be coded as 'partially incorrect,' since *despite* shares phonological material with *in spite of*. If one item could prime another on the basis of shared phonological material, irrespective of morphosyntactic structure, this could be considered evidence that activation of component parts outstrips activation of the whole.

To investigate whether the choice of coding convention had an impact on the results, I recoded the data in a way that is focused on general overlaps between target sequence and response, moreso than overlaps that attend to orthographic word boundaries. In the recoded system, responses were considered to be 'partially incorrect' if phonological material from the target sequence appears in a modified form in the participant's response. Thus, in addition to recoding *in spite of* > *despite*, this new coding system considers *out of* > *outside* and *through to* > *into* to constitute 'partially incorrect' attempts. Based on this recoding under a broader definition of partially incorrect responses, the revised percentages for each item are presented below in Table 3.4. (The 'fully correct' tallies are unchanged by recoding, and thus are not included in this alternate table.)

HIGH FREQUENCY SEQUENCES			LOW FREQUENCY SEQUENCES		
Target Sequence	(Recoded) percent partially incorrect	(Recoded) mean performance score	Target Sequence	(Recoded) percent partially incorrect	(Recoded) mean performance score
*in the same **way** as*	*68.75	*-53.13	*to the same **time** as*	*59.38	*-18.75
*all of a **sudden***	35.14	5.41	*all of a **boy***	61.76	-23.53
*as a **result** of*	16.00	*12.00	*as the **name** of*	37.14	*25.71
*for the **sake** of*	11.43	54.29	*for the **child** of*	62.50	-33.33
*in the **middle** of*	32.43	21.62	*in the **style** of*	81.25	-62.50
*on the **part** of*	0.00	34.48	*on the **life** of*	30.56	16.67
*as **soon** as*	0.00	100.00	*as **big** as*	5.26	89.47
*in **spite** of*	*54.05	*-13.51	*in **fear** of*	*32.35	*32.35
*in **terms** of*	9.09	66.67	*in **things** of*	39.47	21.05
*on **top** of*	7.89	84.21	*on **half** of*	65.63	-40.63
***according** to*	0.00	50.00	***looking** to*	30.30	39.39
***back** to*	13.51	70.27	***through** to*	88.24	-79.41
*out **of***	14.29	71.43	*out **to***	42.86	14.29
AVERAGE	**20.20**	**38.78**	**AVERAGE**	**48.98**	**-1.48**

TABLE 3.4. Quantitative results on the basis of recoded data, with a more expansive definition of 'partially incorrect' responses. Pairs marked with asterisks are those in which the observed scores are contrary to the expected pattern.

Statistical analysis of these revised figures indicates that the alternate coding conventions have little effect on the quantitative findings. With respect to the percentage of responses which are, more broadly defined, 'partially incorrect,' the averages for high and low frequency sequences are 20.20 and 48.98, respectively. This difference is again statistically significant. A one-tailed t-test yields $p = 0.0011$ when paired by item, and $p < 0.001$ when paired by participant. Likewise, the Pearson correlation between log spoken frequency and (revised) percentage partially correct is negative and significant, with $r = -0.59$, and $p < 0.001$. Based on the recoded data, the mean performance scores still exhibit very significant differences on the basis of frequency. The average score is 38.78 for high-frequency items, compared with -1.48 for low-frequency items. A one-tailed t-test

yields $p = 0.015$ when paired by item, and $p < 0.001$ when paired by participant. A Pearson correlation test is also significant, with $r = 0.54$ and $p = 0.002$.

In sum, although more than one coding convention is defensible regarding the identification of 'partially incorrect' responses, it seems that the quantitative results are generally not dependent on the particular convention used. Under either approach, the relationship between sequence frequency and likelihood of partially correct responses is significant and negative. Moreover, under either coding approach, there is a positive, significant relationship between sequence frequency and a mean performance score which combines measures of fully and partially accurate responses.

3.2.4.3. **Exceptions to the general pattern, and qualitative results**.

The foregoing analyses provide quantitative evidence that high frequency does, on the whole, have an effect on the nature of responses in a verbatim memory task. High-frequency sequences are generally more likely to be recalled verbatim, and are less likely to be produced in an interrupted or altered form. Yet it is also certainly true that high token frequency alone does not unfailingly predict that subjects will repeat a sequence verbatim. In making pairwise comparisons between high- and low-frequency items, there are several cases in which participant performance was contrary to expectations. Such exceptions are most noticeable in the distribution of fully correct responses, in which 5 pairs (out of 13) exhibited a reversed pattern. With respect to partially incorrect responses, there was a single exception[24] to the expected pattern, and for mean performance, there were two exceptions (see asterisked items in Table 3.3).

[24] Recoding the data with a broader definition of 'partially incorrect' responses adds one more exception: under this coding convention, *in spite of* is more prone to partially incorrect responses than *in fear of*. See Table 3.4.

Of course, the occurrence of exceptions to broader patterns is not especially troubling, given that the nature of the evidence is statistical, and (as discussed in Chapter 1) it is predicted that the same sequence of words may be activated holistically or compositionally, to varying degrees, under different circumstances. Some sources of variability become apparent if we examine the particulars of participant responses.

Toward this end, consider some of the recurring patterns in participant responses, summarized on the following pages in Table 3.5a (high-frequency sequences) and Table 3.5b (low-frequency sequences). These tables lists cases in which the same response appeared at least twice among the 38 participants; the number of occurrences appears in parentheses after each response. Paired items are asterisked in cases where at least one of the assessments indicated an exception to the expected pattern, i.e., if the high-frequency item was outscored in any measure of holistic retrieval by its low-frequency counterpart. The middle column in the tables lists different types of 'replacements.' Replacements are responses in which the target sequence does not appear, either because the sequence was supplanted by another (non-overlapping) word or sequence, or because of a more general constructional change in the participant's response. In the mean performance scores computed above, replacements were considered neither as evidence for nor against holistic retrieval of the sequence. The rightmost column lists partially incorrect responses by participants, i.e., incorrect responses which overlap with the target sequence in at least one word.

TABLE 3.5A
HIGH-FREQUENCY SEQUENCES: RECURRING DEVIATIONS IN RESPONSES

Target sequence	**Recurring replacements (with synonym word/sequence), or constructional change**	**Recurring partially (in)correct responses**
1. **in the same **way** as*	*like* (5)	*the same as* (11) *as* (5)
2. *all of a **sudden***	*suddenly* (3)	*all of the sudden* (9)
3. **as a **result** of*	*due to* (10)	*because of* (2)
4. *for the **sake** of*	*due to* (2)	*for* (*a career*) (2)
5. *in the **middle** of*	*around* (2)	*in* (9) (*on the floor*) *of* (2)
6. **on the **part** of*	*from* (16) *by* (2)	
7. *as **soon** as*		
8. **in **spite** of*	*+despite* (20) *against* (2)	
9. *in **terms** of*		*in* (3)
10. *on **top** of*		*on* (3)
11. ****according** to*	[#](*family*) *say/said* (12) [#](*family*) *claim/claimed* (4) [#](*family*) *argued* (2)	
12. ***back** to*		*to* (4)
13. *out **of***	*+outside* (2)	

*One or more average scores on this item was contrary to predictions, relative to the corresponding score for the matched counterpart.
+Under the alternate coding system (see Table 3.4), this response would be considered 'partially incorrect.'
[#]This response is accompanied by a change in the construction in which the target sequence appears. The response differs from true substitutions/replacements, since use of the original target sequence would be ungrammatical in this context.

TABLE 3.5B
LOW-FREQUENCY SEQUENCES: RECURRING DEVIATIONS IN RESPONSES

Target sequence	Recurring replacements (with synonym word/sequence), or constructional change	Partially (in)correct responses
1. **to the same **time** as*		*to the time of* (5) *to around the same time as* (2)
2. *all of a **boy***		*all* (*the drawings were*) *of a boy* (12)
3. **as the **name** of*		*for the name of* (6) *to be the name of* (3) *as the name for* (2)
4. *for the **child** of*		*to the child of* (3)
5. *in the **style** of*		*in the* (11) *in the traditions of* (2)
6. **on the **life** of*	#(*a screenplay*) *about* (7)	*about the life of* (7) *on* (2)
7. *as **big** as*		*as tall as* (2)
8. **in **fear** of*		*afraid of* (2)
9. *in **things** of*		*in such things* (2)
10. *on **half** of*		*on half* (7) *on over half* (*of*) (5) *on more than half* (*of*) (3)
11. ****looking** to*		*looking for* (4) *look to* (2)
12. ***through** to*	+*into* (13)	*through* (*the floor*) *into* (5) *down to* (4) *through* (*the floor*) *to* (2)
13. *out **to***		*to* (8) *outside to* (4)

*One or more average scores on this item was contrary to predictions, relative to the corresponding score for the matched counterpart.
+Under the alternate coding system (see Table 3.4), this response would be considered 'partially incorrect.'
#This response is accompanied by a change in the construction in which the target sequence appears. The response differs from true substitutions/replacements, since use of the original target sequence would be ungrammatical in this context.

A review of participants' responses suggests that multiple factors may influence performance in the dictation task, including a need for economy under time pressure. In the experiment, participants faced a distractor task, after which they had to reconstruct the sentence as accurately as possible within a short response period. In a few cases, this experimental setup led them to rephrase the context surrounding the target sequence, resulting in a more streamlined sentence. This pattern was noticeable on one stimulus item in particular, provided in (10a).

(10a) TARGET: *Last year, the actor was praised for playing the famous scientist, but **according to** relatives it was not a realistic portrayal*.

Among the participants, 19 responses (50%) included the high-frequency sequence *according to* verbatim, but another 19 responses (50%) omitted the sequence altogether due to a recast of the sentence. In such responses, the participant rephrased the semantics of 'according to' with a communicative verb (*say, claim, argue, state*) in an active construction. An example of such a response is in (10b).

(10b) SAMPLE RESPONSE: *Last year an actor was awarded for a portrayal of a scientist but the family said it wasn't realistic*.

For this item, the tendency for participants to change the constructional context accounts for the poor showing of *according to* on the 'fully correct' measure (50%); participants had more fully-correct responses (70%) on the low-frequency counterpart *looking to*. A recast such as (10b) seems to have little to do with the target sequence, but is instead motivated by participants' preference for a more concise sentential construction. Since *according to* would no longer be grammatical in the recast sentence, the omission of this target sequence reveals little about the nature of processing for the item of interest. Thus,

in computing the mean performance score, responses which changed the surrounding construction were assigned neither positive nor negative points.

Undoubtedly, a tendency to make stimulus sentences more concise also helped motivate many 'partially incorrect' deviations from the target sequences.[25] For instance, consider the high-frequency sequence *in the same way as*, on which participants were especially inaccurate. The stimulus sentence is provided in (11a).

(11a) TARGET: *The new bill increases penalties for white-collar criminals, arguing they should be sentenced* ***in the same way as*** *other criminals.*

Only 16% of responses reproduced this target sequence verbatim, compared with 41% of responses for the low-frequency counterpart, *to the same time as*. For this item (alone among all the stimuli), the high-frequency sequence was also outscored by its low-frequency counterpart with respect to percent partially incorrect responses, as well as the mean performance score.

The low scores for *in the same way as* were mostly attributable to shortened versions of the sequence which rendered the response more concise. For instance, 11 participants shortened the sequence to *the same as*, as in (11b).

(11b) SAMPLE RESPONSE 1: *The new bill* [portion crossed out] *increasing penalties for white-collar criminals saying they should be sentenced* ***the same as*** *other criminals*.

[25] Clearly, not all partially incorrect responses were motivated by economy; in some cases the participant merely substituted one word in the target sequence, while keeping the total number of words unchanged. One striking example of this is the change of *all of a sudden* to *all of the sudden* by 9 participants. Additionally, one participant substituted *all the sudden*. It is noteworthy that in addition to historical uses (*all of the sudden* is the older form), *all of the sudden* has been rising in frequency recently, as shown by a drastic climb since the mid-1980s in the Google Books N-gram Viewer. It may be that certain younger speakers actually perceive the target sequence as the less frequent variant, *all of the sudden*, due to phonological reduction and/or grammatical idiomaticity of *all of a sudden*. However, in the absence of additional information, all such deviations from *all of a sudden* are coded as 'partially incorrect,' for the sake of consistency.

Additionally, five participants shortened the sequence to *as*, as in the following example.

(11c) SAMPLE RESPONSE 2: *The new bill increases penalties for white-collar criminals arguing they should be sentenced **as** other criminals*.

Responses such as the foregoing make it apparent, in retrospect, that certain of the stimulus sentences may have been more cumbersome than would be ideal. The task facing participants in the experiment was to transcribe sentences word-for-word, to the best of their ability. Nevertheless, it is not terribly surprising that in a time-pressured situation they might inadvertently render certain sequences more concise. Moreover, there are isolated instances throughout the data in which participants shortened the target sequences, even in sentences which seem relatively concise: *in the middle of the living room* is shortened to *in the living room*; *on top of the refrigerator* is shortened to *on the refrigerator*, and so on. As discussed in Chapter 1, it is predicted that even as a sequence of words becomes more unit-like, the sequence need not instantaneously lose its internal structure, and the component words can still be perceived and accessed (Bybee 1998, Bybee and Scheibman 1999, Beckner and Bybee 2009). It is thus to be expected that when resources are limited, speakers might tend to adjust multiword sequences (for instance, by removing words) so as to streamline a sentence.

Consider an additional pattern in the responses for the high-frequency sequence *in the same way as*. Five participants (16% of codable responses) replaced *in the same way* as with *like*, as in the following example.

(11d) SAMPLE RESPONSE: *The new bill increases penalties for white collar criminals arguing that they should be charged **like** everyone else*.

In addition to the modifications discussed above, these substitutions further reduced the 'fully correct' score for *in the same way as*. Indeed, substitutions of this sort had a marked effect on most[26] of the high-frequency sequences that were outscored by their low-frequency counterpart. Reversals of the expected high/low frequency patterns arise in part due to the ready availability of synonym words or phrases that are shorter than the target sequence. Consider, for instance, the stimulus sentence for the high-frequency sequence *on the part of*:

(12a) TARGET: *The group has pushed for a more active role, but they have encountered resistance* ***on the part of*** *most family doctors.*

The percent of fully accurate responses for *on the part of* (around 34%) is lower than the percentage for the matching low-frequency sequence (*on the life of*, at 47%). Only 10 participants recalled *on the part of* verbatim; a larger number (16 participants, or 55%) rephrased the sentence so that the single word *from* took the place of *on the part of*. A typical response is as follows.

(12b) SAMPLE RESPONSE: *The group has pushed for a more active role, but has encountered resistance* ***from*** *family doctors.*

In addition to the 16 substitutions which followed this pattern, there were 2 responses which replaced *on the part of* with *by*. It is perhaps unsurprising that in a time-pressured task, participants would be inclined to replace a four-syllable sequence (*on the part of)* with a more concise, single-syllable equivalent (*from* or *by*)[27].

[26] The exception would be *according to*, which as previously noted, fares poorly due to syntactic factors not directly related to the target sequence.

[27] It might also be noted that the substituted items are also generally far more frequent than the target sequence, and thus may be more easily accessible in a time-pressured task. For instance, *from* is about 300

Consider also the similar case of *as a result of*, presented in sentence context in (13a).

(13a) TARGET: *In general the company never liked criticism from employees, but eventually the policy changed* ***as a result of*** *the worker's complaint*.

Again, fully accurate responses for this high-frequency sequence are relatively low (28%), compared with the low-frequency counterpart *as the name of* (63%). However, the low accuracy on *as a result of* is mostly attributable to replacement by a two-word synonym, as in (13b).

(13b) SAMPLE RESPONSE: *In general the company didn't like criticism but eventually the policy changed* ***due to*** *employee complaints*.

Indeed, *due to* was more popular in the responses than verbatim transcriptions of *as a result of* (with 11 and 7 responses, respectively).

A third high-frequency sequence which exhibits a particularly strong tendency toward replacement is *in spite of*, which was previously mentioned in Section 3.2.4.2. The stimulus sentence is provided in (14a).

(14a) TARGET: *Since I was still fond of my old car, I refused to look for a new one,* ***in spite of*** *my mechanic's advice*.

Participants produced *in spite of* in a fully correct form only 15 times (40% of codable responses), compared with 22 (64%) fully correct responses for the matched low-frequency item, *in fear of*. The lower accuracy for *in spite of* is largely due to the tendency to substitute *despite* for the target, as in (14b).

---

times as frequent as *on the part of* in COCA (Davies 2008). Of course, large frequency differences are to be expected in most cases when the substituted phrase contains a smaller number of orthographic words.

(14b) SAMPLE RESPONSE: *Since I still valued my old car, I refused to look for a new one* ***despite*** *my mechanic's advice*.

Among the participants, 20 responses (54%) replaced *in spite of* with *despite*. This particular substitution may be related to an ongoing shift in American English. Historical data from COHA (Davies 2010) and the Google Books corpus show clear trends: *despite* has been climbing in frequency for the last hundred years, and since the 1920s, *in spite of* has been declining. It may be that the concessives *in spite of* and *despite* are currently in competition with one another, and in everyday usage, speakers are increasingly switching to the more economical form. At the very least, it appears that *despite* is becoming increasingly accessible to English speakers, and it offers a natural substitution for *in spite of* in a time-pressured task.

An open question regarding the dictation methodology is how best to interpret substitutions along the lines of *in spite of* > *despite*[28], *in the same way as* > *like*, *as a result of* > *due to*, and *on the part of* > *from*. Such 'full replacements' indicate that in the memory task, participants were able to retrieve a synonym (or synonymous phrase) which does not involve activation of any of the component words from the target sequence. Such a replacement does not provide any evidence that the target sequence is accessed as a series of individual words. Indeed, in calculating mean performance scores, I have assumed as much, because full replacements are assigned no penalty in the summary score.

Moreover, note that replaceability of a multiword sequence constitutes one source of evidence that the sequence comprises a syntactic constituent (Quirk et al. 1985, Fabb

[28] For purposes of the immediate discussion, I have assumed that *in spite of* > *despite* is best regarded as a full replacement of the target sequence. Of course, an alternate interpretation is possible, as I discuss at the end of Section 3.2.4.2.

2012, Beckner and Bybee 2009). Of course, a sequence's status as a syntactic unit is a weaker requirement than assigning it status as a holistic unit; however, unitary syntactic function would be a prerequisite for holistic retrieval. It is interesting that in the data, high-frequency sequences are more prone to full replacements of this sort. This can be demonstrated by examining the distribution of replacements in which the target sequence is replaced by a synonymous word or phrase (containing no component words of the target), while the surrounding syntactic context remains unchanged. (Thus, the present analysis does not include omissions caused by broader recasts or constructional changes.) Among high-frequency sequences, true replacements of this sort occur on 17.8% of all codable responses, compared with 4.4% among low-frequency sequences. In a t-test paired by item, this difference falls short of significance ($p = 0.055$), but when paired by participant, the difference is highly significant ($p < 0.001$). Similar results are yielded if we focus solely on replacements that consist of a single orthographic word, so as to attend more closely to typical diagnostics for syntactic constituency. In this analysis, we thus include substitutions such as *in the same way as > like*, but exclude responses such as *on the part of > due to*, since *due to* consists of multiple words. This analysis indicates that 14.48% of responses for high-frequency sequences consist of a one-word replacement, compared with 3.63% of low-frequency sequences. This difference is not quite significant in a t-test paired by item ($p = 0.078$), but highly significant when paired by participant ($p < 0.001$). Thus, among high-frequency sequences, deviations from fully correct transcriptions are more likely to suggest that the sequence functions as a syntactic unit.

As one final point regarding the dictation task results, we should note evidence in the data that it is not necessary for a sequence to be high in frequency for it to be reproduced quite accurately. Indeed, Schmitt et al. (2004: 131) observe that verbatim recall of a sequence does not prove that the sequence was retrieved holistically; it may simply have been assembled successfully on a word-by-word basis. Moreover, additional factors may aid in the successful recall of a particular sequence, on the basis of abstract grammatical patterns rather than holistic access of a particular, multiword chunk. A case in point is the low-frequency sequence *as big as,* presented here in sentence context.

(15) TARGET*: The little girl struggled to carry the branch to the campfire, because it was really almost **as big as** she was*.

Participants reproduced this target sequence verbatim around 95% of the time, making it the second-most accurate sequence overall, just behind its high-frequency counterpart, *as soon as*. The relatively low frequency of the sequence *as big as* is not crucial, given that the sequence instantiates an English comparative construction having the form *as* ____ *as,* where the open slot is filled by a gradable adjective or adverb. This construction offers a conventional way for speakers to express 'comparison in relation to the same degree' (Quirk et al. 1985: 458), and this conventionality presumably affords assistance to participants in reconstructing the target sentence. This striking level of accuracy again suggests that, in addition to token frequencies, various factors have an effect on participants' performance in the dictation task.

3.3. **Conclusion**.

Despite the null outcome reported in the Schmitt et al (2004) study, the verbatim memory task offers a fruitful methodology for investigating holistic processing of

multiword sequences. Although certain shortcomings persist in the original Schmitt et al. design, reanalysis of their data suggests that high frequency of a sequence is associated with more accurate verbatim memory, and diminished likelihood of interruption or modification of the sequence.

Moreover, the new experiment using the Schmitt et al. dictation methodology gives evidence that higher-frequency sequences are more likely to be retrieved as wholes, rather than on a word-by-word basis. Most strikingly, the data show that in recalling high-frequency sequences, participants are significantly less likely to produce partially (in)correct variants which signal the decomposition of such sequences into their component words.

Notwithstanding quantitative trends in the data, clearly subjects at times interrupt or modify high-frequency sequences in the dictation task. A review of subject responses reveals that there are various reasons why verbatim recall might be unsuccessful; often there are alternate ways of expressing the ideas found in the sentence, and these alternate versions may be more concise.

The quantitative results presented in this chapter provide new evidence that corpus-derived frequent sequences are more likely to be retrieved as prefabs, supplementing prior research in this area. The present experiment was designed specifically to revisit the methodology and findings of Schmitt et al.'s (2004) verbatim memory task. As such, the current experiment is designed so as to control for a single independent variable, token frequency. Thus, possible roles for relative frequency (as considered in Chapter 2) remain an open question to be addressed in future work using

the dictation methodology. However, relative frequency, as measured by Mutual Dependency, forms a central part of the investigation discussed in the next chapter.

CHAPTER 4. HOLISTIC RETRIEVAL OF MULTI-WORD VERBS: STUDIES OF AFFIX POSITIONING ERRORS

4.0. **Introduction**.

The present chapter investigates evidence for prefabricated, multiword units based on the positioning of affixes in speech. Studies of conversational errors show that speakers at times insert an affix at the wrong position in a sentence, as in *It probably gets out a little* → *It probably get outs a little* (Garrett 1980: 202). In speech errors of this type, a speaker mistakenly applies an affix at the periphery of a salient word sequence (e.g., *get-out* + *-s*), rather than inside the sequence as intended. I propose that we might use such an error as evidence that the speaker has accessed the word sequence holistically, and treated it as a 'wordlike' unit with respect to morphology. Errors such as *get outs* provide one example of syntagmatic, inflectional or derivational errors generally called 'affix shifts' by Stemberger and MacWhinney (1986b).

We might note that on occasion, young language learners exhibit speech patterns of this sort. For instance, in (1) – (7) below, I list a collection of novel forms I have observed in my own son's speech, including age(s) of production in year;month format. (Items marked with a '+' were observed on multiple occasions.)

(1) *That what he look likes*. (2;10)

(2) *cool offed*+ (2;10 - 2;11)

(3) *come offed again*+ (3;0 – 3;1)

(4) *Why my mama miss mes*? (3:1)

(5) *stand upped* (4;9)

(6) *make sures*, *make sured*+ (4;11-5;2)

(7) *It show you*s. (5;6)

(8) PARENT: *How's that cleanup coming?*
CHILD: *It's coming! I'm playing while cleanupping!* (6;5)

Utterances such as the foregoing provide one line of evidence that the language learner has inferred different linguistic units than those of adults' grammar[29]. Language acquisition requires the learner to segment speech into various kinds of units, and it is possible children will initially learn some multiword sequences as units that lack internal structure (Peters 1983). For instance, if a child interprets the sequence *stand-up* as a single English word (a verb), it is then natural to apply the English past tense pattern productively, thus creating the novel inflected form *stand-upped*.

In adult speech, affix shift errors exhibit the same pattern, but the underlying mechanisms are more gradient and transitory. When adults occasionally say things like *It get-outs*, this error is not an immediate artifact of language acquisition, but gives a glimpse into the workings of speech production. An adult can, on reflection, be quite aware of the internal structure of a multiword sequence, but nevertheless under-analyze the sequence during online speech processes. An affix shift error hints at holistic retrieval of a multiword sequence, insofar as apparently (1) there is diminished activation of the component words (in the above example, this would include the target verb, *get*, which fails to be inflected as a verb), and (2) prefabricated production of the word sequence as a wordlike unit, which receives a verbal or nominal inflection as a unit.

The relevance of affix shift errors has been previously noted by Quirk et al. (1985) in a discussion of multi-word verbs, that is, multiword sequences which behave

[29] Of course, some of the utterances in (1) – (8), in particular those produced at later ages, could certainly arise from mechanisms more characteristic of adults' affix shifts, rather than being holophrastic errors of acquisition.

'either lexically or syntactically as a single verb' (1150). In a footnote, Quirk et al. remark on speech errors such as 'The editor must do precisely as he *see fits*,' observed during a radio interview; the 'shift of the inflection from the verb to the adjective testifies to a tendency for speakers to perceive the multi-word verb as a single grammatical unit' (1985: 1151, note a). Quirk et al. say that errors of this type 'deserve attention' in the study of multiword units, but do not present a study of such errors themselves (1151, note a). Similarly, Wray (2008) writes that grammatical indications that a multiword sequence has 'morpheme equivalent status' (that is, that the sequence lacks internal structure) 'often come in the form of errors' (119).[30] To illustrate this idea, Wray provides a single example, in which a Kuwaiti official produces the plural form *weapon of mass destructions*, indicating that the four-word sequence is being treated as a single lexical item[31] (2008: 119). Moreover, affix shifts such as *get-outs* parallel a diachronic process in which inflections may come to be 'externalized,' as in *sisters-in-law* > *sister-in-laws*. Haspelmath (1993) writes that as certain expressions 'come to be felt as single words, speakers externalize the inflection' (289). Such changes are not frequent, but they are certainly attested crosslinguistically, and indeed, there are cases observed in which a verbal inflection moves outside a postverbal particle (Haspelmath 1993: 286-287). Thus, previous researchers have observed that affix shifts may provide insights into multiword units in the mental lexicon, but discussion has generally focused on isolated examples.

[30] Given the gradient account of unit retrieval in Chapter 1, I would make no general claim that affix shifts provide evidence of 'morpheme equivalent' units in adult speech. My views in this chapter are more modest: patterns among affix shifts may provide evidence that certain multiword sequences are more likely to be retrieved as a unit.

[31] Wray does not comment on the language background of the speaker in question. It is possible for L2 speakers to learn certain sequences as holistic units, in much the same way as L1 learners (Wray 2002), and such influences could certainly be at work in this speaker's processing of *weapon of mass destruction*.

Based on a review of the literature, it would seem that the present chapter is the first systematic, quantitative study of affix shifts with respect to multiword sequences.

The hypothesis of the present chapter is that in adult speech, affix shift errors will not be distributed randomly, but will be overrepresented with respect to prefabricated phrases. The distribution of errors, of course, is expected to be probabilistic rather than absolute. In addition to influences from prefabs, there may be a variety of factors which in fact motivate the occurrence of affix shifts[32], but as is standard in speech error research, the question is whether certain classes of errors are more likely than others (Fromkin 1973, MacKay 1979). The predictions to be tested, then, are whether increased Token Frequency and/or Mutual Dependency of a multiword sequence correspond to an increase in the rate of affixes that are erroneously shifted outside that sequence.

In the present chapter, I present two studies of affix shift errors, based on data gathered from naturalistic as well as controlled laboratory settings. As a preliminary to this research, it is helpful to reiterate more precisely the class of errors of interest. Garrett (1980) and Stemberger and MacWhinney (1986b) broadly define an affix shift as an error in which an affix occurs earlier or later in the sentence than intended, without distinguishing between the 'earlier' and 'later' cases. However, with respect to the hypotheses of the present chapter, the distinction between these errors is quite important. We may distinguish between what I will call 'inbound' affix shifts, in which a suffix moves inside of its target (*It dead ends* → *It dead+s end*), and 'outbound' affix shifts, where an affix moves outside of its target, thus appearing outside a word sequence of

[32] For instance, there could be priming influences from the surrounding conversational context, or there could be effects arising from 'competing plans' for an utterance (Baars 1980). As one example, the error 'as he see fits' could partially arise from alternate plans for the sentence in which *fit* was activated as a verb.

interest (*gets out* → *get out*+*s*). The notion of 'inbound' vs. 'outbound' errors allows us to be more precise about affix position with respect to some linguistic unit, rather than merely saying a given affix occurs early or late in a particular shift error.[33] However, I will talk about 'inbound shifts' with considerable care, because the terminology implies that a linguistic unit of some kind is involved. With respect to 'early' affix shifts on compound forms (such as *deads end*), an affix moves inside a lexical unit, and the 'inbound' terminology would clearly be appropriate. However, in certain contexts it is not clear whether an affix arriving prior to its intended target has anything to do with linguistic units. For instance, in a speech error by President George W. Bush, a possessive marker arrives two words earlier than intended: *The decision to put people's in harm_ way*. Such errors may simply be errors of morphological anticipation, and if unitary status is not immediately apparent, I will refrain from describing premature affixes as 'inbound.'

The hypothesis of the present study can only directly address outbound affix shifts, since holistic retrieval of a phrase would encourage affixes to attach outside, rather than inside, a chunked group of words. Inbound shifts and other premature suffixes must be attributed to some other mechanism, such as anticipation of a morpheme that has entered the speaker's buffer memory for a later portion of the sentence (Levelt 1989).

Several previous studies have taken note of affix shift errors (both early and late occurrences), but have not related such errors to the study of multiword sequences.

[33] We may extend the principles of affix shift errors to languages other than English, and propose that the difference between 'inbound' and 'outbound' errors depends on the particular grammar of the language. If the shifted inflection were a prefix rather than a suffix, an 'outbound' error would be one in which the prefix occurred one or more words earlier than its target: word1 + [prefix-word2] → prefix- + [word1 + word2]. Such an outbound prefix shift would hint that word1 and word2 form a prefabricated unit. For a related diachronic process, see Haspelmath (1993: 287-288).

Stemberger (1984), Stemberger (1985), and Stemberger and MacWhinney (1986b) analyze naturally-occurring shift errors, but their analyses are focused on the storage of affixes in the lexicon. Stemberger (1984) examines lexical and affix shift errors generally, analyzing 203 conversational errors in which an affix or a word occurs too early or too late in speech, as in *We tried making it* →*We tried it making* or *If it breaks* → *If its break* (289). Stemberger finds that in general, higher-frequency items (whether open-class or closed-class) tend to appear early in shifts[34], hinting that high-frequency items are 'overactivated' during speech production, and thus executed too early (297; see also Stemberger 1985: 87). This pattern would imply that grammatical suffixes (being high in frequency) have a general tendency to appear prematurely in speech, thus providing one mechanism to account for the occurrence of early affix shifts in English, such as *deads end_*.

Stemberger (1985) provides some quantitative results which focus more specifically on affix shifts. Among a collection of 40 affix shifts collected in conversation, Stemberger finds that in 17 errors, an affix appears earlier than its target; in 13 errors, an affix appears after a clitic[35] which immediately follows the target word; and in 10 errors, an affix appears on some other (non-clitic) word following its target (1985:

[34] Stemberger (1984:290) bases his conclusions about frequency on shift errors involving two open-class items, or one open-class and one closed-class item. In the present chapter, a large number of the shift errors involve a misordering of two closed-class items, as in *get out-s*. There are no indications that Stemberger (1984) performed frequency comparisons of shift errors involving two closed-class items (such as a verb particle and a grammatical suffix), so it is unknown whether the higher frequency/earlier activation pattern can be generalized to such cases.

[35] It is not entirely clear what would be included in Stemberger's (1985) 'clitic' category, and a complete list of his errors is not available. He does include the example *look- atting*, which would apparently be an instance of an affix shifting to follow a clitic (1985: 156). In a related discussion, Stemberger (1989: 171, n6) refers to a separate class of affix shifts which follow a particle, as in *tie-upped*. However, is not immediately apparent that *look at* and *tie up* even represent the same type of verbal unit, nor that they should be separated *a priori* from other word sequences. Note that *look at* and *tie up* do not have identical grammatical properties; the former would be classified as a prepositional verb and the latter as a phrasal (transitive) verb (Quirk et al. 1985: 1153 ff).

156). Following the terminology of the present chapter, this dataset implies a general pattern in which outbound (late) affix shifts (n=23) are marginally more likely to occur than early affix shifts (n=17).[36]

Stemberger and MacWhinney (1986b) present a quantitative study of affix shift errors, but devote their attention to the frequencies of individual words plus inflections. The aim in that study is to determine whether stems and affixes are assembled compositionally, even in cases where the surface frequency of a stem-plus-affix form is high. Stemberger and MacWhinney (1986b) thus examine the surface frequency of affixed forms that were involved in affix shift errors; for example, in an error like *telling us→ tell-us-ing*, the relevant measure is the surface frequency of *telling*. If high-frequency stem/affix pairs were retrieved from memory in preassembled form, then the expectation is that these high-frequency pairs should be less prone to misplacement of the inflection, and should thus be under-represented among affix shifts. However, based on a collection of 41 naturally-occurring affix shift errors, Stemberger and MacWhinney find that high-frequency forms are actually over-represented (although the difference is not significant). From this null result, they conclude that even high-frequency morphologically complex words are assembled compositionally from the lexicon.

In sum, the prevalence of affix shifts in speech has been noted by a number of researchers in psycholinguistics, albeit outside the domain of examining multiword units in the lexicon. It turns out that there are three different types of affix error which may be

[36] My analysis of this dataset differs markedly from Stemberger's, since I merge all 'late' affix shifts into an 'outbound' group. As indicated in note 34, Stemberger considers the 13 errors involving a clitic + affix to be qualitatively different: he says these 'should be viewed as early execution of a clitic, and should be distinguished from other delayed execution of affixes' (1985: 212, n4). Thus Stemberger excludes these 13 errors for statistical purposes, and states that 'early execution of an affix is twice as common as late execution,' apparently based on a comparison of the 17 early-shifted affixes with the 10 late-shifted affixes in the 'other' category (1985: 157). However, for purposes of the present study, it would be circular to assume without argument that only certain multiword sequences motivate affix shifts, while others do not.

of interest in the present research project. First, there are the full outbound affix shifts, such as *It get outs* and *She see fits*. Syntagmatic errors of this sort offer the most straightforward evidence in support of the strong holistic retrieval hypothesis, that is, the view that a complex sequence may be retrieved as a whole, while activation is diminished for its component parts. In full outbound shifts, the tendency for inflections to be 'deflected' to the periphery of particular multiword sequences would attest that (1) the sequence tends to be processed as a whole unit with respect to morphosyntax, and (2) the morphosyntactic status of component words within the sequence is diminished.

Secondly, there exist more complex affix errors which provide a somewhat weaker form of evidence regarding holistic retrieval. In the analyses of Sections 4.1 and 4.2, I supplement full affix shifts such as *get-outs* with 'double-marked' errors, such as *It gets outs*. Errors of this type append an affix to a sequence and to one of its words, implying that with respect to morphosyntax, the speaker has concurrently activated the multiword sequence as a whole unit, and as an assemblage of component words. Such a view is entirely coherent with the gradient account of activation described in Chapter 1; if alternate approaches to retrieval fail to resolve in time, it is reasonable that the speech output would give evidence of both types of retrieval. As such, errors of this type will be taken as indicators that the multiword sequence has been activated as a whole, in addition to a competing, more analytical, activation. Moreover, double-marked shifts also parallel a diachronic process, since double-marked affixes are attested as an intermediate (and often overlapping) stage of inflectional externalization (Haspelmath 1993).

Finally, consider more ambiguous errors in speech, such as the examples in (9)-(11).

(9) *I've never had it look at*.

(10) *As I've said, that's only one leg of the stool. And that these other leg of th-, legs of the stool will be rolled out, uh, systematically, uh, in the coming weeks*. — President Barack Obama, January 29, 2009

(11) *What is the object of our study? The object of our study are, broadly speaking, fourfold: pronunciation, grammar, meaning, and attitudes toward language change*. —Seth Lerer, *The History of the English Language*, Part I.

In sentence (9), which I observed in casual conversation, the speaker clearly intended to say *looked at* (although the error went uncorrected). When grammatical affixes are unintentionally omitted, such slips may be described as 'no marking errors' (Stemberger and MacWhinney 1986a). In (10), it is not clear whether President Obama interrupted an affix shift error (*these other leg_ of the stools*) before it was completed, or whether he corrected a no-marking affix error (*these other leg_ of the stool*). In either case, an error of this sort is potentially revealing with respect to the retrieval of multiword sequences. At the time of the press conference of (10), 'leg of the stool' was a recurring metaphor used by Obama in discussing his administration's plans for economic recovery. If 'leg of the stool' —in its uninflected form— was a multiword unit used in planning speech in this context, it would be unsurprising if the sequence proved to be slightly more resistant to having inflections inserted. Recall from Chapter 2 the notion that holistic sequences are relatively difficult to interrupt or modify (Wray 2006: 592), and this principle may very well apply to inserting inflections into a 'prepackaged' sequence.

Thus, it is arguable that no-marking errors may provide one line of evidence for holistic retrieval, on the assumption that inflections are occasionally omitted since a prefabricated sequence is resistant to interruption. However, such evidence must be interpreted cautiously, since no-marking errors are known to be more likely on low-

frequency forms (Stemberger and MacWhinney 1986a, 1986b), apparently because it is easier to assemble (or retrieve pre-assembled) high-frequency base-plus-affix pairs. Moreover, zero-marking of third person singular verbs is a feature of nonstandard varieties of English (Labov 1969, Green 2002), and it is possible that absent verb marking will be a sociolinguistic variable rather than an indicator of holistic processing.

While noting these caveats, in Section 4.2 I include information about loss of verb affixes in the context of a particular experimental task, as a further source of evidence about the processing of multiword sequences. An additional justification for including this evidence is the particular nature of the experimental task, in which participants are supposed to inflect a stretch of speech they have just heard. In situations where speakers are repeating phrases that have just been used in the context (as in examples (10) and (11) above), the most relevant behavior is not the omission per se— it is the failure to interrupt or alter a sequence which is being produced, or re-produced, as a preassembled unit. Indeed, it is possible for the opposite error to occur, if a speaker fails to produce an uninflected form, but instead repeats an inflected sequence which has just been encountered. As an illustration from conversation, I observed just such an error in my own speech. In the conversational context, the noun phrase *canine teeth* had been repeated at least three times before I said example (12) (while pointing at my mouth).

(12) *That's a canine teeth. Canine tooth.*

A speech error of this sort is unsurprising, not just because *canine teeth* was prominent as a recurring unit in the conversation, but also because *canine teeth* is a reasonable candidate for a prefabricated, multiword unit.[37]

[37] Indeed, 'teeth' are usually more salient to people as plural entities, and a quick check in COCA confirms that *canine teeth* occurs about four times as often as *canine tooth*.

A prefab model would predict that certain sequences are more prone to being produced as preassembled, fixed units, and that such sequences may be more difficult to alter (either via addition or deletion) once they have been primed as units. Since the experiment of Section 4.2 requires immediate repetition of word sequences, in some analyses I include data drawn from speakers' failure to alter the stimuli.

Below, in Section 4.1, I present a systematic study of outbound affix shifts collected from conversation, and in Section 4.2, I present an experimental study designed to elicit such errors in the laboratory.

4.1. **Naturally-occurring affix shift errors**.

Although outbound affix shifts have been regularly observed by a range of speech error researchers, these errors are in fact rare enough in speech that quantitative study presents real challenges. For instance, a review of the 191 speech errors Garnham et al. (1981) compiled from the 170,000-word London-Lund corpus turns up no tokens of affix shift errors. In the speech error corpus collected by Stemberger (1985), there were 23 outbound affix shift errors out of a total of 6300 errors. Thus, in Stemberger's data only 0.37% of all speech errors are outbound shifts; that is, less than one out of a hundred speech errors is of interest in the present study.

Given the foregoing difficulties in assembling relevant data, it seems that conversational error collection may need to take place over the course of several years, and it is helpful to look to a variety of sources for data. Toward this end, the present analysis of conversational errors will examine outbound affix shifts gathered from an online speech error database, supplemented with a small set of errors I have collected myself in everyday conversations.

First, the Fromkin Speech Error Database (Fromkin 2000) provides a collection of naturalistic speech errors collected by a variety of researchers over the course of 30 years. The database contains a total of 6398 English speech errors sorted into various categories. A review of errors classified as 'morphosyntactic shifts' or 'morphological shifts' yields a total of 29 affix shifts, of which 18 are outbound affix shifts[38]. The 18 errors appear in Table 4.1.

Apparent intended utterance	Error in context, where available	Type of affix involved in shift
giving us	*a letter from Leo Scarry give us-ing*	Verbal – progressive
paying for	*I'm pay foring it all together.*	
shutting up	*I should be shut upping*	
comes in	*and Rachel come ins*	Verbal – 3PSG –*s*
comes on	*It come ons at...*	
wants to come	*if she want to comes here*	
makes sure	*she make sures*	
adds up	*add ups to*	
ends up	*he end ups*	
comes up	*when someone come ups to me*	
forgotten about	*I'd forgot abouten that*	Verbal – past participle
parts of it	*some part of its are*	Nominal – plural
phones rang	*all the phone rangs*	
EPLs tend	*EPL tends to be*	
Jerry's Pancake	*Jerry Pancake's house*	Nominal – possessive
easily enough	*easy enoughly*	Derivational -*ly*
highly verbal	*What does it mean to be high verbally?*	
logically speaking	*logic speakingly*	Derivational -*ly* (Plus loss of –*al*)

TABLE 4.1. Outbound affix shift errors in the Fromkin Speech Error Database.

[38] Interestingly, these figures indicate that the 18 outbound affix shifts constitute 0.36% of the 6398 total speech errors in the database – a figure quite close to the 0.37% value found in Stemberger's (1985) database. However, it should be noted that I have intentionally excluded two shift errors included in the Fromkin Database. The first of these is *sanitary inspector* → *insanitary spector*, which I exclude because the shifted *in-* syllable is not a productive prefix, and the error may thus be phonological in nature. The second exclusion is *Ralph and my's uncle*. The English possessive marker attaches to phrases, rather than words (e.g., Pinker 1999: 50), and this results in confusion among speakers about prescriptive norms for attaching *'s* to conjoined nouns and possessive pronouns (as shown by numerous queries on Yahoo! Answers). Thus an utterance such as *Ralph and my's* may in fact not be an error in online processing.

As a supplement to the above set of errors, I have been collecting outbound affix shifts in everyday conversation over the course of approximately 6 years. This collection process confirms that outbound shifts are quite rare, since this process yields only 8 additional errors, presented in Table 4.2.

Apparent intended utterance	Error in context, where available	Type of affix involved in shift
going for	*I'll be go foring, going for a run*	Verbal – progressive
gets along	*Everybody just get alongs great*	Verbal – 3PSG –*s*
goes ahead	*He go aheads and reads it.*	
?gets, goes (to) get	*Stay here while Daddy go gets it.*	
goes home	*Everyone goes homes to nap.**	
comes back	*come backs*	
kept you up	*We kept you upped.**	Verbal – past
rides home	*ride homes*	Nominal – plural

TABLE 4.2. Outbound affix shift and double-marked affix errors collected by the author.

Note that my own collection of errors includes two of the 'double-marked' errors described in Section 4.0, that is, cases in which an affix was applied to an individual word in addition to a multiword sequence. These double-marked errors are indicated with an asterisk in Table 4.2. The Fromkin collection contained no double-marked errors, with the possible exception of *I'd forgot abouten that*. In this error, the stem is changed on the verb as expected (*forget/forgot*), but the past participle *–en* shifts onto the following word. In an effort to locate double-marked errors in the Fromkin Speech Error Database, I did a follow-up search for morphological and morphosyntactic perseverations, but located no additional errors of interest.

Collectively, Tables 4.1 and 4.2 contain 26 errors that are relevant to the current study. Unfortunately, for the sake of uniformity we must exclude a few of these errors

from quantitative analysis. Three of the errors (*want to comes*, *part of its*, *kept you upped*) involve cases where an affix moves two words away from its target, rather than one word. Thus, in most instances the relevant multiword sequences are of length 2, but we will exclude these three errors involving sequences of length 3. As discussed in Chapter 3, whenever possible, it is best to make frequency comparisons among n-grams of the same length. The overwhelming majority of the conversational errors involve two-word sequences, and I limit the analysis to this set so that corpus metrics are more consistent across the set of items. Additionally, I will exclude the error *logically speaking* →*logic speakingly*. This error does seem to involve the outbound shift of a derivational affix (-*ly*), but the error stands apart from others in the set because an affix is deleted as well (-*al*).[39]

4.1.1. **General methods and materials**.

After removing the above exceptional cases, we are left with 22 errors which may be used in quantitative analyses. More specifically, we have a collection of 22 two-word sequences (i.e., bigrams) which are to be evaluated quantitatively in this study. The crucial question to consider is whether bigrams having particular corpus metrics are *overrepresented* or *underrepresented* in the collection of outbound shift errors. This question is statistically more complex than it may seem at first. We may divide the bigrams into 'high' and 'low' categories for Token Frequency and Mutual Dependency, but it is not immediately obvious how to identify the expected number of errors in the

[39] An additional analysis is also possible, which focuses on the shift of word roots rather than affix shifts. Unlike most of the errors in the collection, the target utterance in this case has flexibility in the ordering of words: *speaking logically* and *logically speaking* would both be acceptable. This flexibility could naturally lead to competition between plans for the sentence (Baars 1980, 1992), and blending the plans could result in some of the affixed material becoming stranded from the target root (*speaking logic-ally* , *logic-ally speaking* → *logic speaking-ly*).

bins we define. For instance, it is insufficient to divide bigrams into arbitrary bins, and then simply count the raw numbers of errors in each group. Broadly speaking, a random sample of items drawn from speech would be expected to contain a disproportionate number of high-frequency items, because high-frequency items have a greater 'number of opportunities' to be selected under any criteria (Sellen and Norman 1992: 333). More specifically, highly frequent word sequences have a higher baseline probability of occurring as errors: if a particular sequence is said more often, it has a greater chance of being said wrong.

Thus, my methods in this section will use several varieties of corpus analysis to determine what would be the expected number of bigrams in a particular category, assuming a random distribution. If a particular category of bigrams (specifically, a set of affix shift errors) differs markedly from this random distribution, that will constitute evidence that the category is especially likely (or unlikely, as the case may be) to result in outbound affix shifts. The statistical methods described here are similar to those used by Stemberger and MacWhinney (1986b) for individual word frequencies, although they leave many details of the approach unexplained in the paper. In general, such analyses require exhaustively tallying all the units (meeting some criterion) in a corpus, and weighting groups of such units by Token Frequency. In the subsections below, I describe three versions of such an analysis.

Before describing these analyses, it is helpful to have general background about the raw corpus materials that will be cited repeatedly in this section, and in Section 4.2. The automated searches performed in this chapter require access to complete corpus textfiles. Moreover, it is best to have as large a corpus as possible, and to have the corpus

data be based on spoken English, since the affix errors of interest are, of course, spoken. Thus, I combined three different corpora of spoken American English to create a composite spoken corpus, containing approximately 5 million words. The breakdown of this composite corpus is listed in Table 4.3.

Corpus	**Citation**	**Approximate Word Count**
Switchboard Corpus	Godfrey et al (1992)	3 million words
Michigan Corpus of Academic Spoken English (MICASE)	Simpson et al. (2002)	1.8 million words
Santa Barbara Corpus of Spoken American English, Parts 1- 4	DuBois et al. (2000-2005)	378,000 words

TABLE 4.3. Contents of the 5-million word composite spoken corpus.

Additionally, one of the analyses below requires the use of corpus text that has been tagged for word classes. I thus make use of a tagged version of the 1 million-word Brown Corpus (Francis 1965), consisting of written American English.

Finally, as in Chapter 3, I also frequently consult the 450 million-word Corpus of Contemporary American English (COCA, Davies 2008-) to check the findings from other corpora.

4.1.2. **Analysis 1: Comparison to all bigrams in composite spoken corpus**.

First, I describe the simplest corpus-based distributional analysis, in which I find a halfway dividing point among all bigram tokens in the 5 million-word composite spoken corpus. This analysis is performed by a program I have written in Java, taking the following approach:

(i.) Tokenize all bigrams in the corpus, that is, identify all valid two-word sequences (discarding any sequences that cross speaker turns, cross sentences, punctuation, or other pause boundaries, or which are marked as uncertain transcriptions).

(ii.) Count the number of occurrences of each bigram, and sort all bigrams by frequency. There are 687,216 valid, distinct bigrams in the composite corpus, ranging in frequency from 1 (more than 400,000 of the bigrams occur only once each) up to 43,071 (for *you know*, the most frequent bigram in the corpus).

(iii.) Based on the sorted frequency list, find the midpoint for all bigram tokens in the corpus. That is, identify a frequency such that half of all bigrams are less frequent, and half are more frequent. There are a total of 4,187,085 valid bigram tokens in the corpus; the analysis identifies the midpoint frequency such that ~2,093,542 bigram tokens are more frequent, and ~2,093,542 bigram tokens are less frequent.

Performing the above-described analysis yields a midpoint frequency of 105 in the composite corpus. Thus, in a randomly chosen set of 22 bigrams, we would expect 11 to have corpus frequencies of 106 and above, and half to have corpus frequencies 105 and below.

For the 22 naturalistic outbound shift errors in our set, the actual composite corpus frequencies, and the high/low category divisions, are listed below in Tables 4.4a and 4.4b. The current corpus analysis yields 9 error bigrams which are above the midpoint frequency, and 13 which are below. This division does not significantly differ from the expected values (11 items per category). A chi-square goodness of fit test yields a p-value of 0.5271.

'High Frequency' bigrams: N = 9

Bigram	Composite frequency
come up	494
go ahead	474
make sure	453
end up	448
come in	445
pay for	389
come back	384
come on	196

TABLE 4.4A. Error bigrams above frequency midpoint for composite spoken corpus.

'Low Frequency' bigrams: N = 13

Bigram	Composite frequency
give us	100
go get	92
go home	90
get along	76
add up	35
forgot about	21
shut up	20
easy enough	10
phone rang	7
ride(Noun) *home*	1
Jerry Pancake	0
EPL tend	0
high verbal	0

TABLE 4.4B. Error bigrams above frequency midpoint for composite spoken corpus.

As a second attempt using the same basic corpus analysis, we might consider a three-way frequency split across the composite spoken corpus. A slight modification to the Java script indicates that in the composite corpus, one-third of bigram tokens have frequencies between 1 and 21 (low-frequency); one-third have frequencies between 22 and 424 (mid-frequency), and one-third have frequencies of 425 and above (high-frequency). However, comparing with the 22 errors in our set, the results are again null with respect to frequency. There are 5 high-frequency bigrams, 9 mid-frequency bigrams,

and 8 low-frequency bigrams. This three-way distribution does not differ significantly from a randomly selected set of corpus bigrams, yielding a chi-square p-value of 0.5543.

Thus, this initial analysis finds that the bigrams in our error set do not significantly differ in frequency distribution from bigrams in the composite corpus, yielding a null result. However, this first attempt may be rather naive, insofar as the corpus search includes all bigrams, rather than restricting the search to two-word sequences which could conceivably result in an outbound affix shift. Indeed, the 'high-frequency' bigram category described in the above attempts is overrun with many irrelevant sequences for our purposes, including a large number of sequences starting with (non-suffixable) closed-class words (*of the*; *in the*; *and then*; *it was*; *to be; I think*; and so on). A more selective corpus analysis would be in order; two such approaches are discussed in Section 4.1.3 and 4.1.4. In Section 4.1.5, I present a rather different analysis focusing on distinct classes of affix shifts.

4.1.3. **Analysis 2: Comparison to all Verb- and Noun-initial sequences in the Brown Corpus**.

In this second frequency analysis, I attempt to more meaningfully approximate the sample of bigrams against which our set of outbound affix shifts should be compared. Note that all of the errors in the set consist of bigrams beginning with a content word. Each of these content words, of course, can have a suffix appended to it (otherwise, an affix shift would not be possible). There are various ways we might converge on a sample which represents such bigrams in a corpus: perhaps by disallowing function words, or by searching for words ending in particular morphological affixes.

In the current analysis, I make use of an existing part-of-speech markup of the Brown Corpus (Francis 1965). This approach will allow me to restrict the sample of 'valid' bigrams to those that have a particular part of speech as the first word. Most of the suffixable words in our error set (20 of the remaining 22) are nouns and verbs, and in order to ensure uniformity in the sample, it is on these word classes that I will focus the analysis. Thus, for the current analysis, we will need to discard two additional items from the set (*easy enough* and *high verbal*)[40].

A modified version of the Java script described in Section 4.1.2 then processes the tagged Brown Corpus as follows:

(i.) Tokenize all bigrams in the corpus, that is, identify all valid two-word sequences. Impose the additional restriction of only including bigrams which begin with a verb, a common noun, or a proper noun.[41]

(ii.) Count the number of occurrences of each bigram, and sort all bigrams by frequency. The resulting list contains 137,516 distinct noun-initial and verb-initial bigrams.

(iii.) Based on the sorted frequency list, find the approximate midpoint for all bigram tokens included in the search. There are a total of 238,234 valid noun-initial and verb-initial bigram tokens in the corpus. We would like to divide this set of tokens in half, but due to the relatively small size of the corpus, the midpoint frequency is quite

[40] Moreover, note that *easily enough* and *highly verbal* both involve derivational affixes. There is an additional case to be made that derivational affixes should in fact be excluded from all analyses. Given that derivational affixes tend to be more tightly bound to stems (Bybee 1985), and given that derivational affixes and inflectional affixes do not interact in speech errors (MacKay 1979), it is likely that derivational affixation corresponds to a different psychological process than inflectional affixation.

[41] It is appropriate to include proper nouns in this search, given that the error set includes proper nouns (*EPL, Jerry*). Note that the organization of the Brown Tagset already excludes from the search non-affixable verbs and nouns, such as modals, forms of *be*, and pronouns, since these closed-class items are assigned distinctive tags.

low, and the exact midpoint does not lie between values. In this analysis, it turns out that more than half (57.8%) of the counted bigrams have Brown frequencies of 2 or less, and less than half (42.2%) have Brown frequencies of 3 or more.

Weighting the frequency classes appropriately, this means for a set of 20 bigrams, we expect a random distribution to be represented by 8.44 items in the high-frequency category, and 11.56 items in the low-frequency category. However, in fact, what we find is that the high- frequency category is overrepresented: there are 15 bigrams in the high-frequency group, and 5 in the low-frequency group. The categories are listed in Tables 4.5a and 4.5b, together with the appropriate (part-of-speech restricted) Brown Corpus frequencies.

'High Frequency' bigrams: N = 15

Bigram	Brown frequency
come on	28
make sure	27
come back	26
come in	25
pay for	23
go home	16
come up	15
give us	13
get along	10
go ahead	6
go for	6
shut up	5
end up	4
phone rang	4
add up	4

TABLE 4.5A. High-frequency verb- or noun-initial error bigrams in the Brown Corpus

'Low Frequency' bigrams: N = 5

Bigram	Brown frequency
forgot about	1
go get	1
ride(Noun) *home*	0
Jerry Pancake	0
EPL tend	0

TABLE 4.5B. Low-frequency verb- or noun-initial error bigrams in the Brown Corpus

The distribution shown in Tables 4.5a and 4.5b in fact differs significantly from a random selection of verb-initial and noun-initial bigrams. A chi-square goodness of fit test gives a p-value of 0.0061. Focusing our analysis on the sample of verb-initial and noun-initial bigrams, outbound shifts are more likely to occur on high-frequency bigrams, even controlling for overall likelihood of occurrence. This analysis thus supports the prediction that high-frequency word sequences are more susceptible to outbound affix shift errors[42].

The results from the part-of-speech tagged analysis could perhaps be questioned on the grounds that the Brown Corpus is written, rather than spoken, and also rather small. The third analysis below will in part address these concerns. Moreover, this analysis will broaden the scope somewhat so as to include Mutual Dependency as an independent variable.

4.1.4. **Analysis 3: Frequency and Mutual Dependency in verb-initial sequences**.

In this final corpus analysis, I continue the notion of limiting valid bigrams, while extending to a somewhat larger composite spoken corpus. I also make a first attempt to

[42] Note that the significant result in this section's analysis, as opposed to the null result in Section 4.1.1, is not due simply to the exclusion of the two derivational errors on low-frequency sequences (*easy enoughly, high verbally*). For sake of illustration, we can in fact add these two low-frequency errors back into the set without losing significance (chi-square p = 0.0243).

include Mutual Dependency (MD) as a factor in assessing the distribution of affix shift errors. This is a particularly tricky task to attempt. It is not meaningful to evaluate an individual error in isolation; we must examine groups of errors, and be on the lookout for surprising asymmetries in distributions. As noted in 4.1.1, the collection of an error from naturalistic discourse must be assessed in the context of the frequency with which the utterance (in this case, a bigram) occurs. Moreover, the relationship between MD and frequency of occurrence is highly complex: as discussed in Chapter 2, MD is not independent of frequency, but it is not strictly correlated, either. It is possible for a sequence to be relatively low in frequency, but high in Mutual Dependency, as in cases where the component words of a sequence have restricted uses outside the sequence.

However, in the present section, I pursue an analysis based on an approximation of group frequency for categories we will call 'high MD' and 'low MD.' For our present purposes, the divisions between 'high' and 'low,' both for frequency and MD, will be to a certain extent arbitrary, but weighted with respect to overall frequency for the category.

The 'high' and 'low' bins to be used in this discussion are those devised for the experimental task of Section 4.2. In that section, the definition of the bins will be discussed in some detail, and justified on the basis of creating adequate opportunities for matches among experimental stimuli. For the present analysis, it suffices to note the following steps in defining these bigram sets.

(i.) The bins were defined by first searching the composite spoken corpus for bigrams that begin with one of the 250 most frequent verbs in English. Since the corpus is not tagged, this filtering was performed on the basis of matching of wordforms, rather than checking word classes.

(ii.) For each bigram included in the search, Token Frequency and Mutual Dependency were calculated from the composite spoken corpus. The bigrams were sorted for each of these corpus measures. 'High' Token Frequency was defined as the top 10% of all bigram token frequencies, and 'high' MD was defined as the top 10% of all bigram MD values.

(iii.) The resulting bins were checked against bigram searches in the COCA Spoken corpus (95 million words), allowing high and low category definitions to emerge in a larger corpus.

Since the bigram groupings in this corpus analysis are based on verb forms, it is appropriate to restrict the set of relevant bigram errors to those that begin with a verb. This then further reduces our set of analyzable bigrams to 16. The breakdown of the 16 errors according to 'High' and 'Low' Token Frequency and MD categories is given in Table 4.6 below.

A considerable amount of caution is needed in interpreting Table 4.6. Although I have already argued that naturalistic data must be interpreted in light of frequencies of occurrence, I will repeat that concern here in a more specific context. Because of differing baseline frequencies, it is not appropriate to make any pairwise comparisons of bins in the above 2 x 2 table. For instance, it is tempting to note that there are 13 errors in the High MD/High Frequency group, but 0 errors in the High MD/Low Frequency group. However, note that the High MD/Low Frequency category is a very select group, devised for purposes of the experiment in Section 4.2. This bin is very sparsely populated, both in terms of types (there are far fewer distinct bigram types here than in the High MD/High Frequency bin), and in terms of tokens (naturally so, because by definition, its bigrams

	LOW FREQUENCY			HIGH FREQUENCY		
	Bigram	**Freq**	**MD**	**Bigram**	**Freq**	**MD**
HIGH MD				*go ahead*	10314	21.229
				make sure	10647	21.010
				come back	16991	20.735
				end up	4385	18.532
				give us	5442	18.187
				come up	7074	17.617
				pay for	4689	17.083
				come on	8795	16.716
				go home	2314	15.849
				shut up	774	15.802
				come in	6646	14.696
				add up	665	14.586
				get along	1012	14.457
LOW MD	*forgot about*	137	11.968	*go for*	2278	11.973
				go get	746	10.554

TABLE 4.6. Conversational bigram errors, with Frequency and Mutual Dependency values based on COCA Spoken (95 million words, Davies 2008-).

are low in Token Frequency). Thus, a scan through the composite spoken corpus reveals that as a group, tokens from the High MD/High Token Frequency bin are around 115 times as likely to occur in speech as tokens from the High MD/Low Token Frequency bin. Thus the asymmetry in naturalistic errors between these two bins is not in itself surprising.

That being said, we may now pursue approximate, pairwise comparisons with respect to High and Low groupings of Token Frequency and Mutual Dependency independently. For this analysis, I focus on the compiled corpus metrics used in defining High and Low categories on the basis of the composite spoken corpus. With respect to Token Frequency, we can sum across all the verb-initial bigrams in the High category, and find a total number of 141,528 tokens in the corpus. The same analysis yields 72,204 tokens in the Low-Frequency group. Noting from Table 4.6 that there are 15 High-

Frequency errors and 1 Low-frequency error, this affords the following approximation of error rates for the two groups, computed by dividing the number of errors in the group by the group frequency (cf. Stemberger and MacWhinney 1986b).

	No. errors in sample	Group Frequency	Error Rate
HIGH FREQ	15	141528	.0106%
LOW FREQ	1	72204	.0014%

TABLE 4.7. Comparison of outbound shift rates for High- and Low-frequency categories (Composite Spoken Corpus analysis).

Thus, if we weight categories with respect to baseline frequencies, we find that the outbound shift rate is 7.57 times higher in the High Token Frequency group than in the Low Token Frequency group. Based on the overall frequencies of bigrams in each group, we may estimate expected frequencies of errors of 10.595 for high-frequency bigrams, and 4.405 for low-frequency bigrams. Comparing with the observed distribution of the 16 errors in Table 4.7 allows us to compute a chi-squared statistic (4.422, 1 df) which is significant ($p = 0.0199$). As in the analysis in Section 4.1.3, then, this corpus analysis provides evidence that high-frequency word sequences are more susceptible to outbound affix shift errors.

We may perform a similar analysis with respect to Mutual Dependency, by summing frequencies across all items in the High and Low MD categories. This analysis yields the error rates in Table 4.8.

	No. errors in sample	Group Frequency	Error Rate
HIGH MD	13	118247	0.0109%
LOW MD	3	95485	0.0031%

TABLE 4.8. Comparison of outbound shift rates for High- and Low-Mutual Dependency categories (Composite Spoken Corpus analysis).

Correcting for expected frequencies, we find that high Mutual Dependency is also associated with an increased rate of outbound affix shifts. The values in Table 4.8 indicate that the outbound shift rate is 3.52 times higher in the High MD category than in the Low MD category. Based on the estimated frequencies of bigrams in each group, with respect to the 16 errors we arrive at expected frequencies of 8.852 for high-MD bigrams, and 7.148 for low-MD bigrams. A chi-squared test confirms that the overrepresentation of affix errors actually observed among the high-MD bigrams is statistically significant ($p = 0.037$, chi-square = 4.351, 1 df).

In sum, then, frequency-weighted analyses of verb-initial bigrams in the composite spoken corpus indicate that outbound affix shifts are more likely on sequences that are high in Token Frequency, or high in Mutual Dependency.

4.1.5. **Analysis 4: Comparison of early vs. late affix shifts.**

Note that in the foregoing analyses, we have essentially been comparing the bigrams involved in outbound affix shifts against an entire corpus of bigrams. As one final analysis of naturalistic affix shift errors, we might instead compare outbound affix shifts with a more selective group of bigrams, specifically, bigrams in which a suffix occurs prior to its target.

Recall that in the study by Stemberger and MacWhinney (1986b), all affix shifts (both early and late) were pooled together for purposes of studying the retrieval of affixed forms from memory. However, the hypothesis of the current chapter is that outbound shifts and inbound shifts arise from different mechanisms, and thus multiword sequences should have rather different characteristics in these two error sets. More precisely, as

discussed in Section 4.0, we cannot presume a priori that affix shifts always involve linguistic units. Indeed, it may be that affixes arrive early in an utterance precisely because no multiword chunks prevent this shift from happening. Thus in more general terms, we would predict that there will be striking differences between sequences with early affix shifts and sequences with late affix shifts.

To investigate this hypothesis, we may consider the set of naturalistic, early affix shifts available from the Fromkin Speech Error Database. There are 11 such errors in the database; 8 of these involve two-word sequences, and can meaningfully be compared with the 22 outbound affix shift bigrams.[43]

On the following page, in Table 4.9, I again list the outbound (late) affix shift errors encountered previously, but this time present them alongside the early affix shifts and their corpus measures. Inspection of this table reveals that the outbound affix shifts are characterized by quite different metrics than are the early shift errors. In the COCA spoken corpus, the late shift bigrams have an average frequency of 3779.5, compared with 235.0 for the early shift bigrams. Similarly, late shift bigrams have a higher average Mutual Dependency (13.597) compared with the early shift bigrams (6.278). These cross-category differences are, moreover, statistically significant. The corpus metrics of bigrams are not normally distributed[44], so parametric tests would not be appropriate. Instead, we can use the Mann-Whitney U test (Mann and Whitney 1947), a rank-based test for ordinal data which makes no assumptions about the sizes of samples or shapes of

[43] One of the early affix shift errors is unquestionably an 'inbound' shift, since it actually involves two word roots that occur inside a single, compound word: *print outs* → *prints out*. For sake of comparison, I have treated *print out* as a two-word 'bigram.' To be conservative in the corpus analysis, I have counted compound occurrences of *printout* (or *print-out*), in addition to occurrences where *print* and *out* occur as successive words.

[44] Distributions of n-grams are approximately Zipfian: the *i*th most frequent bigram in a corpus has a frequency that is inversely proportional to *i*. See Manning and Schütze (1999: 213-214).

LATE (OUTBOUND) SHIFTS

Base Bigram	Frequency (COCA Spoken)	MD (COCA Spoken)
come up	7074	17.617
go ahead	10314	21.229
make sure	10647	21.010
end up	4385	18.532
come in	6646	14.696
pay for	4689	17.083
come back	16991	20.735
come on	8795	16.716
go for	2278	11.973
give us	5442	18.187
go get	746	10.554
go home	2314	15.849
get along	1012	14.457
add up	665	14.586
forgot about	137	11.968
shut up	774	15.802
easy enough	48	9.293
phone rang	145	17.943
ride (N) *home*	47	10.898
Jerry Pancake	0	0*
EPL tend	0	0*
high verbal	0	0*
AVERAGE	**3779.50**	**13.597**

EARLY SHIFTS

Error	Base Bigram	Frequency (COCA Spoken)	MD (COCA Spoken)
prints out	*printout*	72	10.040
transducers array	*transducer array*	0	0*
veryest high	*very high*	1618	15.223
build's one	*build one*	52	6.920
workings paper	*working paper*	12	6.167
Joes tell	*Joe tell*	1	0**
keeping suggest	*keep suggest*	5***	3.834
quites get	*quite get*	120	8.042
AVERAGE		**235.0**	**6.278**

TABLE 4.9. Comparison of outbound affix shifts (n = 22) and early affix shifts (n = 8) from conversation. Corpus metrics refer to the <u>base bigram</u> in both parts of the table.

*Mutual Dependency is defined here as a logarithm, which means that the MD of a sequence with frequency zero is mathematically undefined. For sake of comparison, I have assigned an MD score of zero in such cases.

**In this particular case, the equation actually generates a negative MD score (-2.94). Given that items with <u>zero</u> frequency are assigned an MD of zero (see above), it is reasonable to say that in the present analysis, zero should be the minimum allowable score; thus I have set the MD for *Joe tell* to zero.

***The sequence *keep suggest* in fact never occurs in COCA, as one might expect. However, on the chance that the higher frequency of *keep suggesting* is relevant in the present case, I have reported this value (5) here in order to bias the results <u>against</u> my predictions.

distributions. With respect to Token Frequency, a two-tailed Mann-Whitney test yields $U = 36.5$, $p = 0.0168$. For Mutual Dependency, a two-tailed Mann-Whitney test yields $U = 30$, $p < 0.01$.

This result indicates that early and late shift bigrams are significantly different with respect to both corpus measures. Such a finding is indeed what we would expect to see: bigrams which are highly frequent or high in MD are more likely to be retrieved as units, and more likely to result in repositioning a suffix to the periphery. On the other hand, multiword sequences which can be interrupted by a stray affix, as occurs in early shifts, would be expected to be less cohesive. The present corpus analysis verifies that as a group, the word sequences associated with early affix shifts co-occur less often than the word sequences associated with late (outbound) affix shifts.

4.2. **Experimental study of affix positioning errors**.

The foregoing results offer encouraging data in support of the hypothesis that outbound affix shift errors are more likely when words frequently co-occur in usage, whether this co-occurrence is measured using Mutual Dependency or Token Frequency. However, speech errors collected from conversational settings may always be challenged on the grounds that the data are subject to investigators' perceptual limits and biases (Cutler 1981). Ideally, evidence for psycholinguistic phenomena will be drawn from complementary sources, including data from the naturalistic (but uncontrolled) setting of conversations, and from the controlled (but artificial) setting of the laboratory (Stemberger 1992). Experimental investigation of the distribution of affix shifts is therefore appropriate. In this section, I describe an experimental task designed to elicit outbound affix shifts. In addition to seeking additional evidence regarding holistic

retrieval, this study also affords exploration of a novel methodology for observing affix positioning errors.

4.2.1 **Task design**.

The experimental task is designed to elicit affix positioning errors by requiring participants to produce verbal responses as rapidly as possible, expanding upon methodologies used previously by Bybee and Slobin (1982) and Stemberger and MacWhinney (1986a). In the task, participants are instructed to listen to a recorded stimulus sentence, and to repeat back a modified version of the sentence which requires adding the 3rd person singular marker *–s* on the verb. Each sentence includes the pronoun subject *they*; female participants are asked to substitute the pronoun subject *she*, and male participants are asked to substitute the pronoun subject *he*. Thus, for instance, if the stimulus sentence is *Despite the ads about switching to green energy, they depend on contributions from the coal industry*, a correct response would be *Despite the ads about switching to green energy, she depends on contributions from the coal industry*.

The 3rd singular suffix was chosen as the relevant affix in this study for a number of reasons. First, this affix was noted to be the most common one involved in outbound affix shifts in conversation; 12 out of the 26 outbound shifts in Tables 4.1 and 4.2 involve the placement of *-s*. With respect to experiment design, the 3rd singular suffix also has advantages over possible alternatives. For instance, unlike the progressive *–ing* suffix, the 3rd singular marker can be inserted without the addition of auxiliary verbs (e.g., *is running*), which would introduce additional complicating factors from auxiliary errors. Moreover, the 3rd singular marker regularly attaches only to words, not phrases, allowing us to draw clear inferences from how speakers position the affix. This is in contrast with

the possessive marker, which freely attaches to full phrases (e.g., *the cat in the hat's pajamas*), and which thus would provide ambiguous evidence about lexical units (Pinker 1999). Finally, the 3rd singular marker is highly regular, allowing for a wide variety of stimulus verbs that require the insertion of an affix, rather than changes to the stem.

As noted in Section 4.1, outbound affix shifts are quite rare in conversation, and thus the experiment is designed with additional distracting factors in the hopes of increasing the error rate. Thus, as one complication, participants are asked to 'shadow' throughout the course of the experiment (Marslen-Wilson 1973). That is, they are asked to begin echoing back the stimulus sentence immediately, without waiting for the stimulus presentation to end, thus requiring simultaneous listening and speaking during most of the participant's response. Speech shadowing provides one method of overloading verbal capacities, thus providing ongoing interference with participants' abilities to use language introspectively (Hermer-Vasquez et al. 1999), and prompting more automatic, less carefully analyzed speech output. Levelt (1989) argues that speakers monitor their own covert and overt speech, checking for well-formedness of the intended message. Similarly, Laver (1973) hypothesizes that a 'Monitor' component of speech production is constantly on the lookout for errors, and in most cases, manages to correct errors that do occur. Laver (1973) further observes that the Monitor can be impaired under various conditions, including situations in which there are 'competing demands for attention' (140). In the present experiment, then, the additional requirement of shadowing is intended to minimize the speaker's resources for monitoring his or her own planned speech output.

A second distracting element in the experimental task is based on the Competing Plans Hypothesis of Baars (1980, 1992), which suggests that speakers often formulate alternate, parallel plans for an utterance, and competition between these plans sometimes results in errors. A wide range of experiments have been developed in which errors are elicited by 'creating competition between alternative output plans,' either in language or in other motor activities (Baars 1992: 130). Often, such approaches encourage errors (such as spoonerisms) by alternating unpredictably between the type of response required from participants. In the present experiment, we seek to increase the likelihood that participants will, with respect to affixation, chunk multiple orthographic words together. Thus, the experiment is designed so as to intermingle bigram stimuli with distractor items which contain more than one bound root, that is, compound verbs. In the bigrams of interest, any inserted affixes will be expected on the verb inside the multiword sequence, as in *gets out*, *depends on*, and *sees fit*. The Competing Plans Hypothesis predicts that we may encourage syntagmatic errors by priming an alternate affixation strategy, in which roots inside a single lexical item must be passed over: *sleep_walks*, *safe_guards*, *play_acts*, and so on. This alternate strategy may be especially influential in cases where the first component of the compound verb may be parsed as a verb (*sleep-walk*, *hang-glide*, *dry-clean*, etc.). The selection of compound verb distractors will be discussed in more detail below.

4.2.2 **Materials and Stimulus design**.

4.2.2.1. **Frequency x Mutual Dependency bins.**

The present affix shift experiment is designed to investigate possible effects from Token Frequency and Mutual Dependency, both separately and together. Thus stimuli are

selected in a 2 x 2 design, consisting of High/Low Mutual Dependency, crossed with High/Low Token Frequency. For purposes of uniformity and simplicity in the design (see Chapter 3), I focus on two-word sequences, i.e., bigrams.

An extensive automated search was undertaken in order to identify a range of suitable candidates. This first step involved identifying, counting, and sorting all appropriate bigrams from the composite spoken corpus (5 million words; see Table 4. 3). For purposes of this affix shift study, the bigrams of interest all begin with a verb. Thus, I wrote a script in Java to scan through the composite corpus, tallying and cross-sorting any bigram beginning with one of the 250 most frequent English verbs, based on a part-of-speech search in COCA (Davies 2008). The list of acceptable verbs excluded forms of *be*, modals, and other verbs which cannot receive a 3rd person *–s* suffix.

For each bigram collected from the corpus, then, the Java script stores a total Token Frequency value and Mutual Dependency score. Mutual Dependency is defined as in Equation 2.4 from Chapter 2, where the size of the n-gram is equal to 2. For convenience, I repeat the definition here (noting that N is the corpus size).

(Equation 4.1) $MD(w_1w_2) = \log_2 \left[\frac{N^* f(w_1w_2)^2}{f(w_1)^* f(w_2)} \right]$

The full collection of bigrams is sorted in two separate lists, according to Token Frequency, and according to Mutual Dependency. Based on these sorts, Token Frequency and Mutual Dependency are each divided into High and Low categories, through a combination of pragmatic and partially arbitrary criteria. To define a high-frequency (or high-MD) class with potentially idiosyncratic features, it may prove helpful to skew the thresholds somewhat toward the high end of the scale (see Gordon and Alegre 1999,

Kapatsinski and Radicke 2009). On the other hand, it was necessary to keep the 'High' categories large enough to allow for an adequate range of candidates to be available in each 2 X 2 bin, for purposes of matching features between bigrams (as described below). Through a process of trial-and-error, I arrived at a division in which 'High Token Frequency' and 'High Mutual Dependency' bigrams are each defined as the top 10% of all types on the appropriate scale. This split then defines High Frequency bigrams as those having a frequency of 28 or more, and High Mutual Dependency bigrams are those having an MD value of 9.18 or more, based on the composite spoken corpus.

These corpus metric classifications were subjected to an additional step in which they were checked in a second, larger corpus. The purpose of this second analysis was to ensure that bigrams classified in High or Low bins truly exhibit similar patterns in a range of contexts, thus avoiding illusory effects arising in a relatively small corpus. The additional corpus I chose was the 95-million word spoken portion of COCA (Davies 2008). Limiting searches to the spoken portion of COCA has the benefit of making the data more spontaneous and naturalistic (although parts of the data are from scripted news broadcasts). Moreover, focusing on the spoken portion helps to avoid an over-representation of academic English, since the composite corpus already contains approximately 35% academic speech (1.8 million words from MICASE).

The full COCA corpus is not downloadable, and thus cannot be exhaustively analyzed in the same way as the combined Switchboard/MICASE/Santa Barbara corpora. Thus, throughout the stimulus selection process, I looked up a wide range of individual candidate bigrams in the COCA spoken corpus. Since the COCA interface allows for searches to be constrained by part of speech, I limited each search to instances in which

the bigram's first word was classified as a verb. This restriction added an additional safeguard insofar as some verb frequencies were quite low, and subject to erroneous classification from matches based on the wrong part of speech (such as noun instances of *fall*, *freak*, *take*, and so on). As a result of these searches, ad hoc category divisions gradually emerged for the COCA spoken corpus: High Frequency bigrams are defined as those having a frequency of 505 or more, and High Mutual Dependency bigrams are those having an MD value of 13.6 or more.

In identifying potential candidates, in most cases there was agreement between the classifications based on the composite spoken corpus and the spoken COCA corpus. I generally discarded items which were not classified into the same bin by the two corpus analyses. However, for two items (*cut out* and *fit in*, noted in the final table below), I was obliged to ignore disagreement between the corpora, due to a paucity of suitable cross-category matches in the High Frequency, Low MD bin. In these two cases, I used the classifications from the COCA spoken corpus, thus overriding the classifications from the smaller (and presumably less reliable) composite spoken corpus.

The process of identifying matches across the 2 X 2 categories involved a careful, hand-selected search for items which were controlled for numerous features. This approach should be contrasted with that of Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), in which items were randomly selected from within statistically-defined bins, without attempting to match for features such as constituency. Ellis et al. (2008) observe that the statistical coherence of high-Mutual Information sequences 'tends to correspond with distinctive function or meaning as well as grammatical well-formedness

as a complete phrase' (380). This observation is quite probably true, on average.[45] However, it is also true that all four bins as defined for this experiment contain some items which are idiomatic, and others which are semantically transparent; all four bins contain some items which cross constituent boundaries, and others which do not. Thus, to the extent possible I sought bigram stimuli which controlled for numerous features, in an attempt to avoid uncontrolled biases in favor of particular bins. The guidelines I used are listed below.

4.2.2.2. **Bigram features matched across bins.**

1. Stimuli across categories were matched with respect to the part of speech of the word following the verb. This approach helped to encourage cross-category uniformity for constituency of the sequence, given that similar sequences (for instance, Verb + Preposition or Verb + Adverb) will tend to have broadly similar structural features. Moreover, this approach helped generate many candidate matches during stimulus selection, which was generally accomplished by scanning or searching the 2 x 2 lists of bigrams created by the Java script. An attempt was made to include a range of structures in the searches (Verb + Pronoun, Verb + Mass Noun), with a special emphasis on the types of patterns known to occur in conversational errors (such as Verb + Preposition and Verb + Adverb; see Tables 4.1 and 4.2).

[45] Indeed, I can report that finding appropriate matches across categories was near-impossible in certain cases, which speaks to the general validity of the observation that there are cross-category differences. However, the difficulties in finding matches arose not because a particular category lacked cohesive bigrams, but because the cohesive bigrams were not of the appropriate type. As one example, it was challenging to find bigrams containing prepositions in the high MD, low Frequency bin—an unsurprising fact, because prepositions are high in frequency, and thus most low-frequency sequences containing a preposition are also low in MD.

2. Additional heuristics were used to ensure that structural features were uniform across all four bins. Beyond broad similarities imposed by matching part-of-speech sequences, it is possible different bigrams will have varying grammatical relationships with the surrounding text, or will have varying degrees of morphosyntactic fixedness. Thus, items in the four bins were matched in categories including the following, using grammatical descriptions of multi-word verb categories by Quirk et al. (1985: 1152-1161).

a. Type I (intransitive) phrasal verbs (*wake up, settle down, get down*). Verb + Adverbial sequence, with no noun object.

b. Type II (transitive) phrasal verbs (*seek out, figure out, tear apart*). Verb + Adverbial sequence which requires a noun object. When the object is a pronoun, it intervenes between the Verb and the Adverb: *seek it out*.

c. Prepositional verbs (*worry about, arrive at, come with, fear for*). Verb + Preposition sequences, requiring a noun phrase as an object of the preposition. The noun phrase groups syntactically with the preposition (*arrive* [*at the station*]), and thus in a traditional syntactic analysis, such sequences cross a constituent boundary. In prepositional verbs, pronouns do not intervene between the verb and preposition (**arrive it at*).

It should be noted that the same word sequence may fall into different categories depending on the context. For instance, *wake up* may be intransitive (*They wake up late*) or transitive (*They wake up <u>the children</u>/They wake <u>them</u> up*) (cf. Quirk et al. 1985: 1158). Where such variation is possible, I was careful to select a stimulus sentence which honored the presumed structural features of each bigram compared with its matched counterparts in other bins.

3. In all four bins, an attempt was made to include bigrams that exemplify a range of idiomaticity. It should be noted that Verb + Particle sequences quite typically exhibit some degree of semantic opacity on the particle. For instance, V + *up* phrasal verbs fall into a variety of groups including completing and finishing, approaching, or beginning, none of which are transparently related to the spatial/movement meanings of 'up' (see Sinclair 1989: 487-488). However, a number of bigrams selected are especially idiomatic, insofar as the two words collectively have a meaning that is unrelated to the literal meaning of the verb on its own. For these more idiomatic cases, matching cases of idiomatic or metaphorical use were found for all four bins.

As was also true for structural features, verb bigrams may exhibit a range of idiomaticity due to polysemy. Some uses of the same sequence may be more idiomatic (*We* [*work out*] [*at the gym*]) than other uses (*We work* [*out in the sun*]). To deal with such cases, in the selection of stimulus sentences, idiomatic uses were matched across categories with idiomatic uses.

4.2.2.3. **Additional requirements on bigram stimuli.**

In addition to the above cross-category matching requirements, the following general requirements restricted the bigram stimuli selected.

1. To allow for uniformity in the pronouns used in the stimulus sentences, verb bigrams (and the verbs themselves) needed to have natural-sounding uses with a human third-person subject (*they, he, she*).

2. Bigram stimuli could not be part of a larger unit which is highly formulaic and predictable. I established a largely arbitrary threshold limiting acceptable bigrams to those that allowed variation in the following word at least 40% of the time, based on searches in the COCA Spoken corpus. For instance, the bigram *feel free* occurs 215 times in this corpus. Its most frequent following word is *to*; the trigram *feel free to* occurs 156 times in the corpus, constituting 72% of all instances of *feel free*. Thus, *feel free* would be disqualified as a potential bigram stimulus, out of concern that the real unit of interest in this case would be the three-word sequence, *feel free to*.

3. Stimuli could not include any verbs that end in a sibilant, so that there would be no cases in which participants had to insert [əz] rather than [s] or [z] (*they miss it > he misses it*). It proved impossible to find four matching candidates consistently so that *–es* was the appropriate allomorph to insert, and it was feared that requiring an additional syllable for the affix could influence the likelihood of an affix shift error. Thus for the sake of uniformity, I altogether ruled out sibilant-final verbs.

4. Finally, once a bigram was added to the stimulus candidate list, the verb in this bigram was not allowed in any additional stimuli. I avoided re-using the same verb, for fear that word-specific priming effects would play a role. More specifically, I was concerned that once a participant correctly inflected a verb (*fall off > falls off*), priming of the inflected form could make subsequent errors on this verb less likely.

There was one controlled exception to the rule that verbs could not be re-used. Most of the stimuli were grouped together and presented randomly during the course of

the experiment. However, following this main experiment block, there was a ‘bonus round,’ in which four verbs were used a second time: *walk*, *look*, *pay*, and *move*. The purpose of the bonus round was to expand the stimulus set by four items, given that the ‘no re-use’ requirement made finding matching items increasingly challenging. By including the bonus stimuli at the end of the experiment, and by matching conditions across the four bins, I effectively balanced out any priming effects. These bonus sentences were intermixed with distractor sentences, and randomized separately.

4.2.2.4. **Listing of bigram stimuli.**

The above selection criteria were used to select a total of 56 bigram stimuli. There were fourteen matched bigrams across the four bins (including the one ‘bonus’ item per bin, which re-used a previously-used verb). The items used are presented in Table 4.10.

The stimulus bigrams were used to construct 56 stimulus sentences. Sentences were loosely based on usages found in the COCA corpus. Sentences were matched in groups of four, with respect to register, and often semantic domain, in order to prevent some sentences from being less accessible than others. The sentences were all approximately matched for length, measured in number of syllables. The verb bigrams of interest were always positioned close to the middle of the sentence, in order to maximize the cognitive demands (from simultaneous verbal listening, remembering, and speaking) on the participant at the time he or she utters the inflected form of the verb. Sentences were all similar in syntactic structure and complexity, with the verb bigram occurring shortly after an introductory, dependent clause. The sentences containing these bigrams are listed in Appendix 4.1.

	LOW FREQUENCY				HIGH FREQUENCY			
	bigram	**Category**	**Freq**	**MD**	**bigram**	**Category**	**Freq**	**MD**
HIGH MD	1. settle down	I Phrasal	204	14.63	1. wake up	I Phrasal	1865	19.28
	2. screw up*	I Phrasal*	119	14.20	2. work out *	I Phrasal*	1839	15.01
	3. freak out*	I Phrasal*	60	13.60	3. hang out*	I Phrasal*	695	15.61
	4. wrap up*	T Phrasal*	276	14.97	4. add up*	T Phrasal*	665	14.59
	5. tear apart	T Phrasal	21	13.76	5. figure out	T Phrasal	4621	20.18
	6. read aloud	V Mod	24	14.36	6.make sure	V Mod	10647	21.01
	7. gain weight	V MassN	93	16.39	7. stay home	V MassN	546	14.59
	8. interfere with	V Prep	383	15.51	8. depend on	V Prep	1103	16.85
	9. fall off	V Prep	211	13.93	9. pay for	V Prep	4689	17.08
	10. arrive at	V Prep	363	14.26	10. worry about	V Prep	3974	19.07
	11. insist on	V Prep	378	13.77	11. talk about	V Prep	28166	21.64
	12. trust me	V Pro	504	14.61	12. call it	V Pro	4677	15.72
	13. recover from	V X (V)	311	14.30	13 need to	V X (V)	28042	19.03
	B. walk through	V Prep	393	14.61	B: look at	V Prep	32791	21.65
LOW MD	1. move up	I Phrasal	214	10.10	1. get down	I Phrasal	986	12.24
	2. give in*	I Phrasal*	296	6.67	2. fit in*+	I Phrasal*	568	12.70
	3. hold off*	I Phrasal*	204	12.28	3. let go*	I Phrasal*	749	11.45
	4. leave out*	T Phrasal*	132	8.42	4. take on*	T Phrasal*	2372	12.96
	5. seek out	T Phrasal	175	12.10	5. cut out+	T Phrasal	567	12.87
	6. smell bad	V Mod	11	8.61	6. look good	V Mod	833	12.59
	7. buy food	V MassN	69	11.10	7 see people	V MassN	708	10.08
	8. walk at	V Prep	22	3.71	8. point to	V Prep	824	12.20
	9. speak in	V Prep	254	8.30	9. know of	V Prep	1631	8.26
	10. fear for	V Prep	136	9.84	10. come with	V Prep	813	10.15
	11. run after	V Prep	35	5.93	11. agree on	V Prep	1330	13.57
	12. offer it	V Pro	84	6.79	12. forget it	V Pro	618	11.79
	13, resolve to	V X (V)	33	4.60	13. hope to	V X (V)	1773	12.51
	B. pay at	V Prep	53	5.01	B. move to	V Prep	1242	11.70

TABLE 4.10. Stimulus bigrams used in the elicitation experiment. Frequency and MD values are based on COCA Spoken data (Davies 2008, 95 million words).

*Matching items which are especially idiomatic (in the stimulus sentences selected).

+Items which were classified as High Frequency, High MD in the composite spoken corpus, but grouped here with High Frequency, Low MD on the basis of part-of-speech constrained counts in the COCA spoken corpus. Even in the composite spoken corpus, these items do have markedly lower MD values than their counterparts in the High MD, High MD bin (*fit in* is lower than *work out*; *cut out* is lower than *figure out*).

4.2.2.5. **Compound distractors.**

As discussed in Section 4.2.1, the distractors in the experiment are verbs which contain multiple lexical roots, that is to say, compound verbs. To help increase the effects of interference from the distractors, it was desirable to locate many compound verbs

whose initial root is a verb. To assist with this process, I searched through numerous lists of English compound verbs in references on word-formation (Marchand 1966, Adams 1976, Cannon 1987). Additionally, I performed searches in the Oxford English Dictionary online, focusing on verbs with 'backformation' listed in the etymology.[46] These searches led to the 56 compound verbs listed in Table 4.11. Asterisks indicate compounds in which the first root is a verb (including roots which can function as a verb, even if this usage is not primary, or if it arises from homonymy).

*test-drive**	*strongarm*	*double-check**	*moonlight*
hotwire	*deepfry*	*bearhug**	*blackmail*
*timetravel**	*fine-tune**	*tie-dye**	*earmark*
*bookmark**	*mastermind**	*overhear*	*copyright**
*spoonfeed**	*freeload**	*leapfrog**	*fundraise**
cherry-pick	*brainstorm*	*zigzag**	*shoplift**
sidestep	*underestimate*	*bench press*	*globetrot*
safeguard	*daydream*	*windsurf*	*flyfish**
*jumpstart**	*wisecrack*	*dryclean**	*waterski*
*handwrite**	*badmouth*	*blowdry**	*backtrack**
*jam-pack**	*panhandle**	*fireproof**	*sleepwalk**
*spotlight**	*house-sit*	*wallpaper**	*hang-glide**
skyrocket	*mass-produce*	*forcefeed**	*freezedry**
*bankroll**	*babysit*	*play-act**	*proofread**

TABLE 4.11. Compound verb distractors used in experiment.

Since the distribution of errors on compounds was not of primary interest to this study, there was no restriction on compounds which end in a sibilant (*flyfish*) or which have an initial root ending in a sibilant (*house-sit*).

The 56 compound verbs were used to construct 56 distractor sentences, matching the semantic domains for the stimulus sentences. These sentences are listed in Appendix 4.2.

[46] Backformation from gerund compounds represents a common path whereby compound verbs enter English, including verb-initial compounds, as in *sleep-walking* > *sleep-walk* (V) (Adams 1976).

4.2.3. **Participants and experiment setup**.

The 56 stimulus sentences and 56 distractor sentences were recorded by a female native speaker of English, who was instructed to read them aloud at a normal, casual rate of speech. The sentence audio files were presented to participants using E-Prime, with randomization as follows. There were 52 stimulus sentences and 52 distractor sentences which constituted the main experiment trial; these were presented in random order to each participant. In addition, there were 4 stimulus sentences and 4 distractor sentences (the 'bonus round') which were randomized separately and presented to participants after the main experiment block was completed. (See Appendices 4.1 and 4.2.)

It was noted during pilot testing that participants often forgot to change the subject from *they* to *he* or *she,* and often mistakenly changed the verb to the past tense. Failing to replace the pronoun would result in loss of data. To a lesser extent, changing to the past tense would also result in lost data, because many of the verbs used among the stimuli are irregular in the past (requiring either no change, or a change to the verb stem rather than an affix). Thus during the instructional phase, participants were explicitly reminded of these pitfalls.

The instructions to participants were as follows:

1. In this experiment, you will hear a series of sentences in the headphones. For each sentence, you will be asked to speak aloud a variation of the original sentence where you have substituted the pronoun *she* [*he*] for the word THEY as the subject of the sentence. For instance, if you hear the sentence *Using some old reel-to-reel equipment, they tape-record the ensemble's performance,* you would respond with: *Using some old reel-to-reel equipment, she* [*he*] *tape-records the ensemble's performance.*

2. Please note that all of the sentences you hear are in the present tense. Some participants are tempted to change the verb to the past tense, but please keep them in the

present. Also note that in every sentence, the word *they* only appears once, and this is the only time you have to insert *she* [*he*].

3. This task will be more difficult than you might expect, because you should begin your spoken response immediately after the audio begins. That is, you will be listening and speaking simultaneously, that is, 'echoing' the sentence you hear, plus making changes to insert the word *she* [*he*]. In order to keep up, you will need to start speaking shortly after the sentence begins. Please know that the task in this study is meant to be pretty challenging, and you are not being 'evaluated' on how good your performance is. Variation in responses is expected, and you should not be excessively concerned if you feel you made a mistake. Please just do the best you can and proceed with the experiment. All that being said, please do your best to 'echo' word-for-word as much of the whole sentence as possible. If you forget part of the sentence, please just repeat back as much as you can in the allotted time.

4. There will only be a short break between successive sentences. Please attempt to respond with your modified version of each of them as quickly as possible, and complete each one before the next sentence begins. Once you hear a low tone, that means you are no longer being recorded, and you should prepare to respond to the next sentence.

Before the experiment began, participants were given six practice sentences to get them accustomed to the echoing and substitution task. The 6 practice sentences are listed in Appendix 4.3. Three of the six practice sentences contain a compound main verb (*pickpocket*, *spearhead*, *breakdance*), matching the distribution of patterns in the main experiment. Since participants had to resist a tendency to make errors such as *she breaksdance(s)*, the inclusion of compound verbs among the practice items helped to introduce interference from the alternate affix-insertion strategy as early as possible.

During the experiment, participants listened to the stimulus sentences on headphones, and gave vocal responses into a digital microphone attached at the collar. Vocal responses were saved as a collection of separate audio files by E-Prime. The main part of the experiment took approximately 20 minutes for participants to complete. A short break was inserted into the middle of the trial; since the experiment required rather rapid speech for 20 minutes, it was felt that a self-timed break would allow participants to

rest their voices, and prevent loss of data due to fatigue. Thus, halfway through the experiment, a screen appeared telling participants that the task was half-completed, and they could resume with a key press whenever ready.

Volunteer participants were recruited from the university's Introduction to the Study of Language course, and received a small amount of course credit for participating. A total of 29 participants enrolled in the study. Out of this group, 27 participants reported that English was their first language. The remaining 2 participants reported that they learned English before the age of 6, speak English fluently, and use it daily. No participants reported a history of speech or hearing disorders.

4.2.4. **Results and Discussion: Affix shifts and other affixation errors on bigram stimuli.**

4.2.4.1. **Participant accuracy.**

The experiment trial encompassed a total of 1624 responses (56 stimuli X 29 participants). I listened to all participant responses and coded them as No Error, Affix-Shifted, Double-Marked, Zero-Marked, or Unclassifiable. Responses with 'No Error' were those that correctly positioned the *–s* suffix (*they arrive at > she arrives at*). I also included cases where the participant mistakenly converted the sentence to the past tense, but correctly positioned a past tense suffix (*they arrive at > she arrived at*).[47] Unclassifiable responses could arise for a number of reasons. First, there were cases in which the participant gave an insufficient response, such as might happen if the participant forgot that portion of the sentence, ran out of time, reworded the sentence,

[47] The crucial question was not whether the affix was correct, but whether or not participants positioned affix(es) correctly. Thus, I also considered responses correct if the participant produced a conglomeration of affixes (for instance, both *–s* and *–ed*) on a single verb, in the correct position. Examples of such affix conglomerations are provided below.

and/or misunderstood one or more words in the bigram of interest. Additionally, the data were unclassifiable when the participant's response rendered moot the positioning of affixes, as in cases where the participant forgot to change the subject pronoun, or produced an irregular past form (*they give in* > *she gave in*).

Experiment participants were highly attentive to the task, and were generally able to give sufficient responses in spite of the time pressure. Overall, 142 responses had to be rejected as Unclassifiable, meaning that 1482 participant responses (91%) were codable. The rejection rate was relatively stable across the four bins in the experiment (chi-square = 4.82, p = 0.185), indicating that further analysis based on comparing raw numbers of errors across categories is justifiable. (See Table 4.12.)

Stimulus bin	**Number of unclassifiable responses (out of 406 responses per bin)**
Low Freq, Low MD	41
High Freq, Low MD	41
Low Freq, High MD	35
High Freq, High MD	25
TOTAL	**142**

TABLE 4.12. Rejection of data across the four categories.

Additionally, participants' responses were, as expected, overwhelmingly accurate. On average, 90% of the responses (1462 of the 1624) were coded as having No Error. The least accurate participants (2 participants out of the 29 total) provided accurate responses 77% of the time, and were deemed to be sufficiently accurate to be retained in the study.

4.2.4.2. **Outbound shift errors, and double-marking errors.**

On the other hand, the methodology was rather successful at inducing the affix errors of interest in this study. The data contain 7 full outbound affix shifts; sentence (13) represents an example response from one participant.

(13) *Since there's not much else to do before it's time to go, she settle downs on the couch...before it's time to go*.

Additionally, the data contain 9 'double-marked' affix errors, in which a suffix appears on the verb as well as on the following word. A typical example is presented in sentence (14).

(14) *To reward the students for completing more tedious... assignments, she reads aloud***s** *. . . from children's books an hour each day*.

One of the 9 double-marked errors in this set is somewhat more complex, and requires some discussion:

(15) *Despite the ads about switching to green energy, she depends onned* ([dəpɛnzɑnd]) *contributions from this, from the coal industry*.

I have classified sentence (15) as a double-marking error, albeit a double-marking that involves two different suffixes. Apparently the speaker activated two distinct suffixes (3rd person singular *–s*, and the past *–ed*), and a failure to resolve competing plans for the sentence caused one affix to be applied to the verb (*depends*), and the second affix to be applied to the bigram (*dependson-ed*). Such an occurrence is not especially surprising, given overall patterns in participant responses. First, it was relatively common for participants to convert sentences into the past tense, in spite of reminders not to do so

during the experiment's instruction phase. Among the stimulus bigram responses, there were 14 responses that unambiguously shifted to past tense (11 regular *–ed* verbs, plus 3 irregular verbs that resulted in uncodable data regarding affix positioning). Additionally, the compound distractors generated 16 past tense responses (14 regular, and 1 irregular). Thus, it seems that participants may have needed to suppress an ongoing temptation to insert *–ed* rather than *–s*. Additionally, there were other participant responses which evinced the activation of multiple affixes, insofar as two affixes were appended to a single verb.

(16) *On the first day of class in the seminar for majors, she talksed about real-world . . .* [no further response from participant]

(17) *According to the case presented by the prosecusors* [← *prosecutor*], *she routinely shoplifts...ted for the thrill it brings.*

The processes at work in example (17) are perhaps ambiguous, since the final [t] of the verb is also repeated for unknown reasons. However, in example (16), it is quite clear that the speaker has appended *–s* and *–ed* in succession.[48] It is thus reasonable to believe that speakers at times retrieved two different affixes for the same verb. Moreover, in two responses other than (15), a participant produced both *–s* and *–ed* on a single complex verb.

(18) *Although the visit is intended to be leisurely, she jams-packed* [ɑɪ]- *each day with errands and projects.*[49]

[48] Although beyond the scope of this study, it may be interesting to consider whether in such errors, the order of the affixes reveals anything about how tightly bound different affixes are with the verb root (Bybee 1985). Although both (13) and (14) exhibit the pattern ROOT + *-s* + *-ed*, I also observed the opposite pattern during pilot testing. A pilot participant made two errors of this type: *calls it* → *call-eds it* ([kal.ɛdz]) and *moves to* → *moveds to* ([muvdz]).

[49] The [ɑɪ]- in sentence (18) is a false start apparently unrelated to the affix positioning error. It is either a lexical error (a partial production of the word *I*), or a phonological error (involving the wrong initial vowel for *each*), although neither type of error has a clear source in the surrounding context.

(19) *There's little chance of tough questions at the press conference, and she spooned feeds the official file policy to reporters*.

The responses on (18) and (19) occurred on compound distractors, and thus are not included among the 9 double-marked errors discussed in this subsection. However, the pattern evident in sentences (15), (18), and (19) (produced by three different participants) seems to be the same. In all three cases, the speaker activated multiple, competing suffixes, and after failing to resolve the conflict in time, produced two different affixes on different portions of the same complex verb (or verblike unit).

Combining the 7 full outbound shifts with the 9 double-marked errors, there were 16 errors of affix placement in the experiment results. Collectively, then, there were 16 errors out of 1624 attempts, yielding an error rate of just under 1% (0.98%). In other words, an affix positioning error occurred on 1 out of every 102 sentences attempted in the experiment. This error rate indicates that the experiment methodology is highly effective at eliciting affix positioning errors, given how rare these errors are in spontaneous conversation. For comparison, note that Deese (1984: 130) estimates that approximately one out of 100 sentences in conversation contains a speech error of any kind; this estimated frequency naturally includes phonological and lexical errors, which are far more numerous than morphosyntactic ones. Similarly, the catalog of 191 slips from the 170,000-word London-Lund Corpus indicates that speech errors occur approximately once every 890 words (Garnham et al. 1981). Again, the London-Lund speech errors are predominantly phonological and lexical in nature, and the Garnham et al. collection contains no errors of affix positioning. Thus the rate at which errors of interest are observed in the present experiment— equivalent to one affix positioning error

approximately every 2,000 words— indeed represents a drastic increase from baseline error rates in conversation.

With respect to the MD and Frequency bins in this study, the affix positioning errors were not symmetrically distributed. Since all four bins are represented by an equal number of stimuli, if the distribution were random we would expect the errors to be spread evenly across four categories. However, this was clearly not the case, as is evident from examining the 16 affix positioning errors presented in Table 4.13.

	LOW FREQ	HIGH FREQ
HIGH MD	12 errors: *gain weights* *settle downs* *settle downs* **wraps ups* *read alouds* **reads alouds* **reads alouds* **reads alouds* **reads alouds* *tear aparts* **tears aparts* **tears aparts*	4 errors: *make sures* *make sures* **wakes ups* **depends onned*
LOW MD	0 errors	0 errors

TABLE 4.13. Distribution of affix shift errors, and double-marked affix errors collected in the shadowing task. Double-marked errors are indicated with an asterisk. There are 16 total errors out of 1483 codable responses.

The most striking feature of the distribution of affix placement errors is that all 16 of them involved bigrams classified as having High Mutual Dependency. We may verify that the effect of MD is statistically significant using a Fisher Exact test. For purposes of this analysis, the relevant comparisons involve the 1483 classifiable responses. A contingency table for Mutual Dependency may be prepared as in Table 4.14.

	AFFIX POSITIONING ERRORS	NO AFFIX POSITIONING ERROR	TOTALS
HIGH MD	16	737	753
LOW MD	0	730	730
TOTALS	16	1467	1483

TABLE 4.14. Contingency table for affix positioning errors on bigrams, High and Low Mutual Dependency.

Based on this Mutual Dependency data, a two-tailed Fisher Exact test yields an extremely significant result ($p < 0.0001$). As expected, bigrams with high Mutual Dependency are over-represented among the set of affix placement errors.

However, inspection of Table 4.13 also reveals that most of the affix positioning errors (12 out of 16) are low in Token Frequency. A contingency table summarizing the Token Frequency data is presented in Table 4.15.

	AFFIX POSITIONING ERRORS	NO AFFIX POSITIONING ERROR	TOTALS
HIGH FREQ	4	743	747
LOW FREQ	12	724	736
TOTALS	16	1467	1483

TABLE 4.15. Contingency table for affix positioning errors on bigrams, High and Low Token Frequency.

With respect to Token Frequency, the distribution of affix positioning errors runs counter to the hypotheses of this study, and counter to the earlier findings based on naturalistic errors. Moreover, in a Fisher Exact test, the effect of low Token Frequency is statistically significant at the 0.05 level ($p = 0.046$).

The foregoing findings may be verified with additional statistical tests. Multiple logistic regression is a natural choice for analysis of data in which a dependent variable is binary (Harrell 2001, Baayen 2008, Jaeger 2008), as is the case for the occurrence or non-occurrence of a speech error. To represent such data situations, logistic regression maps the variables of a linear regression onto the logistic function (Figure 4.1), such that the dependent variable is a probability between 0 and 1. The logistic function provides a mathematical representation of a relatively abrupt leap between the two binary outcomes, with 1 corresponding to the occurrence of the event of interest, and 0 corresponding to non-occurrence (Jaeger 2008). The independent variables may be categorical, or they may be continuous and unbounded (potentially ranging from [-∞, +∞] on the x-axis). Additionally, a combination of continuous and/or categorical independent variables may be used in a multiple regression model (Hosmer and Lemeshow 2000, Harrell 2001).

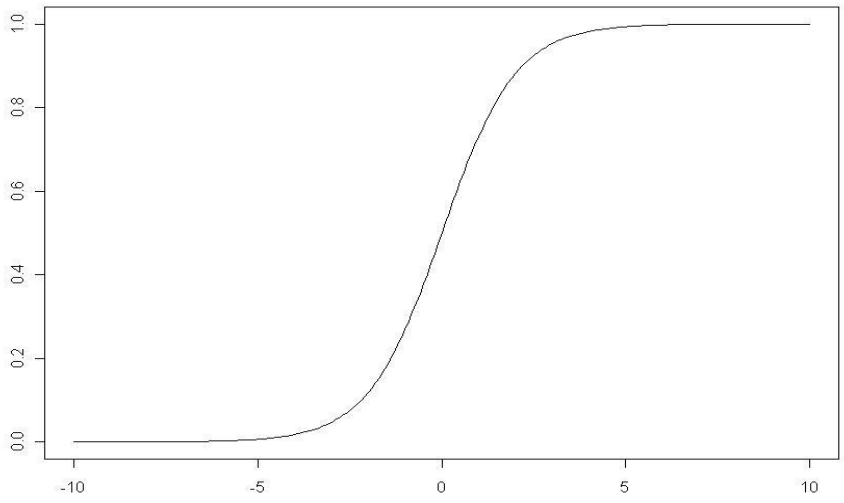

FIGURE 4.1. Logistic function, plotted in R using $f(x) = 1/(1 + e^{-x})$. The x-values may range from [-∞, +∞], and y is bounded by [0, 1].

With respect to the present dataset, it is first necessary to check that multiple logistic regression would provide an appropriate analysis. Logistic regression (or any other kind of regression) will lead to problematic analyses in the event that different independent variables are highly correlated with one another. The risk goes beyond easily-detectible associations between variables, since a broader concern is that one independent variable could largely be a mathematical function of other independent variables. The basic concern is that the independent variables may be characterized by 'collinearity,' or 'multicollinearity.' Mosteller and Tukey (1977: 280) explain that 'the idea is that we get into trouble when we try to treat one piece of information as if it were several pieces. This inevitably leads to arbitrariness about the allocation of the weights to be given the several pieces.' More generally, the concern is that, if there are more than two independent variables, one of these independent variables is in fact a linear combination of other independent variables (Baayen 2008, Belsley et al. 1980). In such cases, the regression coefficients for individual variables will be unreliable (Harrell 2001), and collinearity is exacerbated considerably if variable interactions are included (Aiken and West 2001, Jaeger 2008).

In the present experiment, the principal independent variables of interest are Token Frequency and Mutual Dependency, and it is appropriate to be cautious about collinearity given that MD includes Token Frequency in its mathematical definition (see Equation 4.1). Thus, as a precaution, we may calculate the 'condition number,' κ, for independent variables used in regression analyses (Belsley et al. 1980). The condition number may be calculated using an R function called `collin.fnc()`. Condition numbers of 15 indicate moderate collinearity, and values of 30 or more indicate the

choice of independent variables is problematic (Baayen 2008: 182). The COCA Spoken Corpus values for Mutual Dependency and Token Frequency are available in Table 4.10. Using `collin.fnc()` to check these values[50] for collinearity yields $\kappa = 10.75$, providing a preliminary indication that our independent variables are acceptable components in a logistic regression.

However, collinearity is ultimately a problem of data, not the independent variables themselves. That is to say, collinearity can only be fully assessed once experimental data are available, because a sparse dataset will be more prone to instability arising from collinearity (Belsley et al. 1980: 191, Chatterjee and Hadi 2006: 222). Moreover, introducing variable interactions in a regression analysis often notably worsens problems from collinearity, because variable crossproducts will amplify any collinearity present among main variables (Aiken and West 1991). Thus as an additional safeguard, I will evaluate particular continuous regression analyses for collinearity by considering the Variance Inflation Factor (VIF). The VIF of a regression provides an assessment of how much of the standard error in the analysis is due to collinearity of the independent variables; lower values indicate there is less influence from collinearity (Belsley et al. 1980, O'Brien 2007). I will consult VIF scores for general guidance only, since high VIF scores are neither necessary nor sufficient indicators of variable collinearity (Belsley 1991, Harrell 2001, O'Brien 2007). Moreover, there is no universal agreement about acceptable VIF thresholds (Belsley 1991: 28). Various rules of thumb are proposed; for instance, in a discussion of logistic regression, Menard (2002:76) advises that a VIF value greater than 5 is 'cause for concern', and a value greater than 10

[50] More precisely, my analysis here is based on the $\log_2$ values of Token Frequency counts, as I explain below. By definition, Mutual Dependency is already a logarithm of a ratio. Using the raw Token Frequency counts in collin.fnc actually produces a small decrease in the condition number: $\kappa = 8.18$.

'almost certainly indicates a serious collinearity problem.'[51] Often, a VIF threshold of 4 is used as a rule of thumb (O'Brien 2007), and even more conservatively, Allison (2012) advises caution when the VIF score exceeds 2.5. In general, consideration of VIF metrics should be supplemented by additional safeguards, such as checking the reasonableness of regression coefficients, since collinearity may cause coefficients to be of the wrong sign (Chatterjee and Hadi 2006).

First, I will present a logistic regression model in which the independent variables are evaluated categorically, using the 2 X 2 bins described in Section 4.2.2. In this analysis, however, conventional approaches to logistic regression lead to complications, insofar as two of the four bins contain 'zero cells' (see Table 4.13). In such a situation, the standard method in logistic regression would lead to unstable solutions in analyzing both main effects and interactions (Heinze and Schemper 2002, Faraway 2005). In categorical datasets that are small or sparse, it is common for analyses to be plagued by 'data separation,' a situation in which one of the independent variables perfectly predicts the outcome. The problem is that standard regression methods rely on ratios between bin counts; when a bin contains zero items, a term in the maximum likelihood estimate goes to positive or negative infinity (Heinze and Schemper 2002, Zorn 2005, Gelman and Hill 2007). Such is the case with the data in the present experiment, in which all of the errors belong to high MD bins. This data separation leads to unreliable results in any logistic regression test that uses the maximum-likelihood estimate on the 2 X 2 stimulus categories.

[51] Menard (2002) actually uses the mathematically equivalent concept of 'tolerance,' which is the reciprocal of VIF. Thus, in Menard's approach, a VIF greater than 5 corresponds to a tolerance less than 0.2; a VIF greater than 10 corresponds to a tolerance less than 0.1, and so on.

We may address the data separation problem by using an alternate algorithm for logistic regression analysis of categorical data, namely, the bias reduction method of Firth (1993). This method introduces a corrective term that counteracts unstable consequences from zero cells, while still yielding solutions close to the maximum-likelihood estimates in less problematic datasets (Faraway 2005). Firth's solution has been implemented in an iterative algorithm available as an R package, called `logistf` (Ploner et al. 2010, Heinze and Schemper 2002, Heinze and Ploner 2003).

Using the bias reduction method via `logistf`, logistic regression analysis of the 16 affix positioning errors verifies that MD has an extremely significant effect ($p < 0.0001$). The regression coefficient, β, for MD is 3.49, corresponding to an odds ratio of 33.09 (that is, odds of an affix error increase by a factor of 33.09 in the high MD group compared with the low MD group). The same logistic regression model indicates that the effect of Token Frequency is significant at the 0.05 level ($p = 0.039$). Again, the observed effect for Token Frequency is reversed from the predicted pattern; errors are more likely in the low-frequency group. For frequency, the regression coefficient, β, is -1.07, giving a decreased odds ratio of 0.34 for errors in the high-frequency group.

A second logistic regression model is attainable if we analyze the independent variables as <u>continuous</u> values; that is, we can base the analysis on MD and Token Frequency values for each stimulus (again see Table 4.10), without explicitly delineating membership in 'high' and 'low' 2 X 2 bins. Some explanations are in order regarding the conventions I assume in these continuous regression analyses. First, I will base continuous regressions on frequency counts from the COCA Spoken Corpus, subjected to a log (base 2) transformation. The decision to log-transform frequency counts is not

based on any a priori statistical requirements; there are no distributional restrictions (such as normality) on the independent variables used in a logistic regression (Harrell 2001: 35). Nevertheless, it is common practice to log-transform frequency counts prior to performing a logistic regression analysis (Baayen 2008). As discussed in Chapter 1, one reason to do so is that the impact of frequencies in cognition may be better described by a logarithmic relationship, rather than a linear one.

Moreover, I will follow the convention of not 'centering' the continuous independent variables prior to analysis. As a countermeasure against collinearity, it is sometimes advised that independent variables be centered at zero by subtracting a constant (such as the variable mean) from each value (Jaccard et al. 1990, Aiken and West 1991, Jaeger 2008). However, there is ongoing debate in the literature whether centering independent variables truly counteracts the effects of collinearity. It is true that centering variables would lower the VIF scores reported in this chapter, in addition to the condition number ($\kappa = 10.75$) reported above. Nevertheless, following Belsley (1991: 28), I elect not to center any of the independent variables prior to regression modeling. Belsley (1991: 189-190) demonstrates that centering variables 'throws away information' about the data, thus masking important collinearity diagnostics. More recent research (Echambadi and Hess 2007, Dalal and Zickar 2012) also argues that variable centering does not improve collinear data. Finally, although researchers often center variables in an attempt to improve model significance, the practice actually does little to alter the regression parameters, and does not improve the detectability of variable interaction effects (Kromrey and Foster-Johnson 1998, Shieh 2011).

Use of continuous MD and (log) Frequency values allows us to perform more conventional logistic regression in R using the `lrm()` method in the `rms` package[52] (Harrell 2012; see also Harrell 2001, Baayen 2008 on the earlier `Design` package in R). This analysis again verifies that high MD is associated with an increase in affix positioning errors ($p < 0.0001$), with a positive regression coefficient ($\beta = 0.86$). High (log) Token Frequency is associated with a decrease in such errors ($p < 0.0001$), with a negative coefficient ($\beta = -0.90$).[53] The `lrm` model described here has a reasonably good fit ($p_{gof} = 0.29$) in a le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test (le Cessie and van Houwelingen 1992).[54] Moreover, `lrm` reports a coefficient of concordance, C = 0.92. This coefficient surpasses the expected threshold of 0.80, indicating an acceptable model (Gries 2009). Moreover, the VIF test results indicate that collinearity of (log) Token Frequency and MD is reasonable in this model, yielding values of 3.76 for each variable.

Thus, the logistic regression analyses, as well as the Fisher Exact test, indicate that affix positioning errors are <u>less</u> likely to occur on high-frequency bigrams. This finding is quite surprising in the context of the predictions of this study, and is contrary to

[52] Generally speaking, similar results hold if the bias reduction method (logistf) is applied to the continuous values for MD and Token Frequency, rather than dichotomous values. Where possible, I focus on the more familiar logistic regression analyses from `lrm` in order to include a breadth of analyses.

[53] In the case of continuous logistic regressions, the interpretation of regression coefficients is less intuitive than in the categorical analysis, in part because the continuous scales in each case are logarithmic. But we may use the regression coefficients to estimate the change in odds ratio for each unit increase in the independent variable. In the present model, these coefficients provide odds ratio changes of 2.35 for each unit increase in MD, and 0.41 for each unit increase in log Frequency.

[54] For purposes of a goodness-of-fit test, note that we want values that are <u>above</u> 0.05 (or more conservatively, above 0.10). I use the notation 'p_{gof}' to clarify that this p-value should receive a special interpretation.The described sum of squares test is appropriate to sparse continuous-value data (in contrast with the more familiar chi-square goodness-of-fit test). The p_{gof} values reported here are calculated using a function (`resid(lrm.object, 'gof')`) available in the `rms` package in R (Harrell 2012). For continuous analyses, I report goodness-of-fit scores and the coefficient of concordance, C, but do not report R^2 values, following Hosmer and Lemeshow (2000). They point out that in logistic regressions, R^2 values are typically quite small, and potentially confusing when compared against metrics for linear regression models (2000: 166-167).

the earlier findings regarding affix errors in high-frequency bigrams. However, it is worth noting that the preponderance of low-frequency bigrams among the errors arises entirely from bigrams that are high in Mutual Dependency. All 12 of the low-frequency bigrams among the affix placement errors are also in the high-MD group; no errors are observed when Frequency is low if MD is also low. This pattern hints at an interaction between independent variables in the experimental task, rather than a general effect from low Token Frequency.

We may expand our logistic regression models to investigate the possibility of interactions between variables. However, demonstration of this interaction proves difficult based on the variables as described thus far. Again using the bias reduction method, a categorical analysis (using `logistf`) finds that the interaction term (for MD*Token Frequency) does not have a significant p-value (p= 0.61). Moreover, in this expanded regression model, Token Frequency is not significant (p = 1.0), while MD is still highly significant (p = 0.0002). The regression model is not improved by adding a term for interactions; the 'model fit' measured via likelihood ratios is worsened. Similarly, an interaction model over continuous values is not significant (using either the maximum likelihood model or the bias reduction method). For instance, the `logistf` analysis results in p = 0.02 for (log) Frequency and p = 0.03 for MD, and a non-significant p-value of 0.32 for their interaction. Moreover, the regression coefficients hint at an unstable analysis. Mutual Dependency has a positive coefficient (β = 0.50), and Token Frequency has a negative coefficient (β = -1.38), which are both consistent with the data. However, the interaction coefficient is positive (0.03), which is counter to expectations given the opposing effects of the two independent variables.

Moreover, a VIF test indicates that the interaction model is subject to collinearity problems: the diagnostics are 10.22 for MD, 49.31 for (log) Frequency, and 81.03 for the variables' interaction. Recall that with respect to collinearity, our concern is that no independent variable should be a linear combination of other variables (Belsley et al. 1980). Diagnostics indicate that partial collinearity exists between the two independent variables used in the experiment design (MD and Frequency). However, investigating an interaction in essence introduces a third independent variable (Aiken and West 1991), and the resulting collinearity renders the present interaction model unstable. A common solution to collinearity in regression modeling is to remove an independent variable (Baayen 2008), but in the present case such an approach would not allow us to investigate interactions between variables.

Thus, it seems that a variable interaction between Frequency and MD cannot be verified in the present logistic regression model. Nevertheless, it is quite clear that affix positioning errors are more prone to arise under a particular confluence of variables (Low Frequency and High MD). We may confirm this pattern by comparing the results of models in which the variables are included in isolation, as opposed to jointly as presented above. If we use `lrm()` for a continuous regression with MD as the sole variable (without including Frequency as a separate factor), the p-values are still significant ($p = 0.0125$, $\beta = +0.16$). But the model fit is unsatisfactory: $p_{gof} = 0.03$, and $C = 0.70$. Similarly, a model based solely on Token Frequency has a significant p-value ($p = 0.006$, $\beta = -0.30$), but the model fit is unsatisfactory ($p_{gof} = 0.03$, $C = 0.69$). Yet as reported above, when the model includes both MD and Token Frequency, both variables have significant p-values, and the model fit is good ($p_{gof} = 0.29$, $C = 0.92$). Such a result

indicates that the best description of the error distribution relies on both variables, even though no further regression interaction between these variables can be demonstrated, apparently due to variable collinearity.

4.2.4.3. **Combining no-marking errors with other affix errors**.

As discussed in Section 4.0, there is one additional type of affix error which might be associated with holistic processing of multiword sequences. Specifically, no-marking errors (e.g., *she read aloud*) could be taken as evidence that a speaker has failed to insert an inflection into a prefabricated sequence. As noted previously, a failure to add inflections could be consistent with unrelated phenomena; for instance, in some speech communities, the absence of a third person singular *–s* may be normative.

However, in the present experiment, I proceed with reporting on the occurrence of missing inflections, taking note of some mitigating factors. First, review of the missing inflections among the study's participants indicates that none of the speakers omit *–s* systematically. Among the bigram stimuli, there were 5 total responses involving an omitted *–s*. All 5 errors were by different speakers; that is, no speaker was responsible for more than one of these omissions on his or her 56 responses to the stimuli.[55] Secondly, as noted previously, in the experimental task, participants must initially hear and process a sequence in its uninflected form (*they read aloud*), and a no-marking error may indicate that the sequence is less readily altered or interrupted. If a speaker does indeed activate a particular multiword sequence holistically (*read aloud*, for instance), the study's

[55] Three of the five participants in question did have at least one missing inflection on the compound distractors (e.g., *he fly...fish*). No-marking errors were quite common across all participants on the compound distractors.

hypotheses predict that it may be more difficult to insert an inflection into this sequence in a time-pressured task.

Thus, in Table 4.16 I present the 5 participant responses in which an *–s* suffix was altogether omitted.

	LOW FREQUENCY	HIGH FREQUENCY
HIGH MD	3 errors *she... read aloud* *she gain weight* *she screw up*	1 error *she ca-... call it*
LOW MD	0 errors	1 error: *he conveniently forget it*

TABLE 4.16. Distribution of no-marking errors (n= 5) collected in the shadowing task.

This dataset is quite small, but the distribution of errors across the four bins seems similar to the distribution of the other 16 affix placement errors presented in Table 4.13. Indeed, the Fisher Exact analyses are essentially unchanged if we pool together the 16 affix positioning errors of Table 4.13 and the 5 affix omissions in Table 4.16. In this pooled analysis, the effect of high MD is extremely statistically significant ($p < 0.0001$), and the effect of low Token Frequency is statistically significant ($p = 0.025$).

Table 4.17 presents a quantitative synopsis of the 21 combined affix errors observed in the shadowing experiment.

	LOW FREQUENCY	HIGH FREQUENCY
HIGH MD	15 affix errors/371 codable responses (5 affix shifts) (7 double-marked affixes) (3 no-marking errors)	5 affix errors/382 codable responses (2 affix shifts) (2 double-marked affixes) (1 no-marking error)
LOW MD	0 affix errors/365 codable responses	1 affix error/365 codable responses (1 no-marking error)

TABLE 4.17. Distribution of all affix placement errors (n = 21) collected in the shadowing task, out of 1483 codable responses.

We may supplement the Fisher Exact tests of these pooled results with additional logistic regression analyses. With respect to zero cells, the modified dataset in Table 4.17 provides a small improvement over Table 4.13, insofar as there is now one item in the Low MD/High Frequency bin. However, the dataset as a whole still exhibits 'quasi-complete separation' because there are no errors observed in the Low MD/Low Frequency bin (Zorn 2005). Thus for a categorical analysis, a more conservative approach to logistic regression will again incorporate bias reduction, rather than using the maximum likelihood estimate (Firth 1993, Heinze and Schemper 2002, Heinze and Ploner 2003). Based on the 21 affix errors summarized in Table 4.17 (compared with the 1483 total codable responses), logistic regression of the categorical data (using `logistf`) yields extremely significant positive results for Mutual Dependency (coefficient $\beta = 2.62$, $p = 0.00002$), and negative results for Token Frequency which are significant at the 0.05 level (coefficient $\beta = -0.92$, $p = 0.042$). Similarly, analysis of the continuous data (using `lrm`) yields extremely significant results for both Mutual Dependency and Token Frequency ($p < 0.0001$ for both variables), with an acceptable

model fit (p_{gof}= 0.27, C=0.86). Moreover, collinearity is acceptable, since the VIF diagnostic (2.52) is within range of the most conservative benchmarks (Allison 2012).

For the 21 errors in Table 4.17, logistic regression of variable interaction is again problematic due to variable collinearity. Using the bias reduction method for categorical data, the interaction term (MD*Token Frequency) falls short of significance (p= 0.16). In this model, MD is still extremely significant (p < 0.0001), but Token Frequency is not significant (p = 0.47). The regression coefficients make intuitive sense in this interaction model: MD has a positive coefficient (β = 3.46), and Token Frequency also has a small positive coefficient (1.10), representing the fact that when MD is low (corresponding to a zero term), Token Frequency has a small positive effect on the occurrence of errors. The MD * Frequency interaction coefficient is negative (-2.19), as we expect given that the effects of High Frequency and High MD on errors are opposed to one another. Likelihood ratio tests over the interaction and main effects models indicates that the interaction model (LR = 23.88) represents an improvement over the simplified model (LR = 23.34), but this improvement falls short of significance (chi-square = 1.07, df =1, p = 0.30).

Using `lrm`, an interaction model of the continuous data also fails to reach significance. In this expanded model, MD is significant (p = 0.02), with a positive coefficient (0.68). Token Frequency is not significant (p = 0.20), and has a negative coefficient (-0.63). The interaction is also not significant (p = 0.86), with a very small negative coefficient (-0.005). However, the VIF diagnostics are notably unacceptable for the interaction model, with values of 11.06 for MD, 38.65 for Frequency, and 72.99 for

the variables' interaction, hinting that variable collinearity may interfere with the investigation of interactions.

As in the smaller dataset, even though demonstrating variable interactions remains problematic, the best description of the 21 affix errors considers both independent variables in tandem. That is, errors increase with Low Frequency and High Mutual Dependency, but either variable alone results in an unsatisfactory regression model. For instance, in an `lrm()` model with MD as the sole variable, the p-value is significant ($p = 0.01$, $\beta = +0.14$), but the fit is poor ($p_{gof} = 0.01$, and $C = 0.70$). Likewise, an `lrm()` model of Token Frequency alone has a significant p-value ($p = 0.044$, $\beta = -0.27$), but the model fit is poor ($p_{gof} = 0.04$, $C = 0.68$). Yet as demonstrated above, an `lrm()` model including both variables is significant, and has a good fit ($p_{gof} = 0.27$, $C = 0.86$). Thus, for the expanded set of 21 errors, it again seems that affix errors are more likely to arise under a particular confluence of variables, namely High MD and Low Frequency.

4.2.5. **Post hoc analyses: Examining components of the MD metric**

In sum, the foregoing analyses demonstrate that higher Mutual Dependency of a bigram is associated with increased likelihood of affix errors, implying that such multiword sequences are more likely to be retrieved as units, and/or less amenable to interruption with inflections in a time-pressured task. This result is as predicted, and consistent with the earlier findings from affix errors collected in naturalistic settings. However, these analyses also indicate that higher Token Frequency is associated with a decrease in these same types of affix errors. As noted, this pattern is observable only among bigrams high in MD, but logistic regression models cannot directly verify variable

interactions, apparently due to collinearity in the interaction term, or a relative paucity of error data, or both. Moreover, the 'backwards' frequency effect is a counterintuitive finding which is contrary to the conversational error results, and in need of further investigation. In the following post hoc analysis sections, I address these concerns in several ways.

To help make sense of the Token Frequency data, in this reanalysis I consider the possibility that confounding variables may in fact be driving the reversed effects from frequency. Indeed, retracing the experiment design reveals that additional cross-category frequency differences may be relevant. Specifically, frequencies of the component words in each bigram are a noteworthy factor in the experiment, and asymmetries are apparent if we focus on the set of stimuli which is most prone to affix errors (Low Token Frequency, High Mutual Dependency). For instance, comparing the frequency of the verb in the bigrams in the four stimulus categories reveals that there are rather striking category differences. The rounded average verb frequency (from the COCA Spoken Corpus) for items in the Low Frequency, High MD bin is 3308; this contrasts markedly with averages of 13395, 87953, and 30346 for the other three bins.

Similarly, the second word of each bigram's second word is, on average, lower in the Low Frequency, High MD bin. The average second-word frequency for this bin is 224559; this is markedly lower than the averages in the other three bins (572268, 688406, and 1119300). Listings of the bigrams' component-word frequencies are provided in Appendix 4.4.

All four bins contain a considerable range of verb frequencies, and there are overlaps in values across all the bins, but there is a clear overall trend: on average, lower-

frequency words occur in the category which is most prone to affix errors. Here it should be noted that component-word frequencies were not left to vary in an uncontrolled way in the experiment design. Rather, during stimulus selection it was noted that there would necessarily be cross-category differences in component-word frequencies. More specifically, the bigrams in the Low Token Frequency/High MD bin are, in large part, categorized as such due to the relatively low frequencies of their component words. To understand why this is the case, note from Equation 4.1 that the MD value is calculated on the basis of three corpus measures: frequency of the bigram; frequency of the bigram's first word (that is, F(V)), and frequency of the bigram's second word (F(w_2)).[56] Items in the Low Token Frequency/High MD bin are restricted to those having a relatively low bigram frequency (otherwise, obviously they would be classified as 'High Token Frequency'). Thus the only way for these items to surpass the 'High Mutual Dependency' threshold, while maintaining low frequency for the overall sequence, is for the bigram to consist of lower-frequency component words. More precisely, in each bigram in this bin, F(V) needs to be quite low, or F(w_2) needs to be quite low, or both words need to be moderately low in frequency).

In the analyses below, I present evidence which suggests that the apparent backward effect from Token Frequency is in fact an effect from these low component word frequencies; that is, the overrepresentation of low-frequency words in the Low Frequency/High MD bin accounts for the finding that affix errors are more likely on Low-Frequency bigrams. It is not surprising that component word frequencies may play an important role in holistic access, which necessarily involves diminished activation of

[56] The fourth variable in the equation is the corpus size, N, which is of course constant for all items in the stimulus set, and merely helps scale MD values to be greater than zero.

the individual components. Below I examine frequency effects from the bigrams' first word (the verb), followed by frequency effects from the bigrams' second word. Finally I present alternate analyses based on considering the frequencies of both component words together. These analyses are based on a variety of additional logistic regression analyses that directly include component word frequencies as a variable. In general, rather than creating ad hoc 'high' and 'low' categories for these measures, I focus on continuous analyses in `lrm()` based on log-transformations of the component-word frequency counts.

4.2.5.1. **Post hoc analysis 1: Frequency of the verb.**

With respect to verb frequencies, there are quite intuitive mechanisms which would explain why lower-frequency verbs might be more prone to affix errors in the experiment. In general, it is reasonable to expect that if speakers have more practice with particular verbs (i.e., high-frequency verbs), these items will be less prone to result in affix errors. As I explain in more detail below, relevant influences may arise from online demands in production, and/or comprehension, in the experiment. To investigate such effects, I consider two measures of verb frequency. First, there is the frequency of the verb's base form (abbreviated as F(V)), such as the corpus frequency of bare stems such as *settle* or *talk*. This uninflected verb frequency is the measure used directly in the calculation of Mutual Dependency (represented as F(w_1) in Equation 4.1). It is reasonable that lower F(V) could increase the rate of errors in the present experiment, first of all, because the task requires segmentation of the verb as a precursor to inserting an inflection. If lower-frequency verbs are segmented less readily as separate words (and

activated less as independent units), they may be more prone to errors in which an inflection is delayed or omitted. More generally, the relevant frequency metric might be said to be the verb lemma frequency, that is, combined corpus frequencies of *settle, settles, settled, settling*, and so on, because this combined frequency would be relevant to how readily a verb is segmented from continuous input. However, in the present analysis, I focus on F(V) as a proxy for this more general frequency metric, on the assumption that participants are expecting a bare verb stem.

Secondly, we may consider the frequency of the verb's inflected form, that is, the frequency of *settles, talks*, and so on, which I abbreviate as F(Vs). The frequency of inflected verb-forms loosely correlates (in a log-linear fashion) with the verb's base frequency, but we can identify particular psychological factors with respect to F(Vs) which could be relevant to the distribution of affix errors. Specifically, it is reasonable to anticipate that when verbs occur very frequently in their inflected form, the base + inflection may in fact be retrieved as a unit or as a well-practiced sequence, and thus high-frequency inflected units may be characterized by easier, error-free production (see Stemberger and MacWhinney 1986b).

To investigate these dynamics in the experiment's dataset, it is worthwhile to verify first that these variables have some effect on affix error probabilities. Thus, I initially consider models which incorporate only verb frequency in isolation. Indeed, continuous logistic regression analyses provide some evidence that this is the case, and these models yield negative coefficients which indicate that lower verb frequencies are associated with an increase in affix errors. The effects do not reach significance for the smaller set of 16 affix positioning errors, however. For this dataset, a logistic regression

of bare verb stem frequency, F(V), has a p-value of 0.15 (β=-0.14, p_{gof} =0.73, C=0.62), and inflected verb frequency, F(Vs), has a p-value of 0.11 (β=-0.16, p_{gof}=0.96, C=0.65). The results are somewhat better on the expanded set of 21 affix errors that includes no-marking errors. Here, a regression model of F(V) approaches significance, with p = 0.09 (β=-0.14, p_{gof} =0.70, C= 0.61). A regression model of F(Vs) reaches significance at the 0.05 level, with p= 0.03 (β=-0.19) and an acceptable model fit (p_{gof}=0.91, although C= 0.63).

These findings do support the general prediction that affix errors are more likely on bigrams that start with a low-frequency verb. Of course, an accurate regression model should include all significant variables, and the use of a sole model variable may account for the low coefficient of concordance scores above. In discussions later in this section, I integrate verb frequencies into more inclusive regression models.

4.2.5.2. **Post hoc analysis 2: Frequency of the bigram's second word.**

Let us consider now the possibility that the distribution of affix errors is influenced by F(w_2), the frequency of the bigrams' second word. This account would imply, for instance, that the sequence *settle down* is more prone to being processed as a holistic unit if *down* is a low-frequency word. In fact, post hoc regression analyses imply that F(w_2) is a significant factor, with negative regression coefficients indicating that lower-frequency (second) words are more likely to result in an affix error. This is initially evident if we propose logistic regression models using F(w_2) as the sole independent variable. For the set of 16 affix errors, the coefficient is negative (β=-0.37), and the regression is extremely significant (p<0.0001). This single-variable regression model passes goodness-of-fit diagnostics (p_{gof} =0.21, C=0.85). Similar results obtain for the

expanded set of 21 errors. Again, the regression coefficient is negative (β=-0.34), as expected, and the p-value is highly significant (p <0.0001). Moreover, the model fit is good (p_{gof} =0.63, C=0.80).

These results would indicate that in the experimental task, if a bigram's second word is a low-frequency item, it is activated to a lesser degree as an independent unit (thus resulting in errors such as *tear aparts* and *tears aparts*). This finding is interesting, because it helps to rule out a possible alternate explanation for the affix errors in this experiment. Recall that Stemberger (1984, 1985) presents an account in which certain grammatical units occur early because they are 'overactivated,' and are thus uttered earlier in speech than their targets. This is indeed a plausible explanation for 'early' affix shifts such as *If its break.* With respect to the 3rd singular *–s* inflection, for instance, this suffix is far more frequent than all but the top few words[57], and it is plausible that *–s* might arise in speech in advance of some of these less readily-available words (for instance, *break*).

Consider, then, the occurrence of affix shifts along the lines of *gain weights* and *settle downs*. How can we be certain that such errors are not merely a consequence of early activation? That is, there is a possible interpretation of full affix shifts, in which the second word in the bigram (such as *weight* or *down*) is activated prematurely, resulting in uttering this word prior to the *–s* inflection. However, the frequency analysis of F(w_2) provides evidence to make such an alternate account less plausible. If 'overactivation' of the second word, w_2, were a crucial factor, we would expect higher-frequency words to

[57] For instance, searching the COCA spoken corpus for relevant 3rd singular verbs (following the pattern *s.[v?z*]), and subtracting non-affixed forms such as *is*, indicates that this inflection occurs almost 900,000 times in the corpus. This makes the suffix far more frequent than all English words, with the exception of the following: *the, to, and, a, of, that, I, you, in, it, is,* and *we*.

be more prone to occurring early. But this is hardly the case: in fact, lower-frequency words are more likely to occur early in this error set. Thus it seems more reasonable to postulate that early production of w_2 involves characteristics of the bigram (w_1w_2), rather than characteristics of w_2 as an independent word.

Moreover, note again that the errors of interest in this study include double-marked errors such as *settles downs*, in addition to full affix shifts such as *settle-downs*. There are 9 double-marked errors, along with 7 full affix shifts. Both of these types of error are more likely among the High MD, Low Token Frequency bin, and indeed there is no distinguishable difference in the distributions of the two error types. It is difficult to see how the frequency of w_2 could have an effect on the tendency for affixes to occur redundantly on the second word, in addition to the verb (w_1 in the bigram). More likely, it seems that the distribution of both error types arises from a more general phenomenon involving the bigram characteristics.

4.2.5.3. **Post hoc analysis 3: Component frequencies together**.

Preliminary analyses thus indicate that regressions based on component word frequencies are as we expect: bigrams containing low-frequency words are more likely to result in affix errors in the experiment, implying that such sequences are more prone to being activated as whole units. For a more thorough synthesis, these component-word frequencies should be incorporated into broader regression models. Given the choice among five or more independent variables, a multitude of possibilities present themselves as candidates for regression models to pursue. A naive approach might be simply to add F(w_2) and F(V) (or alternately, F(Vs)) into the regression alongside the design variables, MD and Token Frequency. However, such an approach would be ill-advised; note that

MD is essentially an algebraic combination of the other three variables (Token Frequency, and two different word frequencies). Thus, collinearity would run rampant in such a model; indeed, inspecting this variable combination using `collin.fnc()` yields a diagnostic that is orders of magnitude above usual benchmarks,[58] and `lrm()` actually fails to converge on a result.

Thus, in this section, I focus on two reanalyses which provide an alternate perspective on the apparently paradoxical effects of Token Frequency. First, consider a regression model which includes bigram MD, in addition to component-word frequencies (F(V) and F(w_2)) as an alternative to Token Frequency of the bigram.[59] An initial inspection of MD, (log) F(V), and (log) F(w_2) using `collin.fnc()` indicates that variable collinearity is within acceptable limits, with a condition number $\kappa = 22.74$. This alternate approach generates models with significant results, in which the three regression variables have the appropriate coefficient sign. A regression over the set of 16 affix errors produces a model with a good coefficient of concordance (C = 0.93) and a (marginally) acceptable goodness-of-fit (p_{gof} =0.09). Mutual Dependency has a positive effect on the occurrence of affix shift errors (β=+0.36), as we have generally seen in various regression models, and the effect is very significant (p = 0.0093). Both (log-transformed) component word frequencies have a negative, significant effect on errors, matching our expectation that infrequent words may be processed less readily as independent units. For F(V), β = -0.30, and p = 0.0299; and for F(w_2), β = -0.45, and p <

[58] Specifically, if you're curious, the condition number generated is 16,984,801,961, which is of course somewhat larger than the value of 30 which typically indicates problematic collinearity.
[59] I focus here on F(V) to the exclusion of F(Vs) in order to constrain the wide range of combinatoric possibilities for regression models. Results are generally similar in models based on MD, F(Vs), and F(w_2).

0.0001. Collinearity is not a concern in this reanalyzed model (VIF scores are 1.84 for MD, 1.57 for F(V), and 1.77 for F(w_2)).

Similar results are found using the expanded set of 21 errors. Mutual Dependency has a positive, significant effect on affix errors (β=+0.24, p = 0.0113). Verb frequency (F(V)) has a negative, significant effect (β = -0.23, p = 0.0345), and frequency of the second word has a negative, highly significant effect (β = -0.38, p <0.0001). This regression model's diagnostics are good (p_{gof} =0.59, C = 0.87), and collinearity is not a concern (VIF scores are 1.30 for MD, 1.26 for F(V), and 1.26 for F(w_2)).[60]

The foregoing reanalyses thus present a plausible, alternate account of why in the experimental data, lower-frequency bigrams are more prone to affix errors. The errors of interest are clustered on a particular set of stimuli—those which are high in Mutual Dependency, and low in Token Frequency. Yet these bigrams are also lower in component-word frequencies, and the reanalysis demonstrates that individual word frequencies perform well in regression analyses alongside Mutual Dependency. Moreover, aside from improvements in terms of theoretical plausibility, the models incorporating F(V) and F(w_2) offer other improvements over the original models based on bigram token frequency. Using Likelihood Ratio tests in `lrm()`, we may compare the fit of the original two-variable model (MD and Token Frequency) against the reanalyzed, three-variable model using component-word frequencies (MD, F(V), and F(w_2)). For the set of 16 affix errors, this comparison reveals that the reanalyzed, three-variable model represents an improvement, with a Likelihood Ratio chi-square of 2.98, although this

[60] I will not investigate address variable interactions at length regarding the present reanalyses. As a general observation, in the present context, an analysis of interactions (requiring the inclusion of 5 or more total variables) leads to null results, and moreover, all the main variable effects lose significance when interactions are included. It is likely that variable collinearity plays a role in this problem, since VIF diagnostics often rise to 200 or more.

difference is not quite significant (p = 0.08, with 1 degree of freedom). However, the same comparison across models of the expanded set of 21 errors does reach significance. Here, the Likelihood Ratio chi-square difference is 4.34, with p= 0.037 (with 1 degree of freedom between the two models). This test indicates that the model based on MD and component-word frequencies is not only plausible; it represents a significant improvement over the model based on MD and Token Frequency.

Nevertheless, there are further compelling findings if we perform additional reanalyses which do include Token Frequency as an independent variable. Here, I present the results of analyses which include Token Frequency of the bigram, along with component-word frequencies (F(V) and F(w_2)). As always, all corpus frequency values are log-transformed as a preliminary step. The three variables in this reanalysis are acceptable with respect to collinearity; for Token Frequency, F(V), and F(w_2), a `collin.fnc()` test yields $\kappa = 23.01$.

In the present reanalyses, Mutual Dependency is not included explicitly, but we may think of individual word frequencies as a proxy for this measure; i.e., when combined with Token Frequency, the frequencies of the component words allow for a full mathematical expression of the corpus metrics that vary with Mutual Dependency (see Equation 4.1). To put this another way, the present selection of variables includes all three corpus frequency elements—in atomic form as frequency counts, rather than as a summary ratio — that are used in the definition of Mutual Dependency, Mutual Information, or other relative frequency measures.

This selection of variables yields `lrm()` regressions with significant effects for all three variables. Moreover, the results are rather interesting; once individual word

frequencies are included explicitly in regressions, the apparent backward effect from Token Frequency vanishes. For instance, based on the set of 16 affix positioning errors, the regression coefficient for Token Frequency is now positive (β = + 0.73), and the result is very significant (p = 0.0093). This result obtains, apparently, because the reversed effects associated with Token Frequency are already better accounted for by the other variables in the model. As expected, the component-word frequencies have negative regression coefficients: for F(V), β= -0.67, and p = 0.0068; while for F(w_2), β = -0.81, and p < 0.0001. The fit of this model is acceptable, albeit marginally so (p_{gof} =0.09), and the coefficient of concordance is good (C= 0.93). Variable collinearity is rather high (VIF scores are 11.24 for bigram frequency, 4.86 for F(V), and 11.02 for F(w_2)), although this in itself does not seem to justify rejecting the model (O'Brien 2007).

Similarly, based on the expanded set of 21 affix errors, this configuration of regression variables indicates that higher Token Frequency results in an <u>increase</u> in affix errors. For Token Frequency, the coefficient is positive (β = +0.49) and significant (p = 0.0113). Once again, the component-word frequencies have negative coefficients. For F(V), β= -0.47 (p = 0.0057), and for F(w_2), β= -0.62 (p<0.0001). This regression approach has solid measures with respect to model fit (p_{gof} =0.59, C= 0.87), although variable collinearity does exceed the more conservative benchmarks (VIF scores are 6.75 for Token Frequency, 3.17 for F(V), and 6.29 for F(w_2)).

In sum, then, a reanalysis of the data incorporating component-word frequencies yields expected effects for all variables – including Token Frequency. This approach actually offers a moderate statistical improvement over the original regression models based on Mutual Dependency alongside Token Frequency. Again, we may verify this

using Likelihood Ratio comparisons between the original, two-variable model (MD, Token Frequency), and the three-variable, reanalyzed model (Token Frequency, F(V), and F(w_2)). For the set of 16 errors, the three-variable model represents an improvement, although it is not statistically significant (chi-square = 2.98, 1 df, p = 0.084). For the set of 21 errors, however, there is a statistically significant improvement (chi-square = 4.34, 1 df, p = 0.037).

Thus, there is statistical support for this second approach to reanalyzing the experimental data. The distribution of errors is as expected for all three variables included in the model: frequency of the bigram, frequency of the first word, and frequency of the second word. Lower frequencies of individual words within each bigram are associated with an increase in affix errors. Once these component frequencies are expressly included in the model, we can see that higher bigram frequency is also associated with an increase in affix errors. As noted above, the present data reanalysis offers an alternate way of approaching the measure of interest in Mutual Dependency (or other relative frequency scores), by including the components of this quantity as separate elements. When seen in this light, it becomes apparent that the distribution of errors is indeed as predicted by the theory: errors evidencing holistic processing increase when the whole unit is more accessible, or when its component parts are less accessible.

4.3. **Conclusion: The evidence add ups.**

In Chapter 1, I argued that access units in the lexicon will generally be honed with practice, so as to efficiently retrieve items that tend to co-occur. Indeed, in most cases, experimental investigation of prefabs finds that elements which frequently co-occur are

more prone to fluent and error-free retrieval (e.g., Reali and Christiansen 2007, Tremblay et al. 2007, Arnon and Snider 2010). However, in investigating holistic retrieval, it is also possible to turn this notion on its head, and investigate special cases in which the retrieval of preassembled units actually interferes with a task that requires morphosyntactic analysis. Such is the certainly the case in monitoring studies (Vogel Sosa and MacFarlane 2002, Kapatsinski and Radicke 2009), in which participants are slower to recognize a target word within a well-practiced unit.

Similarly, the investigations of this chapter were proposed on the assumption that holistic units are more likely to be associated with certain inflectional errors, both in naturalistic settings and in an experimental task. These first forays into a systematic study of affix positioning errors have indeed shown that bigrams are more prone to being retrieved as a unit when the two words frequently co-occur. The results are promising, but discussion is needed to reconcile the quantitative findings from the naturalistic and experimental studies.

First, it is encouraging that Mutual Dependency proves to be predictive with respect to affix errors from conversation, as well as those elicited experimentally. In the two analyses of conversational errors which included MD as a variable, this measure is found to be statistically significant. Based on analyses of verb-initial bigrams, high-MD sequences are overrepresented among naturalistic outbound shifts (such as *come backs*). A second analysis indicates that bigrams that occur in conversation with outbound shifts (*come ups*) have significantly higher MD scores than bigrams that contain early shifts (*quites get*). Likewise, in the experimental task, high-MD bigrams are overwhelmingly more likely to prompt shift errors that indicate the sequence is activated as a unit,

including outbound shifts (*gain weights*) and double-marking errors (*wrap<u>s</u> up<u>s</u>*). These positive findings are concordant with earlier studies by Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), which provide empirical support for the related relative frequency measure, Mutual Information. Like Mutual Information, Mutual Dependency provides a mathematical representation of competition between the activation of whole units (the measure's numerator) and the activation of component units (the denominator). The current findings suggest that Mutual Dependency is a useful summary statistic worthy of further investigation.

The findings with respect to Token Frequency turn out to be rather more complicated. Among the conversational errors, high Token Frequency sequences are more likely to prompt the affix errors of interest. High-frequency bigrams are overrepresented among outbound affix shifts, as indicated by analyses of verb- and noun-initial bigrams in the Brown corpus, and verb-initial bigrams in the COCA corpus. Moreover, outbound affix shifts occur on bigrams that are higher in Token Frequency than the bigrams containing early affix shifts, indicating that the former bigrams are more cohesive. These results from naturalistic errors agree with earlier findings that higher-frequency sequences are more likely to be accessed holistically (e.g., Kapatsinski and Radicke 2009).

However, these findings are not immediately borne out among the experimentally-induced errors. Initial quantitative analyses indicate that, contrary to expectations, bigrams that are <u>low</u> in Token Frequency are more likely to prompt affix positioning errors, due to a tendency for errors to arise among items that are low in Token Frequency, but high in MD. Followup analyses (Section 4.2.5) suggest that the

anomalous frequency pattern is in part an artifact of the experiment design. In the experimental stimuli, Token Frequency of the bigram is confounded with frequencies of the component words, and these component word frequencies turn out to be essential to the distribution of affix positioning errors. Indeed, one of the post hoc analyses (Section 4.2.5.3) indicates that higher-frequency bigrams are more likely to prompt affix positioning errors—but this finding is only observable if we take into account the effects of component word frequencies. In the experimental task, bigrams containing infrequent words (an effect involving both the first and second words) are more likely to result in affix errors indicative of holistic retrieval. The importance of component-word frequencies in the experimental task lends support to relative frequency accounts of processing, in which units containing low-frequency components are more likely to be processed holistically (Frauenfelder and Schreuder 1992; Hay 2001, 2003).

These reanalyses raise further questions regarding the integration of findings from naturalistic and experimental data. Among the naturalistic data, to what extent might component word frequency be a contributing factor in the distribution of affix shifts? We have evidence that naturalistic affix errors are more likely on bigrams that are high in Mutual Dependency. Yet a bigram's MD can be high in some cases because the bigram's token frequency is very high, or in other cases because the component words are infrequent. Among experimental affix errors, the indications are that low component frequencies are quite important. For conversational errors, a full analysis of component-word frequencies is beyond the scope of this chapter, but an initial examination hints that these component word frequencies are not as crucial to the occurrence of affix errors in naturalistic contexts. Among the naturalistic errors (Tables 4.1 and 4.2), note the

occurrence of a number of verbs of extreme high frequency: *give us-ing; want to comes; go for-ing.* As a case in point, we may focus on the 12 conversational errors involving a misplaced 3rd singular *–s* inflection. Among this set, some of the same ultra-frequent verbs recur repeatedly: *come* is represented by four errors (*come ins, come ons, come ups, come backs*), and *go* is represented by three errors (*go aheads, go gets, goes homes*). Moreover, if we use corpus analysis[61] to split observed tokens of VERB+*s* into 'high frequency' (the top half of all tokens) and 'low frequency' (the bottom half), 10 out of the 12 verbs are high-frequency. Thus in natural speech, there does not seem to be a tendency for outbound shifts to occur among lower-frequency verbs.

Why might some differences be observed between the patterns of affix errors in natural settings, compared with those induced in the experiment? Consider the demands facing speakers in the time-pressured experimental task, compared with the demands of normal conversational speech. The shadowing methodology explored here requires participants to perceive speech, segment it into words, and almost immediately echo it back, while monitoring continuously for the appropriate site to insert a verbal inflection. Since speech production occurs in such short succession after comprehension (typically with a lag of just 1-2 seconds), it is reasonable that segmentation errors might result in syntagmatic production errors.[62] Of course, speech comprehension (including word segmentation) is still relevant to the study of prefabricated units; indeed, such processes

[61] In this particular case, I used the spoken portion of COCA (Davies 2008), rather than the Brown Corpus, because for single words it is possible to retrieve an exhaustive, part-of-speech constrained list of frequencies from COCA.

[62] By referring to 'errors' of segmentation, I do not mean to imply that there must be a definitive word boundary, which the speaker happens to overlook. As discussed in Chapter 1, boundaries between words are expected to be gradient, and this principle applies during comprehension as well as production.

are at the core of other experimental investigations of holistic retrieval (e.g., Vogel Sosa and MacFarlane 2002, Kapatsinski and Radicke 2009).

However, it is worth acknowledging that the affix errors induced under the present shadowing methodology give a glimpse into the joint effects of comprehension and production. Future work in this area may benefit from a revised methodology that focuses more exclusively on speech production. This may be accomplished by allowing participants to hear the target sentence in its entirety before repeating it back from memory. Along these lines, pilot work indicates that outbound affix shifts may also be induced experimentally, if participants are asked to insert an inflection into a memorized sentence while performing an unrelated distractor task (specifically, phoneme monitoring). However, it remains to be seen whether such an approach will be as effective as the shadowing methodology in prompting affix positioning errors.

One general goal of this chapter has been to investigate a new experimental methodology, with potential for the quantitative study of multiword units. The findings thus far are encouraging, insofar as outbound affix shifts and double-marked inflections are induced on approximately one out of a hundred attempts — orders of magnitude more frequent than what we observe in casual speech. Nevertheless, the collection of experimental data in the current task is labor-intensive, insofar as each response requires a participant to repeat aloud an entire sentence, and many attempts must be made for every successful error elicitation. There thus remains a certain needle-in-the-haystack quality to the elicitation experiment. A much larger set of errors would be useful in addressing a wider range of quantitative questions, such as whether affix shifts are

distributed evenly across the frequency spectrum (or MD spectrum), or whether a U-shaped curve exists (cf. Kapatsinski and Radicke 2009).

Toward this end, to locate a larger number of needles, it may be helpful to supplement experimental work by looking in much larger, text-searchable haystacks. One approach is to gather affix shift errors from large online corpora such as COCA (Davies 2008), which yields a handful of relevant errors from a few searches. Unfortunately, it is not possible to retrieve an exhaustive sample of all affix shifts from a corpus. Note, for instance, that part-of-speech taggers tend to assign word classes unpredictably when affix shifts occur, and existing tags thus cannot be used to identify candidate errors. However, collections of errors can be assembled in a piecemeal fashion, such as by obtaining all sequences with the form _____ *ups* or _____ *upped*, and then filtering to identify actual errors. In expanding the dataset, it will turn out to be useful to include affix errors from written sources. Typos may arise for many unsystematic reasons (e.g., a random finger-slip onto the <s> key), but the patterns of interest could be observable in a large enough sample. Inclusion of written data needs to be selective, however; the texts should be as close to casual conversation as possible, so that typing errors are less subject to offline editing. Various corpora of online discourse (e.g., the 30 billion-word Westbury Lab USENET corpus; Shaol and Westbury 2010) may provide a rich source of data.

On the other hand, the experimental task described in this chapter may have applications that go beyond the original research questions of this dissertation. It turns out that errors such as *gain weights* and *tears aparts* were not the most common ones observed in participants' responses. Inbound affix shifts on distractor items actually outnumbered the outbound errors that motivated this study. I will conclude this chapter

by briefly acknowledging this serendipitous finding, which should prove useful in future studies of word structure.

Collectively, on the distractor sentences, the 29 participants produced 20 'early shift' responses in which the *–s* inflection appears on the first word of the compound verb. An example appears in (20).

(20) *As the office participates in a teambuilding exercise, she plays-act at various situations that might arise.*

Such utterances are true 'inbound' shifts, as introduced in Section 4.0, since the inflection is inserted inside a lexical item. A tally of the errors for each item observed appears in Appendix 4.5. Additionally, there are 37 double-marked errors among the compound verb responses. One example is given in (21).

(21) *There's little chance of tough questions at the press conference, and she spoonsfeeds... official policies to reporters.*

Responses of this sort are, in one sense, quite different from double-marked errors observed among the bigram stimuli (*she wraps ups*), since double-marked compounds require that an inflection interrupt a normative lexical unit. On the other hand, the double-marking errors also speak to a certain commonality between the bigram stimuli and the compound verb: when such errors occur, they indicate the speaker is activating the complex unit as a whole (*spoonfeed-s*, *wrap up-s*) and as an assemblage of parts (*spoon-s + feed*, *wrap-s + up*).

Among the compound verbs, the shadowing methodology proves to be surprisingly effective at prompting early shifts, as in (19), and double-markings, as in (20). For ease of reference, I will refer to both of these error types as 'inbound shifts,'

since in both cases an inflection intrudes into a complex lexical unit. Almost every participant in the experiment produced at least one inbound shift; the average number of shifts is 1.93 per participant. Moreover, almost half of the compound verbs prompted an error, with 19 out of the 56 compounds producing at least one inbound shift.[63] Thus, the affix-insertion task may afford a rich source of information about the online processing of compound words, which would be of interest in current research on retrieval and decomposition of compound words (Badecker 2001, Libben 2005, Baayen et al. 2010).

A full analysis of the existing compound data is beyond the scope of this dissertation. However, a few preliminary observations are possible, offering a kind of mirror-image correspondence with the error patterns observed among bigram stimuli. First, inbound errors are more likely to occur on compounds that are low in frequency: *playact, spoonfeed, flyfish, globetrot*. Such a distribution is unsurprising; if a complex form such as *globetrot* has a weaker representation in memory, it will be more challenging to correctly inflect the whole unit. Secondly, if the first compound-internal word is very frequent as a verb, then the unit is more likely to be parsed into two words, and the compound-internal component is more prone to attract an inflection: *play**act**, blow**dry**, sleep**walk**, fly**fish*. Further, we can expect that the components of each compound will be activated to varying degrees, and there will be competition between holistic retrieval ([*playact*]$_{\text{VERB}}$) and compositional retrieval ([[*play*]$_{\text{VERB?}}$ + [*act*]$_{\text{VERB?}}$]). Due to competition between parts and wholes, then, we would predict inbound shifts to be most prevalent if the compound form is infrequent, or the first word within the

[63] Note further that 14 of the 56 compound verbs are poor candidates for prompting or detecting inbound *–s* inflections, since these items contain a sibilant at the word-internal morpheme boundary: *force#feeds, wise#crack, side#steps, baby#sits*, etc. I have thus excluded these items from the table in Appendix 4.4, and they are excluded from the quantitative analysis below.

compound is a frequent verb, or both.[64] These predictions can be verified by investigating an independent variable that represents competition between wholes and parts, calculated by dividing the (log) frequency of the compound's first word (e.g., the COCA frequency of [*play*]$_{\text{VERB}}$) by the (log) frequency of the compound (e.g., the combined COCA frequencies of *playact, play-act,* and *play act*). A logistic regression using this summary statistic in fact yields significant results for the distribution of inbound affix shifts. The regression coefficient (β= 0.39) is positive, indicating that inbound shifts increase as the first word's frequency increases in relation to the compound's frequency. The p-value for the regression is significant (p = 0.0005), and the fit is acceptable (p_{gof} = 0.72).

Further study of experimentally-induced inbound affix shifts is warranted, and may proceed in tandem with additional investigations of affix errors on multiword sequences. It is hoped that inbound-type shifts (on compounds) and outbound-type shifts (on multiword sequences) can be encompassed under a broader theory of gradient analyzability for complex units. The evidence suggests that frequent complex units (whether bigrams, or compounds) are generally more prone to being accessed as wholes. When components of the complex unit are themselves frequent, competition from these components makes compound-internal affixation more likely (or, in the case of verb-initial bigrams, makes affixation of the verb occur readily). Conversely, when components of the complex unit are infrequent, diminished activation of these components makes holistic retrieval more likely.

---

[64] There are undoubtedly additional important factors, including, for instance, the phonotactic boundary within the compound word. Inbound shifts also appear to be more likely in cases where an improbable phonotactic transition (*leap#frog, black#mail, spot#light, book#mark*) encourages analysis of the compound as a sequence of separate words (cf. Hay 2001).

CHAPTER 5. CONCLUSION

Generative approaches to syntax have persistently held to the view that whenever possible, speech will be produced or comprehended following abstract rules (Pinker and Ullman 2002), such that frequencies (or probabilities) are largely irrelevant to the grammar (Chomsky 1957, 1969). In this view, the mental lexicon contains nothing more than a list of 'exceptions,' that is, information that cannot be predicted by rule (Chomsky 1995: 6, 235). However, it is increasingly recognized that speakers have much more fine-grained linguistic knowledge than a parsimonious storage model would predict, and redundancy is rampant in the lexicon (e.g., see Jackendoff 2010: 590). Moreover, empirical evidence in numerous domains demonstrates that frequency is an essential factor in language processing and language change (Ellis 2002, Bybee 2007), including the representation of multiword sequences in memory.

The view emerging from current research makes no demand that we do away with abstractions. However, it is clear that alongside more abstract generalizations must exist a complex system that is influenced by the frequencies of various units. In Chapter 1, I presented an argument that if emerging units are not represented at intermediate stages (in some form or another), it is not clear how they can ever get stored. That is, if frequency of a unit ever makes a difference in cognition, frequency must always make a difference. Every experience has some effect, albeit small, on mental representations (Bybee 2006). This is hardly to say that frequency of a complex unit is the only factor relevant in storage and processing, since component frequencies may also be important. Along these lines, this dissertation set out to investigate behavioral correlates of two different approaches to measuring frequency of co-occurrence—involving absolute and

relative measures—and found evidence for both of them, supplementing other psycholinguistic evidence for prefabricated multiword units.

The dictation experiment of Chapter 3 is focused on token frequency (that is, absolute frequency) of multiword sequences. In the verbatim memory task, significant differences are evident in performance between high- and low-frequency sequences. Most strikingly, among high-frequency multiword sequences, it is on the whole less likely subjects will retrieve only incomplete or disconnected parts of the sequence. These results indicate that absolute frequency indeed plays a role in the development of prefabs. Based on the current body of research, it would seem premature to reject token frequency as a measure of interest for multiword sequences. Although there are claims of null findings for token frequency in the experimental literature (Schmitt et al. 2004, Ellis et al. 2008), these claims must be considered alongside similar studies showing that token frequency is a significant factor in processing multiword units, including Kapatsinski and Radicke (2009), Arnon and Snider (2010), and Chapter 3 in this dissertation.

There is thus experimental support for the basic insight from usage-based theory that frequency of exposure is associated with the development of units (Langacker 1987). All the same, research into complex units may benefit from a broader perspective that allows for the possibility that frequencies of parts and wholes interact and compete. Various relative frequency measures may be used to investigate such dynamics quantitatively, but many prior psycholinguistic studies (Chapter 3 among them) do not include these metrics. Exceptions may be found in the studies by Ellis et al. (2008) and Ellis and Simpson-Vlach (2009), which control for Mutual Information alongside Token Frequency. The Mutual Information results are quite promising, but it should again be

noted that there are potential pitfalls in the use of MI due to a scoring bias in favor of low-frequency events. This problem can be partially mitigated by including only high-MI items which are also relatively high in absolute frequency (Evert and Krenn 2001), and indeed, such an approach was implemented (without comment) in Ellis et al. (2008) and Ellis and Simpson-Vlach (2009). It seems that the relative frequency of multiword sequences is not entirely separable from absolute frequency, since these measures are entangled in multiple ways: relative frequency is defined on the basis of token frequencies, and token frequency must further be taken into account to avoid spurious results.

One alternate approach is to allow an additional boost from token frequency, as in the Mutual Dependency measure explored in this dissertation. Mutual Dependency would seem to be a promising summary statistic, since it provides a relatively intuitive representation of the competition between wholes and parts, while also being mathematically sound. In the studies of Chapter 4, Mutual Dependency offers a useful account of the distribution of outbound affix shifts and related errors. Across the errors collected from naturalistic contexts as well as the experimental setting, high MD is the measure which consistently predicts the retrieval of two-word sequences as units.

Nevertheless, the quantitative analysis of the affix error data is unquestionably problematic when Token Frequency and Mutual Dependency are included in the same model—perhaps due to the definitional overlap between Token Frequency and MD, and associated collinearity. Among the experimentally-induced errors, initial analyses indicate that the frequency pattern is contrary to expectations, with affix errors most likely among bigrams that are low in Token Frequency (as well as being high in MD).

Post hoc analysis based on individual component frequencies may provide the clearest picture of the different factors involved in the experimental affix shifts. This followup analysis indicates that a two-word sequence is more likely to be processed as a unit when the component words are infrequent, and the sequence itself is frequent. That is, Token Frequency is positively associated with holistic retrieval when component-word frequencies are taken into account. Indeed, this regression model has the best fit when all component frequencies are included (frequency of the whole sequence, and frequency of the two component words), and the model fit is superior to that of the original model including Mutual Dependency alongside Token Frequency. Such componential frequency analyses thus seem to be promising for future research.

In sum, the studies presented in this dissertation provide evidence that the frequencies of complex units, in addition to the frequencies of their component parts, are registered in cognition. It is reasonable to maintain that absolute and relative frequencies both have effects on the processing and retrieval of multiword sequences. Some of these effects may be overlapping, given that relative frequencies can be said to arise from the competition between absolute frequencies of different units. Other effects may be separate, based on evidence from the grammaticalization of complex units that contain highly-frequent parts. As argued in Chapter 1, it may be possible for multiword sequences that are of extreme high frequency to be retrieved holistically, irrespective of high frequencies among component words (Bybee 2010).

Moreover, the simultaneous tracking of absolute and relative frequencies may actually be part of a bigger, and more complex, picture. Note for instance that syntactic constructions can be primed (Bock 1986), and that language processing is sensitive to the

frequencies of particular syntactic constructions (Jurafsky 1996). Thus, the occurrence and frequencies of more abstract grammatical patterns must be registered in cognition as well, and syntactic competence may actually arise from higher-order statistical operations over abstract types (Seidenberg et al. 2002). Perhaps most importantly, the registering of higher-order co-occurrence patterns may be essential to the development of new semantic-pragmatic associations, and the development of new multiword units. The recurrence of multiple items in sequence is, in its own right, the impetus for gradual change (Bybee 2002), but certainly the registering of higher-order patterns (e.g., observing when units X and Y occur in sequence, in context Z) may allow associations to form with a larger communicative context (Bybee 2010).

The multifaceted account of frequencies sketched here may seem to be needlessly baroque, if one starts from the viewpoint that frequency should be excluded from grammatical knowledge. Yet these various dynamics are matters for empirical inquiry, and explanation is needed when quantitative patterns of usage (inferred from corpus analysis) correspond to observable differences in behavior, whether in casual speech or in experimental settings. The studies in this dissertation support the basic insight that language structure and language usage are intertwined, and this interrelationship includes the gradual development of new multiword units.

APPENDICES

Appendix 3.1. Spoken BNC frequencies of the target sequences in Schmitt et al. (2004). The experiment data from Schmitt et al., previously presented in Table 3.1, is re-presented here for ease of comparison.

target cluster	Spoken BNC Frequency	Produced correctly	Partially incorrect	Not produced	Schmitt Mean performance	Reanalysis: Mean performance
to make a long story short	0	23	3	4	1.633	66.67
I don't know what to do	45	27	2	1	1.867	83.33
to give you an example	7	8	10	12	0.867	-6.67
as a matter of fact	56	21	4	5	1.533	56.67
from the point of view	54	19	5	6	1.433	46.67
in the same way as	39	3	11	16	0.567	-26.67
is one of the most	18	27	2	1	1.867	83.33
in the middle of the	172	17	2	11	1.200	50.00
aim of this study	0	2	16	12	0.667	-46.67
it's not too bad	50	16	11	3	1.433	16.67
I see what you	71	3	25	2	1.033	-73.33
you've got to have	151	16	10	4	1.400	20.00
as shown in figure	0	3	17	10	0.767	-46.67
what I want to	111	21	6	3	1.600	50.00
it was going to	72	21	6	3	1.600	50.00
as a consequence of	15	13	6	11	1.067	23.33
in a variety of	14	15	11	4	1.367	13.33
in the number of	22	18	9	3	1.500	30.00
in addition to the	32	18	10	2	1.533	26.67
night and day	11	16	1	13	1.100	50.00
on and off	66	25	0	5	1.667	83.33
something like that	923	16	5	9	1.233	36.67
go away	350	28	0	2	1.867	93.33
for example	1106	18	0	12	1.200	60.00
you know	30814	24	0	6	1.600	80.00

Appendix 3.2. Stimulus sentences for the dictation experiment of Section 3.2.

Practice sentences:
P1. *Later that night we drove along the famous Sunset Boulevard, which had an amazing view although the traffic was extremely slow.*
23 + 44 = ___

P2. *Yesterday the Chief Justice simply announced that the case was pending, to the disappointment of those who had hoped for a quick ruling.*
56 + 37 = ___

Trials:

1. *In general the company never liked criticism from employees, but eventually the policy changed* ***as a result of*** *the worker's complaint.*
14+ 38 = ___

2. *Once Sam realized he had misplaced his glasses, he turned the car around and went* ***back to*** *the bank.*
82+19=___

3. *One of downtown's most memorable landmarks is an elaborate church, which dates* ***to the same time as*** *the famous courthouse.*
33 + 17 =___

4. *Even though dad complained about the constant clutter, the kids left their shoes lying* ***in the middle of*** *the living room.*
21 + 56 = ___

5. *These days, architects see the appeal of traditional building materials, and are* ***looking to*** *the past for more creative alternatives.*
14 + 77 = ___

6. *The panel accused the organizers of using bribes to influence the decision, including one scholarship* ***for the child of*** *an official.*
38 + 26=___

7. *The two neighbors were talking over the backyard fence, but they were interrupted when*
all of a sudden *their dogs started barking.*
57 + 41=___

8. *I do not yet know if I will attend the meeting, but I will give you an answer* ***as soon as*** *I can.*
34+67=___

9. *I tried to talk with my neighbor about politics and current events, since he showed interest*
in things of *that nature.*
67+ 29=___

10. *As the violent blizzard raged on, the farmer still had to walk* ***out to*** *the shed to get firewood.*
44 + 49=___

11. *The group has pushed for a more active role, but they have encountered resistance* ***on the part of*** *most family doctors.*
15+37 =___

12. *Since I was still fond of my old car, I refused to look for a new one,* ***in spite of*** *my mechanic's advice.*
77+18= ___

13. *When the phone rang, I set my book* ***on top of*** *the refrigerator, then spent most of an hour trying to find it again.*
25+59 =___

14. *The musicians asked the audience to shout out suggestions, and they chose the best one* ***as the name of*** *their band.*
13+47 = ___

15. *After the tenants overflowed the tub, the water flooded over the bathroom floor and dripped* ***through to*** *the basement.*
57+16= ___

16. *The new bill increases penalties for white-collar criminals, arguing they should be sentenced* ***in the same way as*** *other criminals.*
33+38=___

17. *Although it shows increasing signs of a change, Mobile is a city rooted firmly* ***in the style of*** *the Deep South.*
64+ 27=___

18. *Last year, the actor was praised for playing the famous scientist, but* ***according to*** *relatives it was not a realistic portrayal.*
82+17=___

19. *The study finds that more people are deciding to postpone having children, usually making this choice* ***for the sake of*** *a career.*
24+ 58=___

20. *The current gallery exhibit seems oddly familiar to me, because the drawings are*

all of a boy *who lived in my neighborhood.*
49 + 29=___

21. *The ongoing research project will be the largest in the department's history, if we describe it* ***in terms of*** *total expenses.*
16+56=___

22. *The little girl struggled to carry the branch to the campfire, because it was really almost*
as big as *she was.*
24+ 74=___

23.*When his savings and work schedule permitted it, the baker would spend time* ***out of*** *the country visiting family.*
37 + 24=___

24. *The president argued that there was no real peace in the region if small countries lived*
in fear of *more powerful neighbors.*
47 + 36=___

25. *In the garden, we found insect damage* ***on half of*** *the plants, so we decided that we might have better luck next year.*
68 + 23=___

26. *The actor took time off from his usual profession, and wrote a screenplay based* ***on the life of*** *his own father.*
18+ 45=___

APPENDIX 4.1. Listing of the 56 stimulus sentences containing the verb bigram stimuli. Sentences are presented in matched groups of four. The following bin labels apply throughout:
(a.) Low Token Frequency, Low Mutual Dependency
(b.) High Token Frequency, Low Mutual Dependency
(c.) Low Token Frequency, High Mutual Dependency
(d.) High Token Frequency, High Mutual Dependency

1. Intransitive Phrasal; Verb + Adv. Domain: miscellaneous domestic.

a. *When the apartment fills with the odor of burnt pasta, they move up to the roof to breathe some clean air.*
b. *Although gardening used to seem like an awful chore, these days they get down in the dirt, happy to pull weeds.*
c. *Since there's not much else to do before it's time to go, they settle down on the couch with the remote control.*
d. *Because the noise from the train tracks is almost nonstop, they wake up in the morning feeling groggy and dazed.*

2. **Intransitive Phrasal; Verb + Adv (Idiomatic). Domain: jobs/employment.**

a. *It's a constant source of frustration around the office, but they give in whenever the boss asks for overtime.*
b. *The restaurant is short of cooks during the night shift, so they fit in very well with the rest of the team.*
c. *By not remembering the time zone difference, they screw up the time for the teleconference.*
d. *It's hard to find time to exercise during the week, so they work out during lunch breaks whenever possible.*

3. Intransitive Phrasal; Verb + Adv (Idiomatic). Domain: miscellaneous social, leisure.

a. *Although it's tempting to go on vacation immediately, they hold off until gas prices start to decline again.*
b. *Once the yanking on the leash becomes truly painful, they let go and the dog chases after the squirrel.*
c. *The long hike turns out to be easier than expected, but they freak out about having no cell phone service.*
d. *Instead of going to the library as originally planned, they hang out at the beach to enjoy the weather.*

4. Transitive Phrasal; Verb + Adv. (Idiomatic/metaphorical). Domain: news, politics, legal.

a. *In the editorial about the upcoming political race, they leave out the fact that the candidate was later found not guilty.*
b. *Although it may prove to be costly in the election, they take on the issue of raising the minimum wage statewide.*
c. *Even though the topic of energy has not been addressed, they wrap up the closed*

session with state lawmakers.
d. *Now that the testimony from both sides is complete, they <u>add up</u> the evidence for and against the defendant.*

5. Transitive Phrasal; Verb + Adv. Domain: miscellaneous domestic.

a. *When the neighborhood party is left without a coordinator, they <u>seek out</u> potential volunteers to be in charge.*
b. *When the date for the annual yard sale approaches, they <u>cut out</u> yellow letters to create distinctive signs.*
c. *After the thunderstorm knocks out the electricity, they <u>tear apart</u> the junk drawer searching for a candle.*
d. *After checking in with the city planning department, they <u>figure out</u> the reason that street traffic is increasing.*

6. Verb + Modifier (Adj or Adv). Domain: miscellaneous social, informal.

a. *After spending the afternoon turning the backyard compost, they <u>smell bad</u> enough to draw stares from everyone on the bus.*
b. *The agency spends a week editing photos to send to the magazine, but they <u>look good</u> in just about all the pictures to start with.*
c. *To reward the students for completing more tedious assignments, they <u>read aloud</u> from children's books an hour each day.*
d. *While planning the surprise for the kids' birthday party, they <u>make sure</u> no one is watching before sneaking off to the store.*

7. Verb + mass noun direct object. Domain: holidays, parties, social.

a. *To avoid having to wait in long lines at the store, they <u>buy food</u> for the party during the previous weekend.*
b. *Attending the concert no longer seems reasonable once they <u>see people</u> camped out waiting for tickets.*
c. *Because the treats back home are always so delicious, they <u>gain weight</u> over the holidays with little regret.*
d. *Although New Year's Eve used to be a huge deal every year, now they <u>stay home</u> all evening with the two sleeping kids.*

8. **Verb + Prepositional Phrase. Domain: political, legal**.

a. *To avoid the throngs of reporters on the steps, they <u>walk at</u> a brisk pace toward the packed courtroom.*
b. *As the senator listens and jots a few notes, they <u>point to</u> a variety of problems with the old law.*
c. *To stall for time while the bill is being revised, they <u>interfere with</u> the process of bringing it to the floor.*
d. *Despite the ads about switching to green energy, they <u>depend on</u> contributions from the coal industry.*

9. Verb + Prepositional Phrase. Domain: miscellaneous social, informal.

a. *As the conversation veers onto more personal topics, they <u>speak in</u> rapid bursts with*

infectious enthusiasm.
b. *It might not be necessary to post a listing on the internet, because they know of someone who likes to collect antiques.*
c. *Suddenly the game erupts into a pillow fight, and they fall off the bed, laughing and unable to breathe.*
d. *The school group's spirits are noticeably raised once they pay for everyone to play a round of miniature golf.*

10. Verb + Prepositional Phrase. Domain: arts, media, pop culture.

a. *In this age of Twitter, Facebook, and nonstop viral videos, they fear for the future of serious art and literature.*
b. *Suddenly the late night studio is crowded with groupies and agents, because they come with an entourage to every public appearance.*
c. *Even though the interview is before 9:00 on a Sunday, they arrive at the hotel acting surprisingly perky.*
d. *Although the new movie project is attracting a new audience, they worry about abandoning faithful viewers of the series.*

11. Verb + Prepositional Phrase. Domain: academic interactions.

a. *As students dodge one another in the busy hallway, they run after the professor to learn when grades would be posted.*
b. *Although the specific details will have to be debated, they agree on the timetable for syllabus revisions.*
c. *Based on the overall performance on the midterm exam, they insist on having a review session before the final.*
d. *On the first day of class in the seminar for majors, they talk about real-world experiences that are relevant.*

12. Verb + Pronoun. Domain: social, food.

a. *When no one seems interested in the day-old doughnut, they offer it to the dog sitting under the table.*
b. *Though bringing dessert is part of the dinner party ritual, they conveniently forget it when the expert is invited.*
c. *After the success of last year's Thanksgiving turkey, they finally trust me enough to put me in charge.*
d. *When inviting friends over for the following Sunday, they call it a brunch even though it will be mid-afternoon.*

13. Verb + X (+ Verb). Domain: travel, vacation.

a. *The Pacific Crest Trail has always seemed very alluring, and they resolve to hike it by the end of the decade.*
b. *After many hours of driving the kids are restless, and they hope to find a hotel that has an outdoor pool.*
c. *It's been a long trip on airplanes and rental cars, so they recover from traveling by resting all day.*
d. *Because the blinding rain floods onto the highway, they need to drive at a crawl for*

more than an hour.

14. [Bonus round.] Verb + Prepositional Phrase. Domain: miscellaneous informal.

a. *When it suddenly becomes urgent to leave, they <u>pay at</u> the register to speed things along.*
b. *Since the hometown job prospects are not promising, they <u>move to</u> another city to try to find work.*
c. *The old neighborhood seems like a distant memory as they <u>walk through</u> an overgrown parking lot.*
d. *During a relaxing stroll in the orchard, they <u>look at</u> all the buds opening on the trees.*

Appendix 4.2. Listing of the 56 distractor sentences containing compound verbs.

Although it's unlikely any cars will be affordable, they test-drive a couple new models to try them out.
In the car theft roleplaying video game, they hotwire the sportscar while the owner is distracted.
In the series of young adult science fiction novels, they timetravel using what appears to be an old couch.
After learning about the genealogical website, they bookmark the page for future reference.
There's little chance of tough questions at the press conference, and they spoonfeed the official policy to reporters.
The department claims to be free of any bias, but they cherry –pick the data to reach a conclusion.
During the interview, the scandal is unavoidable, but they sidestep all those questions rather abruptly.
By phoning sources and verifying information, they safeguard against journalistic carelessness and fraud.
At first it seems like the ignition is completely dead, but they jumpstart the car with old cables from the garage.
After rejecting the idea of sending an email, they hand-write a thank-you note since it seems more personal.
Before heading to the barbecue at the park, they doublecheck the rules about whether pets are allowed.
Although the visit is intended to be leisurely, they jam-pack each day with errands and projects.
The afternoon radio show features diverse music, and today they spotlight the new album from a ska band.
After years of performing in relative obscurity, they skyrocket to fame on the basis of one hit single.
When the grant money fails to materialize as planned, they bankroll the exhibit to save it from being canceled.
With an angry tirade about not returning past favors, they strongarm the club owner into booking a gig.
After considering many ways to cook the chicken, they deepfry it in spite of all the health concerns.
The pie recipe has evolved over many years, and they still fine-tune it every once in a while.
Based on years of planning organic meals at home, now they mastermind intricate dinners at the cafe.
Despite not being invited to the winetasting, they freeload samples from the event in the lobby.
To decide on possible directions for the paper, they brainstorm ideas with the professor for an hour.
Although a fifteen-page limit seems long enough, they underestimate the depth of the paper topic.
In the middle of a long philosophy lecture, they daydream about adventures during

Spring Break.
While chatting during the break in the seminar, they wisecrack about having to purchase the teacher's textbook.
In private conversations when everyone is honest, they badmouth the corner neighbor for being meddlesome.
Despite the security hired by the business owners, they panhandle aggressively out in the parking lot.
After earning a reputation for doing a good job, they housesit when anyone in the complex travels.
To prepare for all the impending rental turnovers, they mass-produce welcome packets for new tenants.
Sometimes if there's a late-night gathering, they babysit the kids to earn some extra cash.
The party is festive and full of warmth, and they bearhug all the arriving friends and relatives.
To throw together some clothes for the festival, they tie-dye a few shirts in the washing machine.
In the middle of the crowded and noisy bar, they overhear an acquaintance telling a partial truth.
While drinking several pots of strong coffee, they leapfrog from topic to topic in a long discussion.
To learn how to maneuver the old kayak, they zigzag all over the pond for an entire morning.
After four weeks of the intense exercise program, they bench press almost twice as much as before.
Although the open ocean might be more exciting, they windsurf in the bay where waves are moderate.
Though usually it doesn't seem worth the effort, they dryclean the heirloom jacket to be extra careful.
After soaking and scrubbing the stain on the carpet, they blowdry it carefully so no one will notice.
In anticipation of a dangerous, dry season, they fireproof the house by replacing the roof tiles.
So that the guest room seems more inviting to children, they wallpaper the room with a cloud and bunny design.
Whenever the copier mechanism jams, they forcefeed the blank paper in manually.
As the office participates in a teambuilding exercise, they playact at various situations that might arise.
In addition to having more steady employment, they moonlight in a country band on most weekends.
According to a rumor in the company, they always blackmail the boss to receive promotions.
Even though the practice has been much criticized, they earmark funds for a special project in two districts.
Based on advice from the book agent's lawyer, they copyright the manuscript before mailing it out.

Because the candidate is a long-time acquaintance, they fundraise for the campaign as volunteers.
According to the case presented by the prosecutor, they routinely shoplift for the thrill it brings.
Because the frequent flyer miles will expire soon, they globe-trot for weeks to seize the opportunity.
The cabin rental by the river is miles from town, and they flyfish all afternoon to provide dinner.
Now that the boat restrictions have been lightened, they waterski every year at the artificial lake.
After missing the exit ramp on the dark highway, they backtrack for many miles before finding the road.
Distractors for bonus round at end of experiment:
According to a series of local legends, they sleepwalk to the river when the moon is full.
In fulfillment of a longstanding ambition, they hang-glide over the valley on a spring day.
Because there is such an enormous harvest this year, they freezedry most of the berries from the garden.
In response to the embarrassing typo on the cover, now they proofread every manuscript twice before printing.

Appendix 4.3. Practice sentences used before the experiment.

Both for the sake of exercise and reducing expenses, they bike the 6-mile commute every weekday.
Though it's usually better to schedule an appointment, they also accommodate more spontaneous visits.
To draw attention to less familiar menu items, they choose something to feature as a nightly special.
As part of a crusade against spending waste, they <u>spearhead</u> the investigation into failed programs.
On the busy corner throughout the summer season, they <u>breakdance</u> in front of excited spectators.
Out in plain view but escaping everyone's notice, they <u>pickpocket</u> sneakily through the crowd of tourists.

Appendix 4.4. Re-presentation of the 56 bigram stimuli, including frequency counts of the bigram components (from Spoken COCA). Corpus searches pertaining to F(V) are constrained by part of speech.

	LOW FREQUENCY					HIGH FREQUENCY				
	BIGRAM	**BIGRAM FREQ**	**BIGRAM MD**	**F(V)**	**F(w_2)**	**BIGRAM**	**BIGRAM FREQ**	**BIGRAM MD**	**F(V)**	**F(w_2)**
HIGH MD	1.settle down	204	14.63	1734	90508	1.wake up	1865	19.28	2225	233482
	2.screw up	119	14.20	308	233482	2 work out	1839	15.01	36216	270881
	3.freak out	60	13.60	102	270881	3.hang out	695	15.61	3391	270881
	4.wrap up	276	14.97	971	233482	4.add up	665	14.59	7344	233482
	5.tear apart	21	13.76	679	4478	5.figure out	4621	20.18	6326	270881
	6.read aloud	24	14.36	18227	143	6.make sure	10647	21.01	103464	49506
	7.gain weight	93	16.39	1857	5165	7. stay home	546	14.59	22242	51792
	8.interfere with	383	15.51	550	546640	8.depend on	1103	16.85	1459	673655
	9.fall off	211	13.93	4413	61799	9.pay for	4689	17.08	20493	737067
	10.arrive at	363	14.26	1584	405469	10.worry about	3974	19.07	6638	411223
	11.insist on	378	13.77	1446	673655	11 talk about	28166	21.64	56472	411223
	12.trust me	504	14.61	4374	220828	12.call it	4677	15.72	27512	1401667
	13.recover from	311	14.30	1390	329733	13.need to	28042	19.03	54108	2590553
	B.walk through	393	14.61	8675	68125	B.look at	32791	21.65	76959	405469
	MEAN	**238.6**	**14.49**	**3307.9**	**224559.2**	**MEAN**	**8880.0**	**17.9**	**30346.4**	**572268.7**
LOW MD	1. move up	214	10.10	17015	233482	1.get down	986	12.24	211305	90508
	2. give in	296	6.67	52448	1560158	2. fit in	568	12.70	2963	1560158
	3. hold off	204	12.28	12947	61799	3.let go	749	11.45	112634	169428
	4. leave out	132	8.42	17881	270881	4.take on	2372	12.96	99948	673655
	5. seek out	175	12.10	2455		5.cut out	567	12.87	15082	270881
	6. smell bad	11	8.61	1051	28028	6.look good	833	12.59	76959	139555
	7. buy food	69	11.10	13644	15179	7.see people	708	10.08	139252	316865
	8. walk at	22	3.71	8675	405469	8.point to	824	12.20	5301	2590553
	9. speak in	254	8.30	12507	1560158	9.know of	1631	8.26	401401	2055523
	10. fear for	136	9.84	2606	737067	10.come with	813	10.15	101723	546640
	11. run after	35	5.93	19821	96897	11.agree on	1330	13.57	20638	673655
	12. offer it	84	6.79	4341	1401667	12.forget it	618	11.79	7350	1401667
	13.resolve to	33	4.60	1648	2590553	13.hope to	1773	12.51	19776	2590553
	B.pay at	53	5.01	20493	405469	B.move to	1242	11.70	17015	2590553
	MEAN	**122.7**	**8.1**	**13395.1**	**688406.3**	**MEAN**	**1072.4**	**11.8**	**87953.4**	**1119300**

Appendix 4.5. Table of inbound errors (early shifts and double-marked) on compound distractors. Frequency counts are drawn from COCA (450 million words). This table excludes 15 items which were unsuitable for quantitative analysis: 14 contained sibilants at the morpheme boundary, and yielded no detectable errors. One additional item (*finetune*) was excluded, even though it generated 9 errors. Post-experiment discussions with participants indicated the item was often perceived as *find tune*, thus making quantitative analysis of *fine* questionable.

COMPOUND VERB	FREQUENCY OF COMPOUND	FREQ. OF COMPOUND'S 1ST WORD (as a verb)	# EARLY SHIFTS (*plays-acts*)	#DOUBLE MARKINGS (*plays-acts*)	TOTAL INBOUND SHIFTS
playacts	18	85877	6	5	11
blowdry	245	8242	2	3	5
blackmail	880	154	2	3	5
leapfrog	306	2229	2	3	5
bookmark	237	696	1	3	4
spoonfeed	25	978	0	4	4
flyfish	63	16944	3	0	3
sleepwalk	66	20310	1	2	3
spotlight	4780	4051	0	3	3
globetrot	5	0	1	1	2
handwrite	19	1840	0	2	2
bankroll	234	660	0	2	2
timetravel	523	3382	1	0	1
backtrack	315	6370	0	1	1
jampack	8	1219	0	1	1
deepfry	85	0	0	1	1
underestimate	1658	0	0	1	1
bearhug	250	13729	0	1	1
tie-dye	106	6381	0	1	1
hotwire	95	0	0	0	0
dryclean	62	4807	0	0	0
fireproof	171	6758	0	0	0
wallpaper	1812	6758	0	0	0
moonlight	2831	167	0	0	0
earmark	353	1	0	0	0
copyright	8090	2351	0	0	0
fundraise	60	4491	0	0	0
shoplift	88	3602	0	0	0
hangglide	13	16041	0	0	0
proofread	458	82	0	0	0
cherrypick	77	0	0	0	0
safeguard	1536	0	0	0	0
doublecheck	373	3885	0	0	0
skyrocket	360	22	0	0	0
strongarm	340	0	0	0	0
mastermind	773	2429	0	0	0
freeload	24	6460	0	0	0
daydream	451	10	0	0	0
badmouth	97	0	0	0	0
panhandle	892	1232	0	0	0
overhear	422	5	0	0	0

REFERENCES

Adams, Valerie. (1976). *An introduction to modern English word-formation*. New York: Longman.

Aiken, Leona S., and Stephen G. West. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Alegre, Maria and Peter Gordon. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40: 41–61.

Allison, Paul D. (2012). *Logistic regression using SAS: Theory and application*, Second Edition. Cary, NC: SAS Institute.

Anderson, John R. (1978). Arguments concerning representations for mental imagery. *Psychological Review* 85(4): 249-277.

Anderson, John R. (1982). Acquisition of cognitive skill. *Psychological Review* 89: 369–406.

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

Baars, Bernard J. (1980). The Competing Plans Hypothesis: An heuristic viewpoint on the causes of errors in speech. In H.W. Dechert and M. Raupach (eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Baars, Bernard J. (1992). A dozen competing-plans techniques for inducing predictable slips in speech and action. In Bernard J. Baars (ed.), *Experimental slips and human error: Exploring the architecture of volition*, 195-215. New York: Plenum Press.

Baayen, R. Harald. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald and Peter Hendrix. (2011). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. Paper presented at LSA workshop, Empirically examining parsimony and redundancy in usage-based models, January 2011.

Baayen, R. Harald, Victor Kuperman, and Raymond Bertram. (2010). Frequency effects in compound processing. In Sergio Scalise and Irene Vogel (eds.), *Cross-disciplinary issues in compounding*, 257-270. Philadelphia: John Benjamins.

Badecker, William. (2001). Lexical composition and the production of compounds: Evidence from errors in naming. *Language and Cognitive Processes* 16(4): 337-366.

Barlow, Michael. (2004). Collocate, version 1.0. Houston: Athelstan Publications.

Beckner, Clay, and Joan Bybee. (2008). A usage-based account of constituency and reanalysis. Paper presented at the Language as a Complex Adaptive System conference, Ann Arbor, MI. Podcast available online at http://www.wiley.com/bw/podcast/lang.asp.

Beckner, Clay, and Joan Bybee. (2009). A usage-based account of constituency and reanalysis. (2009). *Language Learning 59* Supplement 1: 27-46.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. (2009). Language is a complex adaptive system. *Language Learning 59* Supplement 1: 1-26.

Belsley, David A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley and Sons.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley Series in Probability and Mathematical Statistics.

Biber, Douglas. (2010). Corpus-based and corpus-driven analyses of language variation and use. In Bernd Heine and Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 159-191. Oxford: Oxford University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bock, J. Kathryn. (1986). Syntactic persistence in language production. *Cognitive Psychology* 18: 355-387.

Bod, Rens. (2000). The storage and computation of three-word sentences. Paper presented at Architectures and Mechanism of Language Processing Conference (AMLAP-2000), Leiden, The Netherlands. Slides available online at: http://staff.science.uva.nl/~rens/amlap00.ps

Bod, Rens. (2006). Exemplar-based syntax: How to get productivity from exemplars. *The Linguistic Review* 23: 291-320.

Bolinger, Dwight. (1976). Meaning and memory. *Forum Linguisticum* 1(1): 1–14.

Bouma, Gerlof. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede (eds.), *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, 31–40, Tübingen, Gunter Narr Verlag.

Boyland, Joyce Tang. (1997). Morphosyntactic change in progress: A psycholinguistic approach. Ph.D. dissertation, University of California at Berkeley.

Bybee, Joan. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia: John Benjamins.

Bybee, Joan. (1998). The emergent lexicon. *CLS 34: The Panels*: 421-435. Reprinted in Bybee (2007), 279-293.

Bybee, Joan. (2001). *Phonology and language use*. Cambridge, UK: Cambridge University Press.

Bybee, Joan. (2002). Sequentiality as the basis of constituent structure. In T. Givon and Bertram F. Malle, (eds.), *The emergence of language out of pre-language*, 107-132. Amsterdam: John Benjamins. Reprinted in Bybee (2007), 313-335.

Bybee, Joan. (2003). Mechanisms of change in grammaticization: The role of frequency. In Richard Janda and Brian Joseph (eds.), *Handbook of historical linguistics*, 602-623. Oxford: Blackwell. Reprinted in Bybee 2007: 336-357.

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

Bybee, Joan. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, Joan. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan, and Clay Beckner. (2010). Usage-based theory. In Bernd Heine and Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 827-855. Oxford: Oxford University Press.

Bybee, Joan, and Mary A. Brewer. (1980). Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua* 52: 201-242. Reprinted in Bybee 2007: 41-73.

Bybee, Joan, and Paul Hopper (eds.) (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Bybee, Joan, and James L. McClelland. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. In Nancy A. Ritter (ed.), The Role of Linguistics in Cognitive Science, Special Issue of *The Linguistic Review* 22(2-4): 381-410.

Bybee, Joan, and Joanne Scheibman. (1999). The effect of usage on degree of constituency: The reduction of *don't* in English. *Linguistics* 37: 575-596. Reprinted in Bybee (2007), 294-312.

Bybee, Joan, and Rena Torres Cacoullos. (2009). The role of prefabs in grammaticalization: How the particular and the general interact in language change. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali, and Kathleen Wheatley (eds.), *Formulaic language Vol. I*, 181-217. Amsterdam: John Benjamins.

Cannon, Garland. (1987). *Historical change and English word-formation: Recent vocabulary*. New York: Peter Lang.

Chatterjee, Samprit, and Ali S. Hadi. (2006). Regression analysis by example, Fourth edition. New York: John Wiley and Sons.

Chomsky, Noam. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. (1969). Quine's empirical assumptions. In Donald Davidson and Jaakko Hintikka (eds.), *Words and objections: Essays on the work of W.V. Quine*, 53-68. Dordrecht: D. Reidel.

Chomsky, Noam. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Church, Kenneth W., and Patrick Hanks. (1989). Word association norms, mutual information, and lexicography. *ACL* 27: 76-83.

Croft, William. (2001). *Radical Construction Grammar*: *Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Cutler, Anne. (1981). The reliability of speech error data. *Linguistics, 29*, 561-592.

Dalal, Dev K., and Michael J. Zickar. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods* 15(3): 339-362.

Davies, Mark. (2004-) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

Davies, Mark. (2008-). *The Corpus of Contemporary American English (COCA): 450 million words*, 1990–present. Available at http://www.americancorpus.org.

Davies, Mark. (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.

Deese, James. (1984). *Thought into speech: The psychology of language*. Englewood Cliffs, NJ: Prentice-Hall.

De Vaan, Laura, Robert Schreuder, and R. Harald Baayen. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon* 2: 1-23.

Du Bois, John W., Wallace Chafe, Charles Meyer, Sandra A. Thompson, Nii Martey, and Robert Englebretson. (2000-2005) Santa Barbara Corpus of Spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Echambadi, Raj, and James D. Hess (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science* 26: 438–445.

Ellis, Nick. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24: 143-188.

Ellis, Nick C., Matthew Brook O'Donnell, Ute Römer, Stefan T. Gries, and Stefanie Wulff. (2009). Measuring the formulaicity of language. Paper presented at the American Association of Applied Linguistics Annual Conference 2009, Denver, CO, 21-24 March. Slides available at http://ctr.elicorpora.info/files/0000/0126/AAAL2009_Ellis_et_al_NEWx.pdf.

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3): 375-396.

Ellis, Nick C., and Rita Simpson-Vlach. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5(1): 61-78.

Elman, Jeffrey L. (1990). Finding structure in time. *Cognitive Science* 14: 179-211.

Erman, Britt, and Beatrice Warren. (2000). The idiom principle and the open choice principle. *Text* 20(1): 29-62.

Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 188-195.

Fabb, Nigel. (2012). *Sentence structure*. London: Routledge.

Fano, Robert M. (1961). *Transmission of information: A statistical theory of communications.* New York: MIT Press.

Faraway, Julian J. (2005). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: CRC Press.

Firth, David. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80: 27-38.

Fontenelle, Thierry, Walter Bruls, Luc Thomas, Tom Vanallemeersch, and Jacques Jansen. (1994). Survey of collocation extraction tools. Technical report, University of Liege, DECIDE MLAP-project 93-19.

Francis, W. Nelson (1965). A standard corpus of edited present-day American English for computer use. *Literary Data Processing Conference Proceedings*, September 9. 10, 11, 1964.

Frauenfelder, Uli H., and Robert Schreuder. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 161-183. Amsterdam: Kluwer.

Fromkin, Victoria A. (1973). The non-anomalous nature of anomalous utterances. *Language* 47 (1): 27-52. Reprinted in Victoria Fromkin (ed.), *Speech errors as linguistic evidence,* 215-242. Paris: Mouton & Co.

Fromkin, Victoria A. (2000). Fromkin Speech Error Database, Max Planck Institute for Psycholinguistics. Available online at http://www.mpi.nl/cgi-bin/sedb/sperco_form4.pl

Garnham, Alan, Richard C. Shillcock, Gordon D.A Brown, Andrew I.D. Mill, and Anne Cutler. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics* 19 (7-8): 805-817.

Garrett, M.F. (1980). Levels of processing in sentence production. In Brian Butterworth (Ed.), *Language production*, vol. 1 (pp. 177-220). New York: Academic Press.

Gelman, Andrew, and Jennifer Hill. (2007). *Data analysis using regression and multilevel/ hierarchical models.* Cambridge: Cambridge University Press.

Gluck, Mark A., and Catherine E. Myers. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning*. Cambridge, MA: MIT Press.

Godfrey, J.J., E. C. Holliman, and J. McDaniel. (1992) Switchboard: Telephone speech corpus for research and development. *Proceedings of the ICASSP*, Vol. 1: 517-520.

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

Goldinger, Stephen. (1996.) Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(1):166-183.

Graybiel, Ann M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory* 70: 119-136.

Green, Lisa J. (2002). *African American English: A linguistic introduction*. Cambridge, MA: Cambridge University Press.

Gregory, Michelle, William D. Raymond, Alan Bell, Eric Fosler-Lussier, and Daniel Jurafsky. (1999). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35: 151-166.

Gries, Stefan Th. (2009). *Statistics for linguistics with R*. Berlin: De Gruyter Mouton.

Gurevich, Olga, Matthew Johnson, and Adele Goldberg. (2010). Incidental verbatim memory for language. *Language and Cognition* 2-1: 45-78.

Haiman, John. (1994). Ritualization and the development of language. In William Pagliuca (ed.), *Perspectives on grammaticalization*, 3-28. Philadelphia: John Benjamins.

Harrell, Jr., Frank E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer Verlag.

Haspelmath, Martin. (1993). The diachronic externalization of inflection. *Linguistics* 31: 279-309.

Harrell Jr, Frank E. (2012). rms: Regression Modeling Strategies. R package version 3.4-0. http://CRAN.R-project.org/package=rms

Hay, Jennifer. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6): 1041-1070.

Hay, Jennifer. (2002). From speech perception to morphology: Affix ordering revisited. *Language* 78(3): 527-555.

Hay, Jennifer. (2003). *Causes and consequences of word structure*. London: Routledge.

Hay, Jennifer, and R. Harald Baayen. (2002) Parsing and productivity. In G.E. Booij and J. van Marle (eds), *Yearbook of Morphology 2001*, 203-235. Kluwer Academic Publishers, Dordrecht.

Hay, Jennifer, and R. Harald Baayen. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7): 342-348.

Heinze, Georg, and Ploner, Meinhard. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine* 71: 181-187.

Heinze, Georg, and Michael Schemper. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21(16): 2409-2419.

Hermer-Vasquez, Linda, Elizabeth S. Spelke, and Alla S. Katsnelson (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology* 39: 3-36.

Hoffmann, Sebastian. (2005). *Grammaticalization and English complex prepositions: A corpus-based study* (Routledge Advances in Corpus Linguistics 7). London: Routledge.

Hopper, Paul J. (1987). Emergent grammar. *Berkeley Linguistics Society* 13: 139–157.

Hopper, Paul J. (1991). On some principles of grammaticization. In Elizabeth C. Traugott and Bernd Heine (eds.), *Approaches to grammaticalization* (Vol. 1), 17-35. Amsterdam: John Benjamins.

Hosmer, David W., and Stanley Lemeshow. (2000). *Applied logistic regression*, Second Edition. New York: Wiley and Sons.

Huddleston, Rodney D., and Geoffrey K. Pullum. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Jaccard, James, Robert Turrisi, Choi K. Wan. (1990). *Interaction effects in multiple regression*. Newbery Park, CA: Sage Publications.

Jackendoff, Ray. (2010). The Parallel Architecture and its place in cognitive science. In Bernd Heine and Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 583-605. Oxford: Oxford University Press.

Jaeger, T. Florian. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and toward logit mixed models. *Journal of Memory and Language* 59: 434-446.

Jurafsky, Daniel. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2): 137-194.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D. Raymond. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229-253. Amsterdam: John Benjamins.

Kapatsinski, Vsevolod, and Joshua Radicke. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali, and Kathleen Wheatley (eds.), *Formulaic language Vol. II: Acquisition, loss, psychological reality, functional explanations*, 499-520. Amsterdam: John Benjamins.

Klein, Krystal, and Chen Yu. (2009). Joint or conditional probability: Why decide? Paper presented at COGSCI 2009: The Annual Meeting of the Cognitive Science Society, VU University, Amsterdam, July 29-August 1, 2009.

Kromrey, J. D., and L. Foster-Johnson. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement* 58: 42-68.

Krug, Manfred. (2003). Frequency as a determinant in grammatical variation and change. In Gunter Rodenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 7-67. The Hague: Mouton de Gruyter.

Labov, William. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715-762.

Langacker, Ronald W. (1987). *Foundations of cognitive grammar, Volume 1*. Stanford: Stanford University Press.

Langacker, Ronald W. (1991). *Concept, image, and symbol: The cognitive basis of grammar.* The Hague: Mouton de Gruyter.

Laver, John D.M. (1973). The detection and correction of slips of the tongue. In Victoria Fromkin (ed.), *Speech errors as linguistic evidence*, 132-143. Paris: Mouton & Co.

Le Cessie, S, and J.C. Van Houwelingen. (1992). Ridge estimators in logistic regression. *Applied Statistics* 41:191–201.

Levelt, Willem J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Libben, Gary. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics* 50: 267-283.

MacKay, Donald G. (1979). Lexical insertion, inflection, and derivation: Creative processes in word production. *Journal of Psycholinguistic Research* 8(5): 477-498.

Mann, Henry, and Donald Whitney. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1): 50–60.

Manning, Christopher, and Hinrich Schütze. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marchand, Hans. (1966). *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. [Alabama Linguistic and Philological Series #13.] University of Alabama Press.

Marcus, Gary F., S. Vijayan, S. Bandi Rao, and P.M. Vishton. (1999). Rule learning by seven-month-old infants. *Science* 283: 77-80.

Marslen-Wilson, W.D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature* 244: 522-523.

Menard, Scott. (2002). *Applied logistic regression analysis: Sage University Series on Quantitative Applications in the Social Sciences*, Second Edition. Thousand Oaks, CA: Sage.

Miller, George A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2): 81–97.

Mosteller, Frederick, and John W. Tukey. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Nieder, Andreas, and Katharina Merten. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience* 27(22): 5986-5993.

Nosofsky, Robert M. (1988). Similarity, frequency and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 14: 54-65.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. (1994). Idioms. *Language* 70: 491-538.

Oakes, Michael P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

O'Brien, Robert M. (2007). A caution regarding rules of thumb for Variance Inflation Factors. *Quality and Quantity* 41: 673-690.

Pawley, Andrew, and Frances Hodgetts Syder. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt (eds.), *Language and communication*, 191-226. New York: Longman.

Pazzani, Michael J., and Stephen D. Bay. (1999). The independent sign bias: Gaining insight from multiple linear regression. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 525–530.

Pelucchi, Bruna, Jessica F. Hay, and Jenny R. Saffran. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition* 113: 244-247.

Perruchet, Pierre, and Stéphane Desaulty. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition* 36: 1299-1305.

Peters, Ann M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

Pierrehumbert, Janet. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee and Hopper (2001): 137-158. Amsterdam: John Benjamins.

Pierrehumbert, Janet. (2002). Word-specific phonetics. In Carlos Gussenhoven and Natasha Warner (eds.), *Laboratory Phonology* 7, 101-139. Berlin: Mouton de Gruyter.

Pinker, Steven. (1999). *Words and rules*. New York: Perennial.

Pinker, Steven, and Michael T. Ullman. (2002). The past and future of the past tense. *Trends in Cognitive Sciences* 6(11): 456-463.

Ploner, Meinhard, Daniela Dunkler, Harry Southworth, and Georg Heinze. (2010). logistf: Firth's bias reduced logistic regression. R package version 1.10. www.meduniwien.ac.at/msi/biometrie/programme/fl/index.html

Quirk, Randolph, and Joan Mulholland. (1964). Complex prepositions and related sequences. *English Studies* 45: 64–73.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985). *A concise grammar of contemporary English*. New York: Harcourt Brace Jovanovich.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reali, Florencia, and Morten Christiansen. (2007). Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology* 60(2): 161-170.

Sachs, Jacqueline S**.** (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* 2(9): 437–443.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. (1996). Statistical learning by 8-month-old infants. *Science* 274: 1926-1928.

Saffran, Jenny R., and Diana P. Wilson. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy 4*(2): 273-284.

Schmid, Hans-Jorg. (2010). Does frequency in text really instantiate entrenchment in the cognitive system? In Dylan Glynn and Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 101-133. Berlin: Mouton de Gruyter.

Schmitt, Norbert (ed.). (2004). *Formulaic sequences*. Amsterdam and Philadelphia: John Benjamins.

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In Schmitt (2004), 127-151.

Seidenberg, Mark S., Maryellen C. MacDonald, and Jenny R. Saffran. (2002). Does grammar start where statistics stop? *Science* 298: 553-554.

Sellen, Abigail J., and Donald A. Norman. (1992). The psychology of slips. In Bernard J. Baars (ed.), *Experimental slips and human error: Exploring the architecture of volition*, 317-339. New York: Plenum Press.

Seppänen, Aimo, Rhonwen Bowen, and Joe Trotta. (1994). On the so-called complex prepositions. *Studia Anglia Posnaniensia* 29: 3-29.

Shaoul, Cyrus, and Chris Westbury. (2010) A USENET corpus (2005-2010) Edmonton, AB: University of Alberta.

Shannon, Claude. (1951). Prediction and entropy of printed English. Bell System Technical Journal 30: 50-64. Reprinted in N.J. A. Sloane and Aaron D. Wyner, eds. (1993). *Claude Elwood Shannon: Collected Papers*, 194-208. New York: IEEE Press.

Shieh, Gwowen. (2011). Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology* 64: 462–477.

Siegler, Robert S., and Julie L. Booth. (2004). Development of numerical estimation in young children. *Child Development* 75: 428-444.

Simpson, R.C., S.L. Briggs, J. Ovens, and J.M. Swales. (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan. Available online at http://micase.elicorpora.info/.

Sinclair, John. (1989). *Collins COBUILD dictionary of phrasal verbs*. London: Collins.

Sinclair, John. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stemberger, Joseph P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology* 1(4): 281-313.

Stemberger, Joseph P. (1985). *The lexicon in a model of language production*. New York: Garland Outstanding Dissertations Series. [Ph.D. dissertation, University of California, San Diego, 1982.]

Stemberger, Joseph P. (1989). Speech errors in early child language production. *Journal of Memory and Language* 28(2): 164-188.

Stemberger, Joseph P. (1992). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In Bernard J. Baars (ed.), *Experimental slips and human error: Exploring the architecture of volition*, 195-215. New York: Plenum Press.

Stemberger, Joseph P., and Brian MacWhinney. (1986a). Form-oriented errors in inflectional processing. *Cognitive Psychology* 18: 329-354.

Stemberger, Joseph P., and Brian MacWhinney. (1986b). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition* 14(1): 17-26.

Tabachnik, Barbara, and Linda S. Fidell. (2007). *Using multivariate statistics, Fifth Edition.*. Boston, MA: Allyn & Bacon.

Thanopoulos, Aristomenis, Nikos Fakotakis, and George Kokkinakis. (2002). Comparative evaluation of collocation extraction metrics. In *Proceedings of the 3rd Language Resources Evaluation Conference*, 620-625.

Timm, Jason. (2009). Base form mechanisms to complex form opacity: A look at how morphologically complex forms lose compositionality. Unpublished ms., University of New Mexico.

Tremblay, Antoine, Bruce Derwing, and Gary Libben. (2007). Are lexical bundles stored and processed as single units? Paper presented at UWM Linguistics Symposium on Formulaic Language, Milwaukee, WI, April 18-21, 2007.

Vogel Sosa, Anna, and James MacFarlane. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language* 83: 227-236.

Wedel, Andrew. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23(3): 247–274.

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

Wray, Alison. (2006). Formulaic language. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, vol. 4, 590-597. Oxford, UK: Elsevier.

Wray, Alison. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Wray, Alison, and Michael R. Perkins. (2000). The functions of formulaic language: An integrated model. *Language and Communication* 20: 1-28.

Zorn, Christopher. (2005). A solution to separation in binary response models. *Political Analysis* 13(2): 157-170.