UNIVERSITY OF CALIFORNIA

Santa Barbara

The Distributional Learning of Multi-Word Expressions: A Computational Approach

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Linguistics

by

Alexander Robert Wahl

Committee in charge:

Professor Stefan Th. Gries, Co-Chair

Professor Fermín Moscoso del Prado, Co-Chair

Professor Patricia M. Clancy

Professor Luca Onnis, Nanyang Technological University

June 2015

The dissertation of Alexander Robert Wahl is approved.

_____

Patricia M. Clancy

_____

Luca Onnis

_____

Stefan Th. Gries, Committee Co-Chair

_____

Fermín Moscoso del Prado, Committee Co-Chair

May 2015

The Distributional Learning of Multi-Word Expressions: A Computational Approach

by

Alexander Robert Wahl

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help of many people from my professional and personal lives. These people include my doctoral committee co-chairs, Stefan Th. Gries and Fermín Moscoso del Prado. I thank Stefan for his generosity with his time, his statistical acumen, and his guidance and partnership in the development of many of the key insights in this work. I thank Fermín for his invaluable assistance in the design of my experiments, and for his keen attention to how I might position my own work at the vanguard of research in my field. I also thank the two other members of my doctoral committee, Patricia M. Clancy and Luca Onnis. Pat's meticulous eye for editing and revision, and her extensive help with my discussion of background theory in chapter 2, greatly improved the quality of the final product. Luca provided invaluable expertise in cognitive modeling, psycholinguistics, and distributional learning theory, and for this I thank him.

I also thank the broader community of graduate students and faculty in the Linguistics Department at the University of California, Santa Barbara for their friendship and intellectual engagement with me over the years. Through this close-knit and committed scholarly community, I gained a deep understanding of many different subfields of linguistics, and an enduring appreciation for many different methodologies. Among the members of this community, I wish to thank one fellow linguistics graduate student in particular—Brendan Barnwell. Brendan provided generous and patient mentorship while I was learning the Python programming language, and he enthusiastically discussed with me and helped me in developing key ideas pertaining to this dissertation.

Finally, I thank my parents, Bob and Vicki, for their unwavering support and confidence in my ability to succeed, even in the most challenging moments during my graduate school career.

VITA OF ALEXANDER ROBERT WAHL
June 2015

EDUCATION

Doctor of Philosophy in Linguistics, University of California, Santa Barbara, June 2015 (expected)

Master of Arts in Linguistics, University of California, Santa Barbara, June 2011

Bachelor of Arts in Linguistics and Portuguese, Tulane University, New Orleans, Louisiana, May 2005 (summa cum laude)

PROFESSIONAL EMPLOYMENT

2009 – 2015: Teaching Assistant, Department of Linguistics, University of California, Santa Barbara

2015: Teaching Assistant, Department of Linguistics, University of California, Los Angeles

2013 – 2014: Online Supplemental Textbook Materials Developer, *How Languages Work*, Carol Genetti (Ed.), New York: Cambridge University Press

2011 – 2013: Instructor, Department of Linguistics, University of California, Santa Barbara

2012 – 2013: e-Learning Lesson Developer, OpenEnglish.com, Coconut Grove, Florida,

2006 – 2007: English Language Instructor, São Paulo, Brazil

PUBLICATIONS

"Intonation Unit Boundaries and the Entrenchment of Bigrams: Evidence from Bidirectional and Directional Association Measures," *Review of Cognitive Linguistics*, 13(1): 191 – 219, 2015

Review of *Allah Made Us: Sexual Outlaws in an Islamic African City*, by Rudolf Gaudio. In *Gender and Language* 7(1): 139 – 143, 2013

"The Global Metastereotyping of Hollywood 'Dudes': African Reality Television Parodies of Mediatized California Style." In Mie Hiramoto (Ed.), *Media Intertextualities*. Philadelphia: John Benjamins. Pp. 31 – 55, 2012 (Reprint of Wahl 2010)

"The Many Voices and the Few: Contrastive Models of Authentication and Authorization through Register Talk in the Angolan Music Style Kuduro," Master's Thesis. June 2011

"The Global Metastereotyping of Hollywood 'Dudes': African Reality Television Parodies of Mediatized California Style," *Pragmatics and Society* 1(2): 209 – 233, 2010

FIELDS OF STUDY

Corpus Linguistics, Computational Linguistics, Cognitive Modeling, Child Language Acquisition, Psycholinguistics, Cognitive Linguistics, Sociolinguistics, Linguistic Anthropology, Portuguese Linguistics, Language and Media

ABSTRACT

The Distributional Learning of Multi-Word Expressions: A Computational Approach

by

Alexander Robert Wahl

There has been much recent research in corpus and computational linguistics on distributional learning algorithms—computer code that induces latent linguistic structures in corpus data based on co-occurrences of transcribed units in that data. These algorithms have varied applications, from the investigation of human cognitive processes to the corpus extraction of relevant linguistic structures for lexicographic, second language learning, or natural language processing applications, among others. They also operate at various levels of linguistic structure, from phonetics to syntax.

One area of research on distributional learning algorithms in which there remains relatively little work is the learning of multi-word, memorized, formulaic sequences, based on the co-occurrences of words. Examples of such multi-word expressions (MWEs) include *kick the bucket*, *New York City*, *sit down*, and *as a matter of fact*. In this dissertation, I present a novel computational approach to the distributional learning of such sequences in corpora. Entitled MERGE (Multi-word Expressions from the Recursive Grouping of Elements), my algorithm iteratively works by (1) assigning a statistical 'attraction' score to each two-word sequence (bigram) in a corpus, based on the individual and co-occurrence

frequencies of these two words in that corpus; and (2) merging the highest-scoring bigram into a single, lexicalized unit. These two steps then repeat until some maximum number of iterations or minimum score threshold is reached (since, broadly speaking, the winning score progressively decreases with increasing iterations). Because one (or both) of the 'words' making up a winning bigram may be an output merged item from a previous iteration, the algorithm is able to learn MWEs that are in principle of any length (e.g., *apple pie* versus *I'll believe it when I see it*). Moreover, these MWEs may contain one or more discontinuities of different sizes, up to some maximum size threshold (measured in words) specified by the user (e.g., *as _ as* in *as tall as* and *as big as*). Typically, the extraction of MWEs has been handled by algorithms that identify only continuous sequences, and in which the user must specify the length(s) of the sequences to be extracted beforehand; thus, MERGE offers a bottom-up, distributional-based approach that addresses these issues.

In the present dissertation, in addition to describing the algorithm, I report three rating experiments and one corpus-based early child language study that validate the efficacy of MERGE in identifying MWEs. In one experiment, participants rate sequences extracted from a corpus by the algorithm for how well they instantiate true MWEs. As expected, the results reveal that the high-scoring output items that MERGE identifies early in its iterative process are rated as 'good' MWEs by participants (based on certain subjective criteria), with the quality of these ratings decreasing for output from later iterations (i.e., output items that were scored lower by the algorithm).

In the other two experiments, participants rate high-ranking output both from MERGE and from an existing algorithm from the literature that also learns MWEs of various lengths—the Adjusted Frequency List (Brook O'Donnell 2011). Comparison of

participant ratings reveals that the items that MERGE acquires are rated more highly than those acquired by the Adjusted Frequency List, suggesting that MERGE is a performance frontrunner among distributional learning algorithms of MWEs. More broadly, together the experiments suggest that MERGE acquires representations that are compatible with adult knowledge of formulaic language, and thus it may be useful for any number of research applications that rely on such formulaic language as a unit of analysis.

Finally, in a study using two corpora of caregiver-child interactions, I run MERGE on caregiver utterances and then show that, of the MWEs induced by the algorithm, those that go on to be later acquired by the children receive higher scores by the algorithm than those that do not go on to be learned. These results suggest that, when applied to acquisition data, the algorithm is useful for identifying the structures of statistical co-occurrences in the caregiver input that are relevant to children in their acquisition of early multi-word knowledge.

Overall, MERGE is shown to be a powerful computational approach to the distributional learning and extraction of MWEs, both when modeling adult knowledge of formulaic language, and when accounting for the early multi-word structures acquired by children.

TABLE OF CONTENTS

ILLUSTRATIONS

FIGURES

TABLES

1. Introduction to Distributional Learning Algorithms

As computer power has experienced meteoric growth in recent decades, there has been a concomitant explosion in research on computational algorithms that take some input corpus, perform statistical operations on co-occurrences among certain transcribed units, and acquire knowledge of other linguistic structures that are latent in the input—that is, present yet not annotated or transcribed. I will refer to such computational approaches as distributional learning algorithms.

These algorithms are diverse, varying along a number of dimensions. First, the types of units among which co-occurrences are tracked, and the kinds of structures induced, can vary. For example, an algorithm may induce word boundaries (in a corpus in which word boundaries have been removed) by tracking co-occurrence regularities between phonemes or syllables (e.g., Goldwater et al. 2009). Alternatively, an algorithm may track co-occurrences between words, and thereby induce multi-word grammatical structures (e.g., Da Silva 1999).

The second dimension along which distributional learning algorithms may vary is whether the research goal is to model human cognitive processes, or whether there are some other non-cognitive research goals, with the algorithm being a means to this end. In such non-cognitive applications, the algorithm is not necessarily designed to model human cognition. As an example, an algorithm that distributionally learns multi-word structures may be used to identify useful idiomatic sequences for inclusion in a learner's dictionary. As we will see, however, often cognitive and non-cognitive distributional learning algorithms exhibit similarities in their architectures, despite different goals. For this reason, it is in

theory possible for a particular algorithm to be employed for both cognitive modeling and non-cognitive applications, even if this is not typically done.

Third, distributional learning algorithms differ in whether they focus on the acquisition of knowledge that is sequential/syntagmatic in nature, categorial/paradigmatic, or a combination of both. Finding word boundaries based on phoneme co-occurrence patterns or multi-word structures based on word co-occurrence patterns is a prototypically syntagmatic learning task. Conversely, an algorithm may cluster together words that occur in similar distributions of surrounding words to form emergent parts-of-speech (e.g., Reddington et al. 1998). This is a prototypically paradigmatic learning task. These paradigmatic and syntagmatic types correspond to three basic algorithmic architectures that are relevant to the current discussion: extraction and ranking algorithms, which learn syntagmatic representations; category acquisition/unsupervised part-of-speech tagging algorithms, which learn paradigmatic representations; and parsing/grammar induction algorithms, which typically learn both syntagmatic and paradigmatic representations. In the next section, I will go into some detail describing each of these architectures.

Finally, algorithms vary in terms of how much knowledge of linguistic information they have 'built-in.' Depending on whether an algorithm is designed for cognitive or non-cognitive purposes, there may be different motivations for why particular kinds of information are built-in or excluded. Thus, an algorithm designed to learn particular multi-word structures to be used as dictionary entries during lexicographic research may use a part-of-speech-tagged corpus in order to lower the rate at which erroneous, non-meaningful word sequences are learned. However, if the intent is to model some aspect of child language acquisition, it may not be desirable to presume that the child has existing

knowledge of parts-of-speech. That said, even in the case of cognitive questions there may be theoretical differences in terms of how much information should be built into a model: there has been a long debate in linguistics about whether and how much knowledge of grammar is innate versus induced through distributional learning. This is an issue whose literature I discuss in detail in chapter 2.

The centerpiece of the present dissertation project is a distributional learning algorithm. Entitled MERGE (Multi-word Expressions from the Recursive Grouping of Elements), it is an extraction and ranking algorithm that tracks word co-occurrences to learn particular kinds of multi-word, syntagmatic structures. Crucially, the algorithm works with a minimum of built-in linguistic information: it generates its output multi-word representations using an input corpus containing only word boundaries (as well as optional larger boundaries such as between speaker turns or at the boundaries between written texts). Because of the simplicity of the corpora it can be used on, the algorithm is viable for corpus-based research on under-documented languages with limited corpus resources. In addition, when used with child language data, the algorithm is designed to be able to acquire representations of multi-word structures that are compatible with the structures acquired by children learning language for the first time. Thus, MERGE can also serve as a tool in cognitive-based research.

1.1 Why is MERGE Needed?

In the last section, I outlined some major dimensions along which distributional learning algorithms vary, and I briefly situated the algorithm presented in this dissertation by reference to these dimensions. In this section, I provide further detail, elaborating on the

combinations of features along these dimensions that have been embodied in different algorithms in the literature. Through this elaboration, I will provide motivation for why MERGE embodies the particular combination of features that it does, and why its unique design represents an important and timely contribution to the field. The reader should note that the different combinations of features that the various models represent is complex, yet it is necessary to delve into this complexity in order to understand the gaps in the existing research that MERGE addresses. The reader should also note that, with respect to the types of linguistic units among which co-occurrences are tracked, the approaches addressed in the proceeding discussion are virtually entirely based on words.

1.1.1 Non-Cognitive Algorithms

As I mentioned in the previous section, algorithms have typically been distinguished according to whether they are used for cognitive or non-cognitive research goals (even if ultimately MERGE is designed to serve both goals). Within the non-cognitive domain, perhaps the two most significant foci of research are unsupervised part-of-speech tagging and collocation extraction and ranking algorithms.

1.1.1.1 Unsupervised Part-of-Speech Taggers

As indicated above, unsupervised part-of-speech taggers cluster together words that occur in similar distributions of surrounding words, with the clusters being equivalent to lexical categories (e.g., words that frequently occur after *the* and *a* and before a relative pronoun may be clustered together, forming an emergent noun category; Brown et al. 1992; Clark 2003; Biemann 2006). The resulting corpus, annotated with part-of-speech tags, may

then be useful for a wide array of corpus-based research across a variety of linguistic subfields as well as natural language processing applications. (Note that such 'unsupervised' taggers differ from 'supervised' taggers, which have access to gold-standard lists of word types and their parts-of-speech and then assign tags to tokens based on this list; such supervised taggers would not count as distributional learning algorithms). Unsupervised taggers are particularly useful when such gold-standard information may not already available, such as when working with an underdocumented language. These unsupervised taggers only require built-in word boundary information to operate.

1.1.1.2 Collocation Extraction and Ranking Algorithms

While part-of-speech taggers represent non-cognitive paradigmatic learning algorithms, collocation extraction and ranking algorithms embody non-cognitive syntagmatic learning. But what is a collocation? The term is common within corpus-linguistic research, but it is roughly interchangeable with a number of popular terms, including multi-word expressions (or MWEs, used here), multi-word units, formulas, lexical bundles, *n*-grams, and phraseologisms, among others. Prototypical examples include *as a matter of fact*, *apple pie*, or *lift up*. While precise definitions vary from researcher to researcher, the phenomenon is broadly characterized as a sequence of words that exhibits some degree of formulaicity and conventionalization within the language. Note that such multi-word expressions, or MWEs, are not necessarily coextensive or synonymous with syntactic constituents; discourse particles *I mean* and *I know* are traditionally considered to cross constituent boundaries, yet they are typically considered to be MWEs. In chapter 2, I

will review in detail literature addressing the definition, identification, and processing of MWEs.

Collocation extraction and ranking algorithms work by generating a list of MWEs present in a corpus that is ranked according to the scores assigned to each MWE by some statistical metric. The statistical metrics used as the basis of ranking are numerous, and they take into account in different ways combinations of information about the co-occurrence frequency of the target MWE, the individual frequencies of the component words of the MWE, and the size of the corpus (see Evert 2005; Pecina 2009). Such co-occurrence frequency-based measures function as a proxy for the criterion of word sequence formulaicity and conventionalization, a correlation that has empirical backing and which, again, I discuss in chapter 2. The resulting ranked MWEs then have numerous applications, e.g., for lexicographers identifying multi-word sequences that should be included as dictionary entries (e.g., Schone and Jurafsky 2001); second language acquisition researchers and pedagogical development experts interested in the conventional word sequences that make language native-like, thus enhancing second language learning materials (e.g., Simpson-Vlach and Ellis 2010); and researchers interested in using MWEs as a diagnostic for dialectal/genre/variety differences, as well as language change in progress (e.g., Gries and Mukherjee 2010).

Typically, MWEs of one or more particular sizes (often bigrams and/or trigrams) are extracted, scored, and sorted. Furthermore, the component words of the MWEs are typically adjacent. In such an approach, MWE size and component word adjacency represent linguistic information built into the algorithm. However, there has been great recent interest in algorithmic approaches that do not specify a priori the size of the MWE to be extracted,

nor that the words comprising it must be directly next to one another (e.g., Da Silva et al. 1999; Gries and Mukherjee 2010; Brook O'Donnell 2011). On the one hand, MWEs can come in all sizes—not just sizes specified to be extracted. Restricting extraction to specific sizes can result in sequences that are in fact merely fragments of actual, larger MWEs, or that contain actual MWEs but which are themselves not MWEs. On the other hand, MWEs may contain one or more gaps of different sizes, and extraction of only adjacent sequences misses MWEs of these types. And while there has been some work in the development of collocation extraction and ranking approaches that learn MWEs of various sizes with and without gaps intervening between component words, only a few models have been proposed. Moreover, it is not yet clear which, if any, of these algorithms perform better than others in relation to one or more of these issues.

1.1.2 Cognitive Algorithms

In contrast to non-cognitive algorithms, which are designed to learn latent structures in corpora for other research applications, cognitive algorithms are designed either to embody theorized design features of human distributional learning, or at least to acquire representations that are compatible with the output of human distributional learning. Ultimately, to the extent that these algorithms acquire human-like representations, evidence is provided for the usefulness of the theorized mechanisms. Conversely, if these algorithms do not perform as hypothesized, evidence is provided that something needs to be revised: either the algorithm parallels human learning, but the input it is receiving does not, or the algorithm does not parallel human learning, and thus theories about the design and operation of particular distributional learning mechanisms in humans may need to be adjusted.

With these fundamentally different goals, one might expect major differences in the algorithmic approaches taken between non-cognitive- and cognitive-oriented algorithms—and there are numerous differences between these two research domains. At the same time, however, there are important areas of overlap. This is because architectures that work well for non-cognitive applications may operate in cognitively realistic ways or yield results that are compatible with cognitive findings; conversely, cognitive algorithms may function well for non-cognitive applications.

1.1.2.1 Category Acquisition Algorithms

In no area is this overlap more obvious than in the case of category acquisition algorithms (e.g., Reddington et al. 1998; Parisien et al. 2008; Alishahi and Chrupala 2009; Chrupala and Alishahi 2010), which are essentially a cognitive rebadging of the previously discussed non-cognitive unsupervised part-of-speech taggers. Both types of algorithms work in the same fundamental way: words are clustered together into paradigmatic categories on the basis of similar distributions of surrounding context words. Nonetheless, there are certain noteworthy differences, and these differences have to do with optimization for practical non-cognitive applications, on the one hand, or cognitive realism, on the other. For example, part-of-speech taggers function in a batch manner (i.e., they process the whole corpus at once), and they usually use clustering methods that require the researcher to specify the number of part-of-speech categories beforehand. In contrast, category acquisition algorithms often function in an incremental, utterance-by-utterance fashion, and they induce the number of part-of-speech categories in a bottom-up way; both of these attributes reflect more closely the acquisition task faced by a child.

Other than these differences, however, category acquisition and part-of-speech tagging algorithms are largely the same in design, and ultimately they may learn very similar output representations. This illustrates the potential viability of a particular architecture for both non-cognitive and cognitive applications.

1.1.2.2 Parsers/Grammar Induction Algorithms

While paradigmatic learning approaches find close parallels in both cognitive and non-cognitive approaches, the syntagmatic learning-based extraction and ranking architecture—so fundamental to collocation research—has found minimal penetration in cognitive frameworks; in fact, up until now, cognitive and non-cognitive approaches to syntagmatic learning have remained fairly distinct. This may be because the mechanisms engendered by the extraction and ranking architecture represent a fairly high level abstraction away from the conditions in which children are actually learning.

Thus, parsing/grammar induction algorithms have been popular instead, as their design more closely approximates these conditions of children's acquisition (e.g., Klein and Manning 2002; 2004; Solan et al. 2005; Bod 2009). First, the batch approach in which extraction and ranking algorithms typically operate is not the manner in which children acquiring language for the first time receive linguistic input. In contrast, parsing algorithms are typically incremental, operating utterance-by-utterance as they move through the corpus. Second, the typical a priori specification of *n*-gram sizes to be extracted is not ideal for modeling human learning, as humans are able to determine the various lengths of multi-word structures in a bottom-up fashion. However, parsing algorithms are able to learn structures of varying lengths in this way. Finally, while extraction and ranking algorithms

learn multi-word, syntagmatic structures, parsing algorithms also typically acquire knowledge about paradigmatic relations between these syntagmas. In so doing, they learn a grammar, which can be used not only to parse existing utterances but also to generate novel ones. Clearly, a child must likewise ultimately arrive at such a productive knowledge of language.

The general procedure of a parsing/grammar induction algorithm is as follows: for a current utterance (or other current incremental window), it generates multiple ways of exhaustively parsing smaller units into larger units. Then, it scores the candidates according to particular criteria and chooses a winner. Within this general framework, there are numerous approaches to how this parsing procedure can be more specifically implemented. One very popular and successful model that exemplifies parsing is Unsupervised Data-Oriented Parsing (UDOP; Bod 2009). This incremental (utterance-by-utterance) grammar induction algorithm finds the best parse of an utterance by first operating over a training partition of a corpus; generating, for each utterance, all possible hierarchical binary-branching trees over the words; and then storing each subtree fragment. These stored subtrees may be individual words, or they may be multi-word structures. They may be continuous, or they may contain gaps. Then, during test, this bank of subtree fragments is used to generate candidate parses of test utterances, with the winning parse for an utterance being the most probable one—probability is calculated based on the frequency of the subtrees from training. Importantly, as mentioned above, parsing/grammar induction algorithms learn both syntagmatic and paradigmatic representations. While the lexicon of subtrees functions as knowledge of individual words and multi-word units (syntagmatic knowledge), the hierarchical tree structure over these subtrees—which governs the ways in

which they can be combined—functions as knowledge of categories (paradigmatic knowledge).

1.1.2.3 Extraction and Ranking for Cognitive Goals

Despite the dominance of parsing models in cognitive approaches to distributional learning algorithms that acquire syntagmatic structure—and their less abstract relationship to the environmental conditions faced by the child—extraction and ranking approaches can still offer powerful cognitive insights. One extraction and ranking algorithm used to addressed cognitive questions, by Swingley (2004), is not used to explore the acquisition of multi-word structures from individual words, but rather the acquisition of words from the co-occurrence of syllables. Word segmentation algorithms such as this one are popular within distributional learning-based cognitive research, yet, like computational approaches to multi-word knowledge, they fall almost entirely within the conventional parser framework, with the exception of this approach. Swingley's algorithm works by extracting syllable *n*-grams of different sizes from a corpus, scoring, and ranking them. Higher scoring *n*-grams are then shown more reliably to be words than lower scoring *n*-grams, thus providing evidence for the statistical tracking of syllable co-occurrences as potentially informative for word boundary acquisition.

It is important to point out how this extraction and ranking approach provides evidence in a fundamentally different way than a parsing approach to segmentation does. Parsers score candidate parses, choose a winner, and then their performance is evaluated by comparing the winning parse against some gold standard. Generally, there is no attempt to correlate the model's commitment to parses (quantified by parses' scores) with their

closeness to the gold standard. In other words, runner-up parses are not checked to see whether they get farther and farther from the gold standard as their score decreases.[1] Yet this is precisely the kind of logic that underlies extraction and ranking algorithms. Proof for the compatibility of the output of the algorithm with human-like representations is provided by the correlation between scores of induced representations and their varying closeness to a gold standard. Moreover, this is actually likely to be closer to what humans do when learning new structures and assigning parses: certainly, a winning structure is selected, but humans remain aware of alternative structures and their degree of certainty in making a choice. Thus, the scored output of extraction and ranking approaches is compatible with the varying certainty/strength of human representations of linguistic structures. But despite the unique perspective that extraction/ranking can provide on cognitive questions, this architecture has not been used to investigate multi-word acquisition.

In addition, unlike parsing/grammar induction algorithms, extraction and ranking algorithms do not necessarily render an exhaustive parsing of the input corpus into learned units. Specifically, if there is a minimum statistical threshold above which MWEs must be scored to be extracted, there will be stretches of the corpus that will not correspond to any extracted MWEs (where would-be MWEs were scored too low to be extracted). Along similar lines, at the early stages of acquisition, it is likely that children leave sizable stretches of input unparsed, and only later are they able to exhaustively assign structural representations to each utterance heard. In this way, the output of extraction and ranking

[1] Because parsing algorithms score multiple candidate parses in order to choose a winner, it is not to say that runner-up parses could not be checked to see whether they get farther and farther from the gold standard as their score decreases. Simply, in parsing research this approach is not typically taken.

approaches is again perhaps compatible with human-like representations (at an early stage of acquisition) in a way that the output of parsing/grammar induction algorithms is not.

Ultimately, it is important to remember that the incremental nature of parsing/grammar induction algorithms allows them to typically be understood to be *models* of human cognition—that is, they are designed to emulate moment-my-moment processes in the mind. In contrast, the lack of a temporal dimension in extraction and ranking approaches represents a high level of abstraction away from the learning conditions faced by children, and thus such approaches may not typically be thought of as true cognitive models. And while the algorithmic procedure itself may thus not be tightly compatible with human processing, the acquired output representations may indeed be compatible with human knowledge. To the extent that this is true, I have argued that extraction/ranking algorithms offer a unique and potentially highly informative perspective on cognitive questions surrounding human distributional learning, which conventional parsing/grammar induction approaches do not offer.

1.1 Brief Overview of the Architecture of MERGE

In 1.1.1.2 and 1.1.2.3, I ended by highlighting what I see to be two underexplored areas in the existing literature on distributional induction algorithms. In the non-cognitive literature, there is still inadequate development of collocation extraction and ranking algorithms that induce MWEs with gaps, and whose length is not specified beforehand. In the cognitive literature, there is a dominance of parsing/grammar induction approaches for the modeling of multi-word learning, while there are to my knowledge no extraction and ranking algorithms used to examine cognitive questions, despite Swingley's (2004)

demonstration of the virtue of such an approach to addressing such questions in his work on word segmentation. Thus, traditionally non-cognitive collocation extraction and ranking techniques are called for to address this latter underexplored area in cognitive research. Conversely, I suggest, traditional cognitive parsing approaches to multi-word learning offer innovative solutions to address the former underexplored area in non-cognitive research: remember that grammar induction algorithms are typically able to induce structures whose size is not pre-specified, and in some cases with gaps (such as in the case of UDOP).

I argue, then, that a singular architecture that draws features from extraction and ranking as well as parsing approaches is ideally suited to address both of these underexplored areas simultaneously. Of course, a large degree of overlap in cognitive and non-cognitive approaches is not without precedent; as mentioned in 1.1.2.1, at the level of paradigmatic learning, unsupervised part-of-speech taggers and category acquisition algorithms are nearly totally convergent in their design, with only minimal differences.

Here, I thus extend this trend of convergence to the realm of syntagmatic learning through the development of the MERGE algorithm. The algorithm embodies an extraction and ranking architecture that works iteratively using a batch approach. It operates by first extracting all continuous and discontinuous bigrams from a corpus (up to some maximum discontinuity size). Next, it scores and ranks them according to a statistical measure of association. Then, it merges the top-ranked bigram into a single 'lexical' item. For example, the adjacent words *for* and *example* might become the single word *for example*. At the final step, the algorithm replaces all corpus tokens of the bigram with the new merged representation. Then, steps 1 through 4 are repeated until some minimum score threshold is reached. Because the output of previous merges can be the input to later merges, the

algorithm can learn MWEs of theoretically any length, with or without one or more discontinuities. Ultimately, MERGE generates a list of scored and ranked MWEs of various lengths, and with possible gaps of varying sizes.

MERGE's approach of recursively joining bigrams to find larger and larger units exhibits parallels to UDOP, in which multi-word grammatical structures are learned using binary-splitting trees. There are a couple of advantages that such an approach offers over other plausible approaches, such as one in which building blocks of different sizes (not just bigrams) were merged to form larger structures, or an approach in which the algorithm simply considers competing output representations of different sizes (rather than building them up from smaller pieces). First, the current approach is relatively computationally tractable. That is, the set of structures necessary for consideration at any one point in order to arrive at the output representations is relatively constrained, while still allowing for output of virtually any size and shape. Second, MERGE's approach is relatively statistically tractable. Generally, the statistical measures developed for tracking co-occurrence regularities between words are optimized for bigram relationships.

An additional benefit of the MERGE approach is that it is a mechanistically simple approach relying on a minimum of built-in information—it does not rely on knowledge of part-of-speech categories or pre-specified sizes of MWEs. And while, as discussed, the algorithm is not an incremental model of language acquisition, this minimization of built-in information nevertheless allows for more meaningful comparisons between the output of the model and the learned representations of humans (since children likewise do not have access to extensive built-in information). At the same time, it is important to remember that MERGE does rely on some built-in knowledge—specifically, word boundary information.

However, as will be discussed in chapter 2, it is arguably reasonable to assume that a child acquiring a MWE, for example, will already possess some inductive knowledge about word boundaries by the time s/he is learning multi-word structures.

On the other hand, MERGE's mechanistic simplicity and minimum of built-in information gives it great viability as a non-cognitive corpus tool for lexicographers, second language acquisition/pedagogy researchers, scholars of language change and register/genre differences, and others who do collocational research. Specifically, it can be used on very simple corpora that are not enriched with part-of-speech or syntactic information, making it useful for corpus-based research on languages and genres that vary widely in the extent of their available digital resources.

MERGE's powerful approach to building multi-word expressions addresses two of the aforementioned most pressing issues in non-cognitive approaches to collocation extraction and ranking: it acquires sequences of various lengths, as well as sequences that may include gaps in them. At the same time, as a first foray of extraction and ranking-type algorithms into addressing cognitive questions surrounding multi-word learning, the inclusion of these features brings the approach closer to the learning ecology faced by the child than would be the case with a traditional collocational approach in which *n*-gram size(s) and gap count(s) are specified beforehand.

That said, other features typical of extraction and ranking-type algorithms are maintained. For example, MERGE is a batch approach, which is opted for since the statistical measures employed are not designed to work incrementally. In addition, MERGE acquires only syntagmatic representations, while parsers/grammar induction algorithms typically acquire both syntagmatic and paradigmatic knowledge, and thereby are able not

only to parse existing utterances, but also generate novel ones. Finally, MERGE does not commit to a single, exhaustive parse of every utterance. Again, however, the purpose of extraction and ranking is not to do this; rather, it is to quantify the degree of goodness of representations learned from a corpus.

1.3 Looking Ahead

In this chapter, I have introduced the basic theoretical concepts necessary to motivate the MERGE algorithm. In addition, I have given a brief description of how the algorithm functions. In the next chapter, I will provide a more thorough review of the literature relevant to the current study. Then, in chapter 3, I will describe the algorithm in detail.

Chapters 4 and 5 present empirical studies using MERGE. In chapter 4, I report human subject rating experiments designed to evaluate the efficacy of MERGE in producing reasonable multi-word expressions, and comparing the performance of the algorithm to the performance of another collocation/extraction algorithm from the literature. Then, in chapter 5, I report a corpus-based study that evaluates MERGE as a cognitive model of children's acquisition of multi-word expressions, using a corpus of child-directed speech. Finally, in chapter 6, I discuss conclusions and directions for future research.

2. Distributional Learning and Formulaic Language in Linguistic Theory

In the introductory chapter, I gave two major promissory notes. I said I would elaborate on the status of distributional learning within the context of the history of linguistic theory, and on the role of computational modeling in shaping this theory. I also said I would discuss the linguistic unit that MERGE learns, variously referred to as *collocation*, *phraseologism*, and *multi-word expression*, among others (The latter term is the one I will use in my own empirical work with MERGE, while I will use the term *formulaic language* as a coverall for the general linguistic phenomena that all of these terms collectively refer to). In the first two sections of this chapter, I address these promissory notes.

In order to understand the importance of the concepts of distributional learning and formulaic language to linguistics in general and to the current dissertation in particular, it is necessary to begin with a brief discussion of generative linguistics, a theoretical framework that was dominant throughout the second half of the 20th century. Section 2.1 is dedicated to this discussion. Specifically, I focus on two aspects of generative linguistics that are especially relevant to this dissertation: the notions of universal grammar and the poverty of the stimulus, on the one hand, and the notion of distinct syntactic and lexical systems, on the other.

In section 2.2, I then discuss evidence against these cornerstone components of generative theory. First, in section 2.2.1, I provide arguments against universal grammar and the poverty of the stimulus, focusing especially on how distributional learning offers a

strong critique of these ideas. It is here that I discuss previous experimental research on distributional learning. Then, in section 2.2.2, I discuss research on formulaic language. I will begin by elaborating on how formulaic language is defined and how common it is in discourse. I will then review empirical studies of formulaic language, which effectively expose the flaws of the notion of a clear syntax/lexicon divide.

In the third section of this chapter, I begin by discussing what has been (typically implicitly) treated as a theoretical incompatibility between the research threads of distributional learning and formulaic language, and thus a major reason for their belonging to separate empirical traditions. However, I will argue that such theoretical incompatibility is in fact non-existent. Indeed, research on the computational modeling of the distributional learning of formulaic language brings these lines or research together quite successfully. This section is thus primarily devoted to the review of these studies, some of which were mentioned in chapter 1. Through their union of these two important trends, computational models have provided a more complete alternative generative concept of how language is acquired by and organized in the human mind. MERGE belongs to this research trend, and I end the chapter by revisiting the ways in which MERGE addresses gaps in the existing computational research.

2.1 Generative Linguistics

Generative linguistics is a theoretical framework that emerged in the 1950s and 1960s, spearheaded by Noam Chomsky's seminal books *Syntactic Structures* (1957) and *Aspects of the Theory of Syntax* (1965). Generative linguistics became highly influential because, at the time, it offered a model of language differing a great deal from existing

theories, and it seemingly provided greater explanatory power. In its earliest forms, it was conceived of as a set of tools for describing the structure of language, as was not seen as a description of the cognitive substrate of language. Nonetheless, some of its central ideas were later adapted as part of models of how the mind represents and processes language (e.g., Pinker 1984).

There are a number of important concepts within the then developed generative paradigm, including a distinction between surface structure and deep structure, syntactic transformations, the distinction between performance and competence, universal grammar and its relationship to the notion of the poverty of the stimulus, and a distinction between syntax and the lexicon. I will focus on these last two, as they relate most directly to distributional learning and formulaic language.

2.1.1 Universal Grammar and the Poverty of the Stimulus

Universal grammar is the common, innately specified grammatical substrate that underlies all human languages. The notion of such an inborn set of grammatical knowledge became popular in large part because of the issue of child language acquisition. That is, based on the input children receive, how do they arrive at a mature productive grammar?

One possibility that presented itself early was, in fact, distributional learning – that is, that children track co-occurrences among percepts to induce linguistic structure By the time generativism emerged, distributional information had already long been recognized as informative for linguistic categories. Linguists describing languages use distributional tests to identify syntactic constituencies, morphological slots, and phonemic categories, for example. Researchers later extended distributional tests as a possible account of how

language is acquired (Maratsos and Chalkey 1980; Maratsos 1981). However, there was also a strong rejection of distributional learning as a possible cognitive acquisition mechanism (Pinker 1984, pp. 49 – 50). Specifically, it was believed that there was a *poverty of the stimulus*—that the input children receive is too noisy and too incomplete for the tracking of co-occurrences between percepts to be sufficient for the full acquisition of language.

As noted by Reddinton et al. (1998), Pinker (1984, pp. 49 – 50) specifies perhaps the most oft-cited list of arguments against distributional learning as the mechanism of child language acquisition. First, he claims that there are so many possible co-occurrence relationships available in the input that they would overwhelm any learning mechanism based on distributional regularities. In addition, he claims that co-occurrence regularities engender spurious correlations. For example, consider the following three utterances that the child could hear: *John eats meat*; *John eats slowly*; *The meat is good*. Based on these utterances, Pinker argues that the child would have no way of knowing that the utterance *The slowly is good* is ungrammatical. Pinker also argues that looking through all possible co-occurrence regularities is wasteful, since languages only allow certain correlations to be relevant in reflecting structural patterns. Finally, he claims that syntagmatic co-occurrences alone do not reveal the abstract, hierarchical phrase-structure rules that generative grammarians argue underlie all languages.

Because of these alleged fatal flaws of a distributional learning approach to acquisition, for decades the main generative answer to the acquisition problem revolved around what are known as parameters (see Tomasello 2005). These are innately specified, bivalued settings that specify grammatical phenomena theorized to be a part of universal grammar. For example, one of the most widely researched parameters (and one of the very

few parameters that the generative literature could agree on; see below) is the head direction parameter, which specifies head-dependent order in a language. The idea here is that a child will set this parameter early in life (and it will remain then set for the rest of their life). This setting then constrains how the child interprets input: a child will know that, given a phrasal sequence, the head lies at one or the other end of this sequence. In other words, parameters are seen as allowing a child to cut through the unnecessary and spurious distributional correlations that nativists argue make distributional learning unfeasible.

Pinker (1984, pp. 37 – 39) notes, however, that what is known as the bootstrapping (or linking) problem must be solved: that is, how do the surface forms of any of 7000+ natural languages get aligned to the innately specified parameters? The most widely accepted answer to this question is Pinker's theory of semantic bootstrapping (Pinker 1989, pp. 248 – 253). This theory argues for an innate linking between grammatical categories and semantic categories. For example, a child is born with an innate linking rule that specifies that subjects are prototypically agents. In this way, if a child 1) perceives their dog barking, 2) knows their dog is named *Fido*, 3) hears their mother say *Fido is barking*, then 4) because of the agent-subject linkage, they are able to set the head direction parameter.

2.1.2 Syntax and the Lexicon

Another key tenet of generative linguistics has been that there is a sharp distinction between syntactic rules and the lexicon (Pinker 1999, pp. 1 – 19). Syntactic rules serve as the basis for assembling items drawn from the lexicon into sentences. Furthermore, the traditional view has been that repeated usage events do not affect the memorized representations of these rules. Meanwhile, the lexicon is a repository of stored items, which

are traditionally considered by generativists to be words and a relatively small set of multi-word idioms, whose semantic and/or syntactic non-compositionality necessitates their storage rather than creative assembly. Because of the relatively small role accorded to idioms, the generative approach emphasizes the creativity of most linguistic productions.

2.2 Answers to Generative Problems

While from the 1960s the generative framework became dominant, the views just described have fallen into increasing disfavor over the last few decades. This is because a number of logical arguments as well as descriptive, experimental, and computational studies have shown generativist ideas to be largely implausible. In what follows, I review a large body of research that has made this point. First, I discuss problems with the notions of universal grammar, parameters, and the poverty of the stimulus, and I provide a review of literature on distributional learning, which is a viable alternative to innately specified linguistic categories. Then, I review research on formulaic language, which has been largely responsible for dispelling the fallacy of the syntax/lexicon divide.

2.2.1 Distributional Learning in Linguistic Theory

Numerous criticisms have been directed at the generative notion of parameters. In order for semantic bootstrapping to work, there must already be word-meaning mappings (Tomasello 2005). In the example I gave above, the child already knew that *Fido* was a word that referred to the family dog. But most words are not heard in isolation, so some amount of segmentation and word meaning acquisition must go on before semantic bootstrapping can take place. Generativists sometimes admit that distributional learning may be necessary at

early stages to 'get the ball rolling,' but once parameter setting takes place such distributional learning is no longer necessary. They argue that this dual account is not unparsimonious because of the constraints that parameter-setting places on the search space.

However, it appears that even after children acquire phrases with ordered heads/dependents, they still make ordering mistakes later, so it cannot be that a parameter gets suddenly set as soon as the right evidence is provided (Tomasello 2005). Relatedly, young children often find verb arguments reflecting non-prototypical thematic roles as easy—if not easier—to produce than arguments reflecting prototypical roles; this casts into serious doubt the usefulness of semantic bootstrapping for solving the linking problem, and whether linking rules can be considered innate (Bowerman 1990).

The actual developmental path that children follow for the acquisition of productive grammatical structures is as follows: initially, children make few errors; then, there is a period of overgeneralization; later, children correct overgeneralization and converge on a mature grammar. Such a trajectory is not commensurate with sudden parameter setting; instead it reflects a process whereby children learn structures on an item-by-item basis, slowly generalizing across these items (to the point of overgeneralizing), and then eventually correcting for overgeneralization as additional evidence continues to come in regarding irregular forms.

But most damning to parameters is extensive research on typologically diverse languages over the past half-century that has shown that many of the proposed parameters, which have been based primarily on work on English and related languages, do not hold across diverse language families (see Tomasello 2005). Moreover, two of the biggest proponents of parameters have separately proposed lists of parameters that are nearly

entirely disjoint (Baker 2001; Fodor 2003; see Tomasello 2005). Another (previously) strong proponent of this framework has admitted that the observed patterns in language acquisition cannot be accounted for by any one of the proposed parameter-based accounts (Hyams 2011). So criticized have parameters been that major contemporary nativist theories, such as the minimalist program, do not draw on them (see Chomsky 1995). Instead, Chomsky now claims that the one linguistic universal is recursion (but see Everett 2005).

But more interesting and relevant to the current dissertation is how arguments leveled by nativists against distributional learning have come up short. Returning to Pinker's famous criticisms of distributional learning, Reddington et al. (1998) offer logical counterarguments to each. Pinker's argument that looking through all possible correlations is wasteful, since languages permit only certain correlations and not others, can be dismissed on logical grounds. Specifically, generative grammar is itself wasteful given linguistic diversity; if the space of innately specified parameters necessarily accounts for the grammatical possibilities across all the world's languages, then in the mind of the speaker of any single language there will be numerous unnecessary parameters, since no one language makes all possible grammatical distinctions.

Pinker's other major arguments constitute empirical claims, and thus they must be tested (Reddington et al. 1998). First, it is not a priori clear that syntagmatic co-occurrences do not reveal grammatical abstractions. Similarly, Pinker's claim that the vast number of possible co-occurrence relationships in the input will overwhelm the learner's distributional mechanisms requires empirical testing. It may be that that the distributional learning mechanisms that we use are constrained in particular ways so that combinatorial explosions of possible co-occurrences are not an issue.

What is more, both of these claims were based on outdated, theoretically unsophisticated understandings of distributional information. Major advances in statistics over the past half century warrant a reexamination of the distributional learning hypothesis. In fact, there has been a wealth of experimental and computational work that has explored the degree to which distributional learning can serve the acquisition and mental representation of language. To preview the results (which I discuss immediately below), this work has shown time and again that syntagmatic co-occurrences can indeed give rise to speakers' knowledge of linguistic structures, and combinatorial explosions of co-occurrences end up not being an issue. In section 2.2, I discuss experimental studies of distributional learning. Later, in section 2.3, I will return to a discussion of distributional learning in the context of computational modeling (after I have discussed the phenomenon of formulaic language).

2.2.1.1 Empirical Work on Distributional Learning

The vast majority of experimental research on distributional learning has involved artificial languages. In an artificial language experiment, an adult or child is exposed to sequences of linguistic units (often syllables). These sequences have no meaning, at least in the native language of the participants. Crucially, these sequences are designed in such a way that the co-occurrence regularities among units exhibit particular statistical properties. Participants are then tested to see whether they have learned particular kinds of relationships between the units as a function of the statistical properties that obtain.

Arguably the most famous and seminal of such artificial language studies are a series completed by Saffran, Aslin, and Newport (1996a; 1996b; 1998). In these studies,

experimental participants are exposed to syllable sequences in which the syllable-to-syllable transitional probabilities (TPs) are carefully controlled (TP refers to the probability of a next syllable given the current syllable). At test, the participants are then able to recognize trisyllabic sequences that they had heard in which internal TPs were relatively high as cohesive, word-like units. In contrast, trisyllabic sequences containing a low syllable-to-syllable transitional probability are not recognized as words. Saffran and colleagues show this pattern to be true for both infants (1996a) and adults (1996b). The infant results are particularly important theoretically because they show that this distributional learning capacity exits at the developmental stage at which humans could rely on it to acquire their first language.

In a follow-up study, the research group shows that it is indeed the transitional probability between syllables that study participants are attending to, and not to raw string frequency (1998). These results are also important because, at that time that they emerged, the effect of frequency on linguistic representations in the mind had already been established. However, contingency of co-occurrences, measured through more sophisticated metrics such as TP, had not yet been firmly established as an important aspect of language learning. In the present dissertation, contingency-based statistics will be used instead of simple frequency. I return to the issues of frequency effects on mental representation and the statistical operationalization of co-occurrences below.

These watershed studies launched a cottage industry of artificial language experiments in which various scholars explored the possible relevance of statistical correlations between word boundaries and other types of cues besides just syllable TPs. Cues identified as important include prosodic lengthening at the ends of words (Saffran et al.

1996b); stress placement (with stress more often expected on the initial syllable of words in English due to the language's trochaic bias; Curtin et al. 2005; Thiessen and Saffran 2003); phonotactic constraints (that is, which phoneme sequences may occur within words versus at the boundary between words; Hockema 2006; Mattys et al. 2005); using knowledge about already segmented words in an auditory stream to make hypotheses about the segmentation of the residue (Mattys et al. 2005); and combinations of these cues (Johnson and Jusczyk 2001).

One of the main criticisms that has been leveled against this body of research is that simply showing that infants can learn artificial language structures via statistical learning does not mean that they do learn natural languages in this way. To respond to this criticism, recently Pelucchi and colleagues (2009a; b) have launched a new research program based on distributional learning experimental tasks but using natural language stimuli. In one experiment from their 2009a study, infants of around 8½ months who are acquiring English as their first language are exposed to Italian sentences during a familiarization phase. Then, during a test phase, they are exposed to bisyllabic words in isolation. While these words occur during the familiarization phase with equal frequency, the syllables that constitute them do not. As a result, a subset of the target words exhibits an internal syllable-to-syllable TP of 1.0, while the other subset exhibits a TP of just .33. The infants recognize the words with the higher internal syllable-to-syllable TP as more familiar, despite the fact that all words are equally frequent during familiarization. In this way, Pelucchi et al. (2009a; b) are able to show that results similar to those obtained using artificial languages hold true for natural language stimuli. Indeed, infants use co-occurrence probabilities to segment words. In general, results like these are hugely important for verifying that statistical learning is a

fundamental mechanism in language acquisition processes.

As the segmentation of words has become thoroughly explored over the past two decades, attention has shifted to other problems of linguistic induction via statistical learning, including the learning of phoneme categories as well as multi-word structures of grammar (see Gomez and Gerken 2000). Together with the word segmentation literature, these studies have been instrumental in showing that nativist claims regarding the poverty of the stimulus are inaccurate across all levels of linguistic structure.

For the purposes of the present dissertation, it would seem that the research on the distributional learning of multi-word grammatical structures would be most relevant. However, the kinds of multi-word structures studied in the artificial language literature and the kinds we are concerned with here are actually fairly different. The artificial language experiments have tended to focus on how infants can use co-occurrence statistics to learn abstract word classes (Reeder et al. 2013), grammatical phrases comprising these word classes (Thompson and Newport 2007), and rewrite rules for concatenating these phrases into sentences (Saffran 2001). These studies have yielded invaluable evidence against one of generativism's harshest critiques of distributional approaches—that distributional learning performed on syntagmatic structures cannot yield productive grammatical knowledge that can be used to utter sentences never before heard. But while these studies have yielded stunning results, they ironically take a rather conservative, incomplete view of sentence generation, focusing on highly creative utterances in which dependencies between abstract form classes (word classes and phrasal types) define structure, while dependencies between specific word types take a back seat.

However, as we will see below, such dependencies between specific word types

instantiate partially or fully lexicalized multi-word structures, and sentences are often suffused with these types of structures. In fact, the distributional learning experiments perhaps most relevant to the current dissertation are the earlier word segmentation studies. As described above, these studies show that the co-occurrences between some elemental types can be used to learn new, larger types. While the researchers who have conducted these studies have (typically) defined the types rather narrowly as syllables and the emergent types as words, the elemental types could equally be thought of as words and the emergent types as lexicalized multi-word sequences. Thus, while distributional learning experiments have not explicitly dealt with lexicalized multi-word structures, the existing architectures tested in word segmentation experiments are compatible with the statistical dependencies of such multi-word structures. In the next section, I discuss these kinds of multi-word structures.

2.2.2 Formulaic Language in Linguistic Theory

Just as distributional learning research has been important in challenging generative ideas about the necessity of innately specified, universal grammar and claims about the poverty of the stimulus, the phenomenon of formulaic language has been important in challenging the generative idea of separate syntactic and lexical systems. This is because formulaic language, as it is widely understood, calls into question basic assumptions about the contents of the generative lexicon. Under the traditional generative model, the only kinds of stored multi-word sequences in the lexicon are semantically and syntactically non-compositional ones, such as idioms (Pinker 1994, p.142 – 143). Moreover, these structures are considered few in number compared to the dominance of compositional multi-word

sequences assembled online via syntactic rules. However, as I will show in the following subsections, stored formulaic sequences appear to be omnipresent in discourse, and, what is more, they are often semantically and syntactically compositional.

Below, I first discuss what formulaic language is. Then, I turn to the issue of its ubiquity in discourse. Finally, I discuss behavioral research verifying the psychological reality of multi-word, formulaic sequences.

2.2.2.1 What are Formulaic Language and Multi-Word Expressions?

Consider the following word sequences: *you know what I mean, less than or equal to, beyond a reasonable doubt, all of a sudden, oh my God, as soon as, sort of, mutual fund*. These sequences were all identified by MERGE as multi-word expressions (hereafter MWEs) in experiments that I will report on in chapter 4, and they were assigned relatively high scores by human raters according to subjective criteria given to them for identifying MWEs. The reader may already have a sense that there is a unifying quality of formulaicity across these sequences; here, I will address the issues involved in defining forms of formulaic language such as MWEs.

MWEs, which will be the term I use in my research, is one among many related terms that have been employed in various studies, all of which may be placed under the umbrella of formulaic language. Wray, one of the foremost researchers on the topic, provides an extensive list of terminology employed by different authors in the literature on formulaic language (2002):

> amalgams – automatic – chunks – clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[dioms] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units (p. 9)

In fact, despite its size, Wray's list is not exhaustive; other influential terms that might be placed under the general rubric of formulaic language include *lexical bundles* (Biber et al. 2004), *collostructions* (Stefanowitsch and Gries 2003), *constructions* (Goldberg 2006, p.3), *n-grams* (Gries 2008), *phraseologisms* (Gries 2008), and, of course, *MWEs*. Part of the reason there is such a proliferation of terms is the fact that, across authors and research traditions, these terms are not all used to refer to exactly the same linguistic phenomenon.

In reviewing research on formulaic language, Gries (2008) identifies six variables involved in how the terms under this umbrella are typically defined. (Note that, while Gries uses the terms *phraseology* and *phraseologism*, the variables he identifies are applicable across the terminology space of formulaic language.) Gries argues that, while these six variables may be explicitly addressed in some authors' definitions of their particular terminological choices vis-à-vis formulaic language, very often one or more of these variables are left untouched, and the reader is left to infer them. Below, I detail Gries' six

variables. I also define my use of the term MWE in the context of MERGE with reference to these six variables.

1. The type(s) of linguistic units involved in the phraseologism. Typically, research on formulaic language has focused on co-occurrences of words, which is how I define MWEs in the present work. Moreover, words, as defined here, are orthographically transcribed and are word forms rather than lemmas. As a point of contrast, Gries' definition of phraseologism includes the possibility of co-occurrences between one or more words and abstract grammatical constructions. For example, under this definition, a co-occurrence between the verb *give* and the ditransitive construction would form the phraseologism *SOMEONE give SOMEONE SOMETHING*.

2. How many of these units are involved. Often, corpus linguists examining collocations using automated extraction techniques focus on bigrams (co-occurrences between two elements), since they are easy to extract and handle statistically (Evert 2005). From the perspective of linguistic theory, this is not ideal, since formulaic language may in principle be of any length. Accordingly, MWEs in this work may consist of two or more words, with no ceiling on their length other than the length of the utterances in which they are contained.

3. How frequent the co-occurrence of units must be in order for it to count as a phraseologism. Frequency is an extremely important correlate of formulaicity. Broadly speaking, the more frequent a sequence is, the more entrenched in memory it becomes. However, the story is not so simple. While the correlation of frequency and storage may be

obvious for certain oft-used sequences like *you know what I mean*, idioms are typically very low frequency, yet clearly memorized. Thus, some researchers eschew raw frequency counts of sequences in favor of more sophisticated statistical measures of contingency. These include measures such as transitional probability (mentioned in the discussion of distributional learning experiments), in which the probability of some item is conditioned on the occurrence of some other item. In addition, there are numerous more sophisticated lexical association measures that have been developed by corpus linguists to quantify the statistical strength of word co-occurrences. This issue of how to count word sequences is an important one, which I return to later in this work. However, for now it suffices to say that contingency-based metrics represent how salient or strong a sequence is in a corpus based not just on the frequency of the sequence, but also on other types of frequency information, such as the corpus size, or the individual frequencies of the component words of the target sequence outside of that sequence. Many of these contingency-based measures are designed to treat not just frequent sequences like *you know what I mean* as strong MWEs, but also low frequency (yet salient) ones like idioms. In the discussion below, I use the term *co-occurrence strength* to refer collectively to sequence frequency as well as to contingency-based measures. In the present study, to be identified as a MWE, a word co-occurrence must exceed a threshold according to a specified lexical association measures. In the behavioral and computational studies I review, some authors use simple frequency counts of word sequences as indices of their formulaicity, while other studies use measures of contingency. And, in some studies, frequency-based criteria for formulaic language may play no role; often, researchers may determine formulaic language based purely on semantic or morphosyntactic criteria, or intuitive notions of what is formulaic.

4. How much syntagmatic distance, if any, there can be between the units. In much of the corpus work dealing with automated extraction techniques, the bigrams that are focused on comprise strictly adjacent elements. Again, this is not ideal; formulaic language often contains discontinuous dependencies between words. Under my definition, therefore, MWEs may (or may not) exhibit one or more internal gaps, with the maximum permissible gap size programmed in the MERGE algorithm as a free parameter. Thus, a nonadjacent collocation is in principle permissible—for example, *as _ as* in the sentence *he's as tall as me*—just as a typical adjacent one is permissible, such as *happy birthday*.

5. How much lexical and syntactic flexibility is permitted among the units. Some degree of lexico-syntactic inflexibility is one of the most typical diagnostics for formulaic language (e.g., Pawley and Sider 1983; Erman and Warren 2000). More specifically, a sequence of units is typically seen as formulaic if it is hearable as more conventional and native-like than some other alternative way of 'saying the same thing.' For example, in English we say *strong coffee* but not *powerful coffee*, and *I want to marry you* but not *I want marriage with you*. Pawley and Sider have termed this phenomenon native-like selection, and historically it represented one of the first major challenges to the traditional generative criterion of semantic non-compositionality for sequence storage. That is, while both *strong coffee* and *powerful coffee* are compositional, only one is habitually used. However, native-like selection has limited uses in cases where it is difficult to identify an alternative way of saying the same thing, with paraphrases often being much longer or only partly synonymous with the target sequence. As a result, native-like selection was unsuccessful in completely overturning the generative notion of what is and is not lexicalized.

Additionally, in some approaches to formulaic language (such as Gries' definition of phraseology), the unit under inspection may include partially-specified paradigmatic slots into which various words may be placed (e.g., *he kicked the bucket* vs. *she kicked the bucket*). And some approaches allow for syntagmatic interchangeability of elements within a given unit, such as the concgrams approach of Cheng et al. (2009), in which *hard work* and *work hard* both instantiate the same concgram. Due to the ways in which MERGE tracks word co-occurrences and assigns scores to them, it does not allow for such kinds of syntagmatic interchangeability in the MWEs that it finds. And, while MERGE does not directly learn paradigmatic representations for particular slots within MWEs, the gaps of different sizes that MERGE can learn (e.g., *put _ down*) may be thought of as functioning as slots for different hypothetical paradigm members.

6. Whether or not the units together must exhibit semantic unity and non-compositionality. This is one of the variables according to which approaches to formulaic language vary the most. For some researchers, non-compositionality is a prerequisite for formulaicity. For others, whether or not a word sequence is compositional is a basis for categorizing formulaic language into different types (e.g., idiomatic versus non-idiomatic formulaic language; see Conklin and Schmitt 2012). And still in other approaches, there may be no direct accounting for semantics at all; instead, frequency-based metrics may be the sole means for identifying formulaic language. This last approach is the one taken by the MERGE algorithm. (In Chapter 4, I will report a rating experiment that may indirectly access human intuition about other sorts of criteria regarding formulaicity, but this is used as a post-hoc evaluation of model performance, which itself is based solely on frequency

criteria.) One should note that semantically-based criteria preclude the possibility of formulaic language that is entirely composed of function words, since function words have little to no semantic content. Thus, a sequence such as *in the* or *the _ of*, are neither compositional nor non-compositional, even though such sequences are very high frequency and thus will typically be represented as formulaic by frequency-based approaches.

Regardless of the variation in how terms are defined according to these different variables, undergirding all of them is a consistent notion of non-creative linguistic processes; that is, formulaic language is accessed from memory rather than creatively assembled online. And, in section 2.2.2.3, I will discuss various studies that have provided empirical evidence for this perspective on the mental storage of formulaic language.

One of the advantages of the wide definitional latitude afforded by Gries' taxonomy is that it provides a principled means by which we can relate the operationalizations of formulaic language used by different authors. Along these lines, it allows for the placement of my definition and those of others within the broader framework of construction grammar, one of the most influential theoretical approaches in contemporary linguistics (see Goldberg 2006 for a thorough introduction). In this approach, constructions represent *form-meaning pairings* in language, and manifest at all levels of linguistic structure and abstraction. Goldberg (2006, p. 5) provides a widely cited typology of constructional types, reproduced in Table 2.1.

Table 2.1. Constructional Types

Morpheme	e.g., *pre-*, *-ing*
Word	e.g., *avocado, anaconda, and*
Complex word	e.g., *daredevil, shoo-in*
Complex word (partially filled)	e.g., [N-s] (for regular plurals)
Idiom (filled)	e.g., *going great guns, give the Devil his due*
Idiom (partially filled)	e.g., *jog* <someone's> *memory, send* <someone> *to the cleaners*
Covariational Conditional	The Xer the Yer (e.g., *the more you think about it, the less you understand*)
Ditransitive (double object)	Subj V Obj1 Obj2 (e.g., *he gave her a fish taco; he baked her a muffin*)
Passive	Subj aux VPpp (PP_{by}) (e.g., *the armadillo was hit by a car*)

In theory, Gries' phraseological taxonomy could be used to account for nearly all of these constructional types (with the exception, perhaps, of morphemes, since they are not typically regarded as representing a co-occurrence of meaningful units). That is, given that constructions are memorized units that are not created online, they instantiate formulaic language. In practice, however, most of the work that is couched in the terminology of phraseology, formulaic language, MWEs, and many of the terms mentioned above, uses these terms to refer to the constructional types that roughly fill in the middle of the table: filled and partially filled complex words, filled and partially filled idioms, and the covariational conditional—in brief, partially to fully lexicalized MWEs.

Such constructions contrast with fully abstract argument structure constructions such as the ditransitive (not typically considered a phraseologism or MWE), which represent perhaps the greatest focus of construction grammar-based research. Thus, one coarse way of understanding the difference between construction grammar versus common phraseological/formulaic language/MWE-based research is in terms of the difference in

these constructional types typically researched (despite actual theoretical overlap in the approaches).

At the same time, however, MWEs as implemented here have more varied compositions than the (partially) lexically specified multi-word constructions in Table 2.1. This is because, again, they are defined by the extraction and ranking computational technique on purely frequency-based criteria. In this way, they are similar to lexical bundles as discussed by Biber et al. (2004). Lexical bundles and MWEs, in addition to being possibly semantically compositional and containing only function words, may not represent complete constituents. For instance, some examples of such lexical bundles identified in Biber et al.'s (2004) work include I *don't know if, I want you to*, and *if you look at*. However, while lexical bundles are defined in terms of raw frequency counts, MERGE uses (as I have stated above) more sophisticated lexical association measures (see below).

2.2.2.2 The Extent of Formulaic Language in Discourse

The proliferation of terms under the rubric of formulaic language not only bespeaks the wide amount of definitional variability observed across authors, but it also hints at the omnipresence of formulaicity in language. In a study examining the extent of formulaic language in discourse, Foster (2001, p. 83) asks seven native English speakers to indicate in a set of corpus transcripts those word sequences they consider to have been produced as fixed chunks. The author then consolidates the annotations, generating a master transcript containing chunks annotations only for those sequences that had been indicated as fixed by at least five of the participants. Consider the following example, adapted from the author's paper:

> (It doesn't matter) (what the circumstances), (she didn't have the right to) (take his life). If she was that er emotionally (you know) er distressed, then she should have- (I don't know) (got out of the situation). (It's difficult to say) when you are not (in the situation) but (at the end of the day) she did (take another human life). (There you go). (p. 83)

In this example, parentheses delimit the boundaries of the fixed chunks agreed upon by at least five participants. What is immediately apparent is how much of the text appears to be formulaic. Overall, according to her method, Foster finds that about 32% of the corpus is formulaic.

Other studies have also aimed to quantify the degree of formulaicity in language, using different criteria for defining formulaic language, different corpora, and yielding different results. Erman and Warren (2000), using the term *prefab*, find that 52% of the written corpus and 59% of the spoken corpus contain such prefabs. Biber et al. (1999), using the term *lexical bundles* (see above), find that roughly 30% of a conversational corpus is formulaic. Rayson (2008) finds that 15% of a corpus is formulaic. Butler (1997), using Spanish data, finds that 12.5% of his spoken corpus is formulaic. At the low end of the spectrum, Moon (1998, p. 57) finds between 4 and 5% of a corpus of >18 million words to be part of what she terms *fixed expressions and idioms*. It is noteworthy that this author's criterion for such fixed expressions including idioms is quite conservative: she checks for occurrences of expressions established as formulaic by the *Collins Cobuild English Language Dictionary*. In contrast, at the high end of the spectrum, Altenberg (1998) finds that 80% of the words in the London-Lund corpus are part of recurrent words combinations. This author's definition of formulaic language is quite liberal, including any continuous sequence occurring more than once.

There is neither the space nor the need to delve into all the differences among these approaches in order to understand why their results vary in the way that they do. In general, however, the studies on the low end tend to adhere to more established idiomatic kinds of formulaicity, in line with traditional generative notions of the kinds of word sequences that are stored in the lexicon. In contrast, the studies on the high end tend to be less beholden to strict ideas about compositionality as definitive of formulaicity, often relying primarily on frequency-based criteria. Thus, under the less conservative approach espoused here, the important point is that much of discourse appears to be formulaic.

And ultimately, because of the apparent omnipresence of formulaic language in discourse, its application has become very important across a number of different research veins. These include syntactic parsing and distributional learning (Bod 2009); the use of formulaic language as the basis for the differentiation between varieties of the same language (Gries and Mukherjee 2010) and between genres within a single language (Biber et al. 2004); creating multi-word dictionary entries in lexicographic work (Sinclair 1987; Schone and Jurafsky 2001); development of native-like abilities in second language acquisition (e.g., Sinclair 1987; Simpson-Vlach and Ellis 2010); creating native-like speech in natural language generation (Lareau et al. 2011); and child language acquisition studies (Bannard et al. 2009).

2.2.2.3 Cognition and Formulaic Language

A criticism of the above-described approaches to quantifying the amount of formulaic language in discourse is that what counts as a formulaic sequence is often defined either in terms of researchers' intuitions, or by reference to dictionaries of formulaic

language, whose own criteria for deciding what is formulaic are not necessarily scientific or even known. Thus, it is desirable to have independent evidence that particular sequences are indeed formulaic in the sense of being psychologically represented as stored, holistic units in memory. Furthermore, in order to offer convincing evidence against the traditional generative syntax/lexicon divide, it must be apparent that the sequences being shown to be stored include not just non-compositional, idiomatic chunks, but also sequences that are more or less compositional—that is, sequences traditionally theorized to be constructed via rules.

In this subsection, I review various studies that provide evidence for this perspective. First, I focus on a number of different methodological approaches that examine the cognitive architecture of formulaic language in adults. I then conclude the subsection by considering a few studies that investigate the role of formulaic language in children's acquisition.

2.2.2.3.1 Adult Processing of Formulaic Language

One interesting source of evidence comes from research on patients who have suffered neurological trauma. In one study, Mondini et al. (2002) generate stimuli that include Italian noun-adjective and adjective-noun compounds, which are test items, and novel noun-adjective and adjective-noun combinations, which are controls. An example of one of the compounds is *febbre gialla* ("yellow fever"). Such compounds are lexicalized, and cannot be modified by additional NP dependents, nor can their words be reordered. In contrast, an example of a novel sequence is *febbre strana* ("strange fever"). Novel sequences can be modified by additional NP dependents, and their component words may be reordered.

Crucially, Italian grammar requires gender and number agreement morphology in both compound and novel noun-adjective/adjective-noun sequences.

The authors then have two non-fluent aphasics perform certain psycholinguistic tasks, including a reading task and a repetition task (see below for discussion of these kinds of tasks), as well as a completion task (where participants read aloud and complete a sentence that contains a blank where the agreement-marking morpheme should be). Overall, participants can successfully mark agreement for compounds, but not for novel combinations, suggesting that, while the former items are stored, the latter are constructed online.

In another neurological-deficit study, Van Lancker Sidtis and Postman (2006) analyze the spontaneous speech of left versus right hemisphere-damaged patients. While the left hemisphere-damaged group use more formulaic sequences than a control group, the right hemisphere-damaged group use fewer formulaic sequences than controls. Accordingly, the authors note that the right hemisphere has been previously tied to formulaic language, and the results of their analysis provide further evidence for this.

While these studies are quite telling about the neurological representation of formulaic language, again, the authors decide in an a priori fashion what counts as formulaic. The primary source of independent evidence for formulaic language is the way in which particular behavioral phenomena that index holistic storage may be correlated with frequency-based measures of formulaic sequences. More specifically, to the extent that such a correlation holds at the level of the whole sequence itself—rather than at the level of the individual words—this is evidence that a representation that spans the entire sequence must be maintained in memory across usage events. Furthermore, because sequencehood is

operationalized in terms of frequency rather than in terms of a simple yes-no decision made by the researcher as to whether a sequence is a formula or not, positive results from such studies more strongly provide evidence against generative notions of the lexicon. This is especially true in light of the possibility that a researcher making yes-no decisions may be biased towards a compositionality-based definition of formulaicity, or towards a controversial native-like selection-based definition.

Finally, the studies to be discussed employ a variety of frequency-based measures, from raw frequency to contingency-based measures such as transitional probability and lexical association measures. Thus, to the extent that these different operationalizations yield similar results, the general findings across studies will be strengthened.

One behavioral indicator of an increasingly holistic, stored representation of a particular sequence is the degree of reduction and coalescence at the levels of phonetics, phonology, and morphology. Generally speaking, the higher the frequency of a sequence, the greater its reduction and coalescence (that is, the greater its status as an integrated unit rather than a sequence of fully separable subcomponents). A number of studies have shown this to be true using data from English corpora. These include: Krug (1998), who examines the alternation between *have* and *'ve* and between *is*/*has* and *'s*; Bybee and Scheibman (1999), who examine lenition of the consonantal onset and the vowel in *don't*; and Gregory et al. (1999), who examine durational shortening as well as tapping and deletion of final alveolar stops among word types containing such segments.

All of these studies investigate frequency-based measures of co-occurrence of these target structures with different adjacent surrounding word types (forming bigrams or trigrams) as predictive of processes of reduction. And, while the manners in which they

quantify co-occurrences vary—Krug (1998) and Bybee and Scheibman (1999) use simple co-occurrence frequencies, while Gregory and colleagues (1999) compare various measures including frequency, conditional probability, and mutual information—all of these studies draw similar conclusions. That is, reduction processes increase as the strength of the co-occurrences between the target words and surrounding words increases.

Taking a somewhat different approach in a previous study (Wahl 2015), I extract all adjacent word bigrams from a spoken corpus and investigate the degree to which the co-occurrence strengths of these bigram types (measured in terms of different lexical association measures) are negatively correlated with the placement of boundaries between intonation units. Put more simply, as the co-occurrence of two words becomes more frequent, the likelihood that a break between successive intonation contours will be placed between those two words decreases. Taken together, these studies suggest that statistically stronger word sequences function as integrated, non-divisible, stored units, and this is reflected in their production.

At the same time, a number of criticisms may be leveled against the statistical approaches employed in these studies of reduction and coalescence. Most worrisome is that the Bybee and Scheibman (1999) study engages in no significance testing of any kind. In addition, the use of a multifactorial model by Gregory et al. (1999) to examine the most predictive type of frequency measure for processes of reduction fails to take into account the obvious problem of collinearity. And overall—with the exception of Krug (1998)—the studies do not take into account the possibility that the indicators of reduction/coalescence that are observed are actually better predicted by the frequencies of the individual words or smaller sequences that function as components of the full sequence. For example, in the case

of Bybee and Scheibman (1999) it could be that the frequencies of the individual words that precede and follow *don't* are responsible for the reduction of *don't*, rather than the co-occurrence frequency of these words with *don't* being responsible. And, even in the case of Krug (1998), who does report some information on individual component word frequencies, and its apparent lack of effect on reduction, he does not engage in significance testing to validate this claim. Thus, while the results of these studies are suggestive of the holistic representation and storage of multi-word, formulaic sequences, their statistical shortcomings leave open the possibility that some other factors, like the frequencies of sequence components, are in fact what are responsible for the observed effects.

Complementing this research on reduction and coalescence is a large body of experimental psycholinguistic work that has examined particular effects of expedited processing as a correlate of frequency-based measures of word co-occurrence strength. And as might be expected, the general finding in this research is that the stronger a word co-occurrence, the stronger the memory representation, and the more quickly the word sequence is processed, providing further evidence for the holistic storage of recurrent sequences.

Methodologically, this body of research has drawn on a number of different kinds of psycholinguistic experimental designs. One such design is the word monitoring task. In a study by Sosa and MacFarlane (2002), participants are asked to listen to sentences and press a computer key as soon as they hear the word *of*. In these sentences, *of* occurs as a part of different adjacent *X*-plus-*of* collocations (e.g., *kind of, sort of*), from four different frequency bands. As the authors hypothesize, slower reaction times and higher error rates are observed for high frequency collocations. Thus, it appears from these results that higher frequency

collocations function more as holistic units that are more difficult to 'look inside.' Crucially, in contrast to the assumptions one might make given the reduction and coalescence studies discussed above, the slower reaction times could not be due to reduced acoustic salience as a result of frequency-correlated phonetic reduction: only a handful of test items were reduced, and these were low frequency collocations.

A more popular experimental design in the literature on formulaic language is different kinds of reading tasks. Under these approaches, experimental participants read sentences or word sequences presented on a computer screen, and the speed with which they read certain sub-sequences within the larger stretch is tracked. Two common implementations of this approach include (A) self-paced reading tasks, in which participants press a button on a keyboard to advance through successive sub-sequences in a larger sequence or sentence (Reali and Christensen 2007; Tremblay et al. 2011), and (B) eye tracking-based reading tasks, in which a special eye-tracking device is used to record the fixations of participants' eyes (and their durations) on particular words while they read a sentence or sequence (McDonald and Shillcock 2003; Siyanova-Chanturia et al. 2011).

The different studies within this body of research use these approaches to examine different kinds of multi-word sequences, including subject pronoun-plus-verb sequences within English object relative clauses (Reali and Christiansen 2007); English verb-plus-noun sequences (McDonald and Shillcock 2003); four-word lexical bundles, which are any four-word sequence exceeding some frequency threshold in a corpus (Tremblay et al. 2011); and conventionalized binomial phrases (e.g., *bride and groom*; Siyanova-Chanturia et al. 2011). Ultimately, this research all shows that participants spend less time reading more strongly co-occurring subsequences than less strongly co-occurring subsequences, suggesting a more

holistic representation in memory that does not require as much online computation. Here, frequency is the most common implementation of subsequence strength, although McDonald and Shillcock (2003) find that the statistical model most predictive of their observed data is one that includes both frequency and the contingency-based transitional probability measure.

Tremblay et al.'s study (2011) also includes another type of experiment that is known as a recall task. While the word monitoring and reading tasks tap into online comprehension, recall tasks are designed to investigate memory processes. In this task, participants are presented a sentence both aurally and visually, followed by a list of distractor words. The sentences contain either frequent lexical bundles or infrequent non-bundles (just as in the other experiment in Tremblay et al. 2011). Following presentation, participants have to type the sentences and as many distractor words as they can remember. As expected, sentences containing lexical bundles are recalled better, likely because it is easier to retain a single item in working memory (i.e., holistically stored lexical bundles) than a collection of items.

But perhaps the most widely used experimental task in psycholinguistics for probing processing is the lexical decision task (and its various derivatives), and versions of it have been used specifically in the investigation of formulaic language. In this kind of task, participants are presented with one or more stimuli, and then they must make a decision as to the linguistic validity of one or more of the stimuli. The dependent variable in such a task is the time it takes for participants to respond. Thus, this task measures processing effort associated with particular stimuli.

In a pair of studies, participants are presented with a random sampling of corpus-extracted 3-, 4-, and 5-word (adjacent) sequences of at least a certain frequency (Ellis et al. 2008)[2], and verb-plus-noun or adjective-plus-noun sequences that had previously been identified as reflecting principles of native-like selection (e.g., *absolutely diabolical* and *fully fledged*; compare to *fully diabolical* and *absolutely fledged*; Ellis et al. 2009).

Participants then have to make decisions as to the possibility/likelihood of these sequences in English. As hypothesized, decreasing reaction time is correlated with increasing co-occurrence strength in both studies (frequency and mutual information are used as measures of co-occurrence strength in Ellis et al. 2008; log frequency is used in Ellis et al. 2009). This provides evidence, once again, for holistic storage of frequent sequences, since it should take less time to access a single item from memory than individual lexical items that then need to be assembled into larger sequences via syntactic rules.

Durrant and Doherty (2010) also use a lexical decision task in what is the only study I am aware of that provides evidence *against* the primacy of frequency-based measures in determining holistic representation and processing of multi-word sequences. Using the Edinburgh Association Thesaurus and a norming study, the authors first determine whether corpus-retrieved adjacent bigrams are psychological associates or not. Psychological associates are words with related meanings that semantically prime one another. Participants are exposed to the initial collocate of the bigram as a prime, and then the second collocate as the target for the lexical decision. They find that when prime exposure is 600 milliseconds, participants react to high mutual information bigrams more quickly than to low mutual

[2] Ellis et al. (2008) also includes a separate production task, wherein participants have to utter sequences presented to them as quickly as possible. This experiment yields comparable results to the lexical decision task.

information ones, and this is true for bigrams in which the elements are or are not psychological associates. However, when exposure is only 60 milliseconds (which would make top-down anticipatory strategizing nearly impossible), only high mutual information bigrams that are associates show significant differences from low mutual information bigrams. The authors' results thus suggest that the facilitation effects elsewhere attributed to frequency-derived representational unity of word sequences might in fact be due to conceptual relatedness of collocates' meanings. At the same time, in the 60-millisecond condition, the results nonetheless approach significance for non-associate bigrams, indicating that the results may not be as definitive as they initially seem.

Compared to the reduction and coalescence studies discussed earlier, these various experimental studies overall offer a higher level of statistical sophistication in the degree to which potential factors responsible for observed effects are controlled for. Specifically, in theory it could be that some behavioral effects are better predicted not by a frequency-based measure of the entire target word sequence, but rather merely by an *n*-gram subsequence within the entire sequence. Or, the effects may be better predicted by the frequency of an individual word within the sequence. Thus, these sub-components need to be controlled for in some way.

For example, in Siyanova-Chanturia et al.'s design (2011), binomial phrase target items such as *bride and groom* are matched with their non-conventional, reverse positionings (*groom and bride*) as control items. Thus, faster reading times associated with the conventionalized sequences over the non-conventionalized ones cannot be a result of word frequency differences.

Similarly, in Tremblay et al.'s study (2011), the authors match frequent lexical bundles with infrequent control sequences, which differ from one another in terms of only one word. Moreover, this word is more frequent in the non-bundle condition. Again, in such a design, the expedited reading and enhanced recall observed for bundles could not be due to this different individual word, since, if the individual word were responsible, one would expect the control sequence to be read and recalled more quickly than the lexical bundle, since the control sequence has a higher frequency word in it as the only difference.

Nevertheless, in both of these cases, the authors do not frequency-match any bigrams or trigrams that contain the target word, so it is in theory possible that what is causing the expedited processing is not the entire lexical bundle, but merely these smaller *n*-grams. That is, in the case of Siyanova-Chanturia et al. (2011), it could be that the observed effects are caused by differences in the frequencies of the bigrams *groom and*, *and bride*, *bride and*, and/or *and groom*. In fact, in none of the experimental studies discussed thus far are frequency-based measures of sub-grams controlled for.

Furthermore, in some of the studies, such as Sosa and MacFarlane (2002), Ellis et al. (2008), Ellis et al. (2009), and Durrant and Doherty (2010), frequency-based measures of neither individual words nor sub-grams are used. In the case of the Durrant and Doherty (2010) study, this casts doubt on the validity of their argument that the expedited processing of formulaic sequences is based primarily not on frequency-based measures of such sequences, but rather on the semantic associations of their component words.

In my view, there are two studies in the experimental literature (that I am aware of) that offer the highest level of statistical controls, as they control for both frequencies of

individual words and sub-grams that are components of the larger target formulaic sequences. These studies are Tremblay and Baayen (2010) and Arnon and Snider (2010).

In the former study, the authors present participants with a list of 4-word sequences (e.g., *in the middle of*), and then participants have to type as many of the sequences as they can remember. The authors find that sequences that exhibit a higher ratio of their overall frequency to the frequency of the sub-sequence forming their first three words (i.e., those whole sequences that are more 'salient' relative to their sub-sequences) are recalled more accurately. This finding points to the storage of these whole sequences.

Furthermore, the authors include another empirical methodology in their study, that of event-related potentials (ERPs). ERP-based studies involve placing electrodes on the scalp of participants and recording spikes in brain electrical activity time-aligned with different points in comprehension and production events. Particular types of spikes in activity are associated with particular cognitive processes. The authors find that early ERP spikes, or components, were modulated by the frequency ratio (described above) of the word sequences to sub-sequences. They argue that this again points to the holistic storage of such sequences, since it would be too early after presentation of stimuli for there to be retrieval of multiple words.

And crucially, in the analyses of both types of data (behavioral and ERP), Tremblay and Baayen (2010) use a multifactorial model that takes into account numerous predictors, including frequency information on sub-grams and individual words. The fact that whole-sequence strength emerges as significant indicates that there indeed must be a representation across the whole sequence that is maintained across usage events.

In the other study, Arnon and Snider (2010) use a variant of a lexical decision task, in which participants are asked to decide whether different four-word phrases are possible in English. These phrases comprise high and low frequency items, with different thresholds for high frequency across different trials. Similar to Tremblay et al. (2011), each high frequency test item is matched with a low frequency item that differs in terms of only one word, and these words are matched for frequency. But unlike Tremblay et al. (2011), bigrams and trigrams that include these differing words are also frequency-matched. Thus, when the results eventually show that participants respond to high frequency items more quickly than to low frequency items, there is strong evidence that the effect is a result of processing facilitation of retrieving from memory a whole, stored representation that spans the entire sequence, rather than smaller word-level, bigram-level, or trigram-level representations. What is more, as the threshold for what counts as high frequency is adjusted, the effect is modulated, which shows that holistic storage and processing of formulaic sequences is a gradient phenomenon.

Overall, the discourse-analytic, neurocognitive, corpus-based and experimental studies discussed provide convergent evidence that multi-word sequences are stored as holistic entities in memory as a function of their co-occurrence strength. Importantly, these sequences do not necessarily have to be semantically or syntactically non-compositional, which, under traditional generative models, was the criterion for whether something was assumed to be stored in the lexicon. The storage of even compositional items is particularly supported by studies like Tremblay and Baayen (2010) and Arnon and Snider (2010), which exhibit tight statistical design and controls and use stimuli that are fairly compositional.

Other studies vary in the role that sequence meaning plays. For example, while Durrant and Doherty's (2010) stimuli are not necessarily non-compositional, the authors do vary whether the words in the test bigrams are psychological associates or not. Moreover, Siyanova-Chanturia et al.'s (2011) use of binomial phrases (e.g., *bride and groom*) and Ellis et al.'s (2009) stimuli (e.g., *fully fledged*), while also not exactly non-compositional, are selected to embody the phenomenon of native-like selection. In most of the other frequency-based studies, the issue of meaning is not taken into account (e.g., Tremblay et al.'s (2011) use of purely frequency-defined lexical bundles). In future work, it would be desirable to more explicitly control for the dimension of meaning. In particular, although Arnon and Snider as well as Tremblay and Baayen's studies are generally non-compositional, a separate factor based on compositionality scores assigned by human raters, for example, may yield interesting results. It would also fortify the empirical basis of the claim that even non-compositional items are stored. For now, however, the results of these various studies do indeed appear to point in this direction.

Both the experimental studies and the corpus-based studies on reduction and coalescence also vary in how they operationalize co-occurrence strength. While most studies use frequency counts, some studies find significant results using probability measures (McDonald and Shillcock 2003; Tremblay and Baayen 2010) and the lexical association measure Mutual Information (Gregory et al. 1999; Durrant and Doherty 2010). On the one hand, these measures treat certain kinds of sequences differently, such as low frequency idioms, which may be represented as strong by contingency-based measures. One criticism of Arnon and Snider (2010), then, might be their less sophisticated use of frequency counts. Nonetheless, frequency and contingency are fairly collinear, so the overall results would

likely not differ dramatically (though this is an empirical question). Furthermore, the very sequences that get ignored by frequency counts—low frequency idioms—are incontrovertibly formulaic. Thus, in these studies, low frequency items that might exhibit some degree of formulaicity would probably work against the observed results, since idioms exhibit a processing advantage (Tabossi et al. 2009). In other words, the processing advantage of low frequency idioms (included among other low-frequency, non-idiomatic items) could decrease processing latencies of a low-frequency condition overall, potentially narrowing the significant difference in processing latencies of low frequency versus high frequency conditions. In contrast, without such idioms, there might be a wider difference between low and high frequency conditions in terms of processing latencies. In summary, it is unlikely that the hypothetical changes to operationalizations of variables that I have described here would have changed the overall convergent results of these various studies.

2.2.2.3.1 Formulaic Language in Children's Acquisition

In addition to the research on adult processing of formulaic language, there has been an important body of research on the role of formulaic language in children's acquisition. For an excellent review, see Bannard and Lieven (2012). Collectively, this research has shown that children's early multi-word utterances are dominated by rote-learned formulaic language. It is from the building up of and abstraction across these formulaic sequences that children gradually acquire a mature, productive linguistic system. Such a model is incommensurate with a sharp division between a lexicon that serves as a storage repository of unanalyzable/non-compositional items, and a syntax comprising fully productive rules. Moreover, because we know from the research discussed above that adults maintain stored

representations of formulaic language, we can conclude that children's representations of formulaic language do not simply disappear once they acquire a mature linguistics system; rather, they are maintained across the lifespan.

There have been a number of studies that investigate formulaic language in children's acquisition using behavioral data. For example, in a corpus-based study, Lieven et al. (2009) use a technique known as the traceback method. Under this method, investigators see how well structures found in child utterances can be accounted for by earlier usages. Lieven and her colleagues employ this approach using data from four 2-year-old children, and they find that between 20% and 40% of child utterances have previously been used by the children. Moreover, between 40% and 50% of child utterances differ from previous utterances only by a single alteration. This alteration consists of an insertion of new material into an existing frame within a structure, and, typically, this new material refers to people, places, or objects—concrete referent-concept pairings that are relatively simple for children to acquire. And, in their ability to interchange semantically related words in simple syntactic slots, children are demonstrating an emergent understanding of lexical categories. Thus, while children appear to be beginning with fixed word sequences, the authors provide evidence of how these sequences later give rise to a more productive system.

In another corpus-based study, Kirjavainen et al. (2009) examine *me*-for-*I* production errors in child speech, in which children use the accusative first person pronoun instead of the nominative one in subject position. Using data from 17 children aged 1;4 to 5;0, they find that the children's error rates correlate with the rate of *me* + verb combinations in caregiver input (that is, in constructions containing a nonfinite verbal complement such as *let me do that*). In addition, they look at the set of verb lemmas that are preceded by both *me*

and *I* in the children's speech. They find that, overall, lemmas preceded by *me* in the children's speech are also preceded by *me* in caregiver input at a rate higher than lemmas preceded by *I* in the children's speech are preceded by *me* in caregiver input. In summary, these results suggest that children are storing pronoun + verb combinations as chunks, rather than creatively generating these combinations via syntactic rules.

In contrast to these corpus-based studies, Bannard and Matthew (2008; 2010) use an experimental approach to examine the role of formulaic language in children's acquisition. In their design, 2- and 3-year-old children are asked to repeat either a frequent sequence (such as *a drink of milk*) or a paired less frequent sequence that is identical to the frequent sequence except for the last word (such as *a drink of tea*). The children make fewer errors and produce more quickly the initial part of the sequence in the frequent condition, despite the fact that this part is identical in the infrequent condition. And, as in Arnon and Snider (2010), both the final words and the final bigrams are matched for frequency, so differences in the final parts of the sequences cannot independently be responsible for the expedited processing. Rather, it must be that there is a representation spanning the entire sequence—a stored, formulaic sequence—that is responsible for the expedited processing.

2.3 Bringing it All Together

While the relevance of frequency as a correlate of storage is strongly affirmed in the preceding discussion of adult representations of formulaic language, I have not delved into why this might be the case. Researchers have used the term *chunking* to describe the cognitive process by which a syntagmatic sequence of linguistic percepts (such as words) comes to be increasingly represented as a holistic, integrated unit (such as a formulaic

sequence) the more times that the particular sequence is comprehended, processed, and produced (Bybee 2010). This chunking mechanism is seen by many to be a basic cognitive capacity of humans, and it is observable through the effects of coalescence, reduction, and expedited processing discussed above (Haiman 1994; Ellis 1996; Bybee 2010).

Functioning as a complement to chunking is analogy, which is seen as another basic cognitive capacity. Analogy is the process by which items which occur in similar contexts are taken to be members of the same abstract category (Bybee 2010).

Crucially, taken together, chunking and analogy may be considered to be the two fundamental components of distributional learning. Indeed, while the distributional learning studies reviewed do not use these terms, the processes investigated in these studies are basically chunking and analogy (even if chunking-associated effects of reduction, coalescence, and expedited processing are not themselves the subject of investigation). On the one hand, the early artificial language studies by Saffran, Aslin, and Newport (1996a; 1996b; 1998), and many of the studies that followed, rely on the participants' chunking of syllables to form emergent words, as a function of inter-syllable transitional probabilities. On the other hand, later artificial language studies that investigate the acquisition of word classes (e.g., Reeder et al. 2013) rely on participants' use of analogy of surrounding contexts to induce abstract word class categories.

Meanwhile, the correlation between frequency and holistic status (indexed by reduction, coalescence, and expedited processing) that is characteristic of chunking is likewise reflected in the adult and child language studies on the representation and processing of formulaic language. This raises the question: why do distributional learning and formulaic language represent fairly distinct research traditions, if chunking is at the

center of both of them? On the one hand, distributional learning and formulaic language primarily respond to two different areas of generative theory: universal grammar and its relationship to the poverty of the stimulus, and the syntax/lexicon divide, respectively. On the other hand, there seems to be a contradiction between the findings of distributional learning and those of formulaic language research when it comes to acquisition. As we saw in the research on children's acquisition of formulaic language, children seem to learn whole formulaic sequences of language at once: that is, they do not appear to start out by synthesizing formulas from individual words. However, a synthetic approach is at the heart of most studies of distributional learning.

Nonetheless, I argue that the seemingly contradictory findings are not actually so. In particular, distributional learning work using natural language data such as that by Pelucchi et al. (2009a; b), reviewed above, shows that children are using smaller units (syllables) to find larger units (words) that are embedded within word sequences. In fact, children use a plethora of strategies to segment words before they even begin speaking and thus using these holistic chunks (see Swingley 2005). Therefore, the mere fact that children produce whole strings intact does not mean that they do not also have analytic knowledge of their component units.

In fact, computational modeling of the distributional learning of formulaic language has shown that even when they are synthesized from smaller units (words), the structures produced are whole formulaic sequences at an early stage of acquisition, and only later become more productive and analytic. In this section, I review the literature on the computational modeling of distributional learning of formulaic language.

Remember from chapter 1 that this research can be cognitive or non-cognitive in nature. First, I review cognitive computational models, which examine the distributional learning processes by which humans can learn formulaic language. Then, I turn to the discussion of non-cognitive models. If we know that language is best theorized as highly formulaic, rather than primarily constructed creatively from individual words via syntactic rules, then we need computational tools for doing corpus-based research that are capable of identifying and representing this high level of formulaicity.

2.3.1 Cognitive Models of Distributional Learning of Formulaic Language

Perhaps the most researched set of computational approaches to the distributional learning of formulaic language are termed fragment grammars (e.g., Johnson et al. 2007a; 2007b; Bod 2009; O'Donnell et al. 2011). These approaches are based on context-free phrase structure grammars, yet they also learn/represent lexicalized multi-word structures by storing entire trees and subtrees (rather than generating all phrase structure trees from individual words). Thus, they are sometimes referred to as lexicalized grammars. Here, I will focus on Bod's Unsupervised Data-Oriented Parsing (UDOP) approach (2009), since it has been most thoroughly applied to the distributional learning of formulaic language in natural language corpora. While I briefly discussed this model in chapter 1, I will review it in more detail here.

The UDOP algorithm begins by taking some group of corpus sentences. Then, for each sentence, all possible unlabeled binary-branching syntactic tree structures are generated. And, for each tree, all possible subtrees are extracted. Importantly, the subtrees may instantiate discontinuous word sequences. These trees and subtrees are then stored.

Consider the sentences *Watch the dog* and *The dog barks*. According to the principles just described, the former sentence can be parsed as *[[watch the] dog]* and *[watch [the dog]*, while the latter sentence can be parsed as *[[the dog] barks]* and *[the [dog barks]]*. Along with these four trees, the subtrees *[watch the]*, *[X dog]*, *[watch X]*, *[the dog]*, *[X barks]*, *[the X]*, and *[dog barks]* are stored with their frequencies.

Next, the total collection of trees and subtrees is used to find the best trees for a group of sentences. This group may be the same group from which the trees and subtrees were extracted, or it may be a new group. There are different ways to find the best tree, but the method used in Bod (2009) involves a combination of finding the shortest derivation for a sentence, and breaking ties when there are multiple shortest derivations. If there is a tie (i.e., one can generate two different trees for a sentence via equally short derivations), the tie can be broken by summing across the probabilities of *all* of the different possible derivations for each these two trees; the tree with the lower sum of the probabilities of its possible derivations is the winner. For the sentence *the dog barks*, the shortest derivations are the two full trees. (When a sentence has been previously seen during training, the shortest derivation will always be a full tree). In tie-breaking, the best tree would end up being *[[the dog] barks]* since the subtree *[the dog]* occurs twice in the collection of extracted trees and subtrees from the toy example. When the extracted trees and subtrees are used to find the best tree for a sentence that contains words not previously seen (as in the case of using parsing sentences distinct from the training sentences), the unfamiliar word can be treated as a wildcard and can match with any word.

The author evaluates the model across a number of case studies, using both adult and child-language corpora. In general, the model performs well in both cases in acquiring

structures and assigning parses that reflect hand-annotated gold standard trees. In the case of the adult corpus, the author trains the model on the whole corpus, and then tests it on this same set of utterances. In an additional step, he compares versions of UDOP that can and cannot learn discontinuous structures; the model that can learn discontinuous structures performs significantly better, highlighting the importance of discontinuous co-occurrences in the distributional learning of linguistic structure.

In the case of the child language corpus, the author implements different incremental versions of the model. In one, he trains the model on the utterances in all corpus files up to some file *k* (child, adult, or child and adult utterances), and then tests the model's ability to assign correct parses to the child utterances in file *k*. He repeats this procedure for each file *k* from the beginning to the end of the corpus. The author also uses the same basic procedure, but for all utterances *k* in one of the corpus files.

In these incremental versions, the author finds that, as learning progresses through the corpus, the structures acquired become more abstract and productive, which reflects the behavioral results discussed above regarding the role of formulaic language in children's acquisition. Furthermore, in working with the child data, the author finds that the model can learn complex fronted auxiliary questions based only on simple auxiliary questions. This is significant because auxiliary fronting is one of the English structures that generativists have long claimed distributional learning-based mechanisms cannot acquire due to the poverty of the stimulus. These child language results are all the more impressive because, while in the case of the adult corpus the model has access to part-of-speech tags, the model can only base its learning on the word forms themselves in the case of the child data.

Solan et al. (2005) develop a different approach that is not based on a phrase structure grammar but rather on a unique, induced graph-based grammar. Their procedure begins by loading a corpus into a directed pseudograph. (A pseudograph is a nonsimple graph that permits both loops and multiple edges between the same pairs of vertices). In the pseudograph, each vertex represents a lexical entry, and each sentence in the corpus defines a separate path over the graph (linking vertices). Furthermore, each path is indexed according to the order of appearance of the corresponding sentence in the corpus.

Next, grammar induction proceeds as the algorithm traverses various search paths. Initially, these search paths correspond to the original sentences. If two or more vertices (i.e., words) have several paths going through them in parallel, this may constitute a significant pattern—that is, a multi-word unit of some kind—and a set of mathematical equations is used to determine which successions of vertices represent such significant patterns. For a given search path, multiple significant patterns may be identified, but only the most significant one is chosen as the winner. The winning succession of vertices is then merged into a new single vertex (i.e., the multi-word sequence is unitized). Importantly, any paths (corpus sentences) that pass through only part of the succession of vertices corresponding to the significant pattern are not merged into the new vertex. Instead, along with their vertices, they are spun off and maintained as separate subpaths.

In addition, the algorithm identifies equivalence classes along the search path. These are regions where a number of paths pass through a vertex, then fan out to different vertices, then immediately fan back into a common vertex again. In other words, equivalence classes occur when there are paths that overlap in terms of two vertices, with one vertex intervening in which these paths do not overlap. Again, mathematical formulae are used to determine the

number of paths that need to exhibit this pattern for there to be an equivalence class. Interestingly, after one equivalence class is merged, this may create the common context for a new equivalence class, which may then create a succession of vertices that represents a now-significant pattern, which can be merged into a new single vertex. Thus, these forms of syntagmatic (significant patterns) and paradigmatic (equivalence classes) learning feed each other. Finally, it is important to note that the best available pattern in each iteration is immediately and irreversibly rewired, and thus the syntax the model acquires depends on the order of presentation of the sentences.

Ultimately, the authors evaluate the model using both artificial language grammars and a natural language corpus. In the latter case, they train the model on a subset of the corpus sentences. Based on the grammar acquired, they then generate novel sentences. These sentences are judged by human raters to be as grammatical as the remaining sentences from the corpus.

Thus, like UDOP, Solan et al.'s (2005) approach is successful in acquiring human-like grammatical representations that take into account the fact that formulaic, multi-word sequences are endemic to language. Furthermore, they do this based on knowledge of smaller word-level representations, which may appear counterintuitive to observations that children begin by producing 'unanalyzed' formulaic sequences. Yet despite this word-level knowledge, the trajectory of Bod's approach nonetheless mimics child language acquisition—the model starts with more formulaic sequences and later develops more abstract representations.

At the same time, these studies are not designed as dedicated models of the earliest stages of children's acquisition of syntax, but rather as general cognitive models of

distributional learning of grammar/formulaic language. Bannard et al.'s (2009) model, on the other hand, is conceived specifically for exploring the problem of children's language acquisition. The authors take two corpora of two children's speech (without adult utterances) and bisect each one into an earlier training and later test partition.

They then induce grammars from the training sets using a computational algorithm. The algorithm proceeds by extracting, for each utterance, all word *n*-grams (from unigrams up to the length of the utterance). Then, for each *n*-gram, the algorithm finds alignments with *n*-grams from previously processed utterances that share some degree of overlapping material. These alignments are then used to generate schemas (i.e., partially or fully lexically specified multi-word formulas), in which the non-overlapping parts of the alignments are replaced by slots. Given this collection of schemas, for any particular utterance there will be many possible parses, which are hierarchical in nature (since the slots of schemas can and will be occupied by further schemas, which themselves may contain slots). Like the above approaches, the induced grammars all contain lexically specific content (i.e., there are no schemas in these child grammars that contain only slots and no specific words).

The authors then use a Bayesian inference procedure for unsupervised grammar induction to find a posterior distribution over possible grammars. After the probabilities of candidate grammars have been calculated, the authors examine the degree to which winning grammars account for child utterances at ages 2 and 3. They find that, at age 2, most of the utterances that the grammars of the two children can account for require a single parse operation. However, at age 3, most of the utterances accounted for by the grammars require two to five operations. In other words, at age 3, both children's language is more productive. Furthermore, while each child's individual grammars poorly account for the other child's

language at age 2, at age 3 each grammar can be used on either child's utterances relatively successfully.

Overall, then, Bannard et al.'s (2009) computational approach demonstrates a trajectory by which children's emergent grammars can start out at a more lexically specified stage, and then later be used to generate more complex, creative hierarchical structures. Moreover, it does this by operating on individual words—remember that, at the first step, word *n*-grams are extracted as the basic units for the induction of grammars. At the same time, however, it does not trace back to the actual learning input, since the grammar induction process is only run on child utterances. Thus, Bannard et al.'s (2009) study does not directly model children's acquisition of grammar in the same way that UDOP does; rather it models progressive maturation in children's grammars as they age.

Across the three cognitive computational models discussed in this subsection, the trajectory of representations learned moves from more lexically specified to more abstract—as a child does—even though the basic unit of which the models have knowledge is the word. In this way, we can assert that children's early production of whole formulas is not incompatible with the generation of these formulas from word-level representations. And, more importantly, facts about chunking and distributional learning can be integrated with facts about formulaic language. Furthermore, while these models generate more abstract structures with increased learning, they still continue to use lexicalized subtrees/schemas/significant patterns even in their mature states, reflecting the research on adult formulaic language processing that shows that widespread formulaic language is not merely an early childhood phenomenon necessary to get grammar off the ground.

These computational studies also complement the behavioral studies discussed above in the evidence they provide against traditional generative notions about the mental representation of lexical and grammatical knowledge. On the one hand, these models help expose the artificiality of a divide between the lexicon and syntax. While they induce productive syntactic structures, they do so via the storage and retrieval of (partially) lexicalized multi-word structures, which are partially redundant with smaller, individual words, and often with larger multi-word structures. That is, despite the compositionality of these structures, they are remembered and used as units by the models nonetheless. On the other hand, these models provide evidence against Pinker's (1984) claims that (1) distributional learning cannot reveal abstract grammatical representations necessary for productive grammar, and (2) unfettered co-occurrence tracking would lead to a combinatorial explosion. And while the behavioral research gave some evidence for why these claims are invalid, the computational research operationalizes specific learning mechanisms that result in productive grammar and a properly constrained search space for co-occurrences.

These algorithms do this, crucially, through implementations of the above-mentioned chunking and analogy mechanisms, theorized as fundamental to distributional learning of language. While subtrees (Bod 2009), significant patterns (Solan et al. 2005), and schemas (Bannard et al. 2009) are the product of the syntagmatic process of chunking, how the subtree nodes combine (Bod 2009), equivalence classes (Solan et al. 2005), and the different items that can fill slots in schemas (Bannard et al. 2009) represent the paradigmatic process of analogy. Overall, then, these two mechanisms, theorized to challenge the generative

model of acquisition and representation, are shown to yield human-like linguistic knowledge when implemented computationally.

2.3.2 Non-Cognitive Models of Distributional Learning of Formulaic Language

Returning to the categorization of distributional learning algorithms discussed in chapter 1, the cognitive approaches discussed above represent parsing/grammar induction algorithms, which learn both syntagmatic and paradigmatic structures, based only on knowledge of word boundaries. In the current section, I focus on non-cognitive algorithms for the distributional learning of multi-word structures. Unlike the parsing/grammar induction algorithms, remember that these extraction and ranking algorithms are designed to learn purely syntagmatic structures. They generate a list of candidate multi-word structures from a corpus, and then score and rank them according to some statistical metric of co-occurrence strength. Those items ranked highest represent the model's best hypotheses for true multi-word formulas, and those ranked lowest represent the model's best hypotheses for what are not multi-word formulas. Unlike parsing/grammar induction algorithms, they are not a 'full' model of human linguistic knowledge, since they do not represent the productivity associated with the combination of items from abstract form classes. Furthermore, they tend not to be implemented in an incremental way as are cognitive algorithms, which are designed to closely simulate a child's exposure to language.

Extraction and ranking of multi-word structures as an approach is in fact quite old. Long before recent research on the role of formulaic language in acquisition, its representation in adult minds, and the sufficiency of distributional learning as an acquisition mechanism, corpus linguists were generating lists of collocations ranked based on their co-

occurrence strengths. Through these scholars' work with textual data, it was widely accepted in this community that mature human knowledge of language must extensively include lexically specific combinations that generative phrase structure-based theories were not powerful enough to account for (e.g., Nattinger 1980; Becker 1983; Pawley and Sider 1983; Biber et al. 2004). In Nattinger's prosaic words, "…language production consists of piecing together the ready-made units appropriate for a particular situation and…comprehension relies on knowing which of these patterns to predict in these situations" (1980, p. 341).

However, the fact that these corpus linguists' purview was primarily textual data meant that these claims would not become accepted as mainstream notions of how cognitive processes involving language function until the empirical studies reviewed earlier in this chapter were conducted. Thus, corpus linguists have long pursued techniques that take into account this extensive lexical specificity of real language, yet they were not attempting to (or accepted as) modeling specific cognitive processes; instead, their techniques were motivated by an understanding that corpus-based non-cognitive research must work with structures that accurately reflect what humans know about their language, and generative approaches were inadequate with regard to formulaicity.

Now, with the emergence of widespread empirical research on formulaic language, as well as efforts to quantify the amount of formulaic language in discourse (see 2.2.2.2), the field of linguistics in general has come to appreciate the fundamental importance of formulaicity. Today, there is widespread interest in computational techniques not just to model the cognitive processes involved in the acquisition and use of formulaic language, but to extract formulas from corpora and rank them for use in other non-cognitive applications. As mentioned above, these include the use of formulaic language as the basis for the

differentiation between varieties of the same language (Gries and Mukherjee 2010) and between genres within a single language (Biber et al. 2004); creating multi-word dictionary entries in lexicographic work (Sinclair 1987); development of native-like abilities in second language acquisition (e.g., Sinclair 1987; Simpson-Vlach and Ellis 2010); and creating native-like speech in natural language generation (Lareau et al. 2011), among others.

There are two major extraction and ranking problems faced by corpus linguists: how to score a co-occurrence of words, and how to choose the collection of words whose co-occurrence is quantified (i.e., size(s) of *n*-grams to be extracted, and whether or not there are discontinuities). Interestingly, there has been a tremendous amount of corpus linguistic effort that has gone to the former problem, while relatively little has been devoted to the latter. In the next subsection, I discuss how corpus linguists go about scoring word co-occurrences. Then, in the subsection that follows, I discuss different approaches to how the size and shape of co-occurrences is computationally specified.

2.3.2.1 Lexical Association Measures

Various statistical approaches have been used in the literature to quantify and compare the strengths of word co-occurrences, including simple metrics such as raw or relative frequency and transitional probability. Most widely accepted, however, are lexical association measures, which have been primarily developed by corpus linguists and are based on contingency tables (Tables 2.2 and 2.3). Note that these contingency tables are based on two-word co-occurrences, or bigrams: lexical association measures are in general optimized for calculating the strength of bigrams (rather than, say, trigrams). To adapt them to larger *n*-grams requires adjustments in how the association measure scores are calculated.

One such method is developed in Da Silva et al. (1999). However, there is not one widely accepted method for doing this.

For every bigram type found in a corpus, two contingency tables can be used to represent across their different cells the various pieces of frequency information relevant to that bigram type. The first contingency table represents the **observed frequency** information for any bigram type with component words *x* and *y* (Table 2.2). In the other contingency table (Table 2.3), each cell again corresponds to the same relationships between the component words *x* and *y*, except that the cells contain **expected frequency** information rather than bigram frequency values actually observed in the corpus. These expected values are calculated by multiplying the observed individual frequencies of the word components, and then dividing by the corpus size.

Generally, lexical association measures are based on various mathematical formulae that compare cell value(s) from the observed frequency contingency table to cell value(s) from the expected frequency contingency table. Using an association measure's formula, one can calculate an association score for each bigram type; these scores may be used to rank the bigrams in a corpus by strength. While each measure's scores represent different units, often a positive value of a score will indicate statistical association between two words: that is, that the two words co-occur more often than might be expected by chance. Conversely, a negative value will indicate statistical repulsion, or that two words occur less frequently than might be expected by chance.

Numerous association measures have been developed over the years—Pecina (2009) reviews 80 different measures—and employed in various corpus-linguistic applications. For example, in studies discussed above, lexical association measures employed have included

Table 2.2. Observed Frequencies, for any Bigram *x,y*

	y present	*y* absent	totals
x present	*a* (freq of bigram "*x*, *y*")	*b* (freq of all "*x*, not *y*" bigrams)	*a* + *b* (freq of all *x*-initial bigrams)
x absent	*c* (freq of all "not *x*, *y*" bigrams)	*d* (freq all "not *x*, not *y*" bigrams)	*c* + *d* (freq of all non-*x*-initial bigrams)
totals	*a* + *c* (freq of all *y*-terminal bigrams)	*b* + *d* (freq of all non-*y*-terminal bigrams)	*a* + *b* + *c* + *d* (size of the corpus *n* in bigrams)

Table 2.3. Expected Frequencies, for any Bigram *x,y*

	y present	*y* absent	totals
x present	*a* (freq of *x* * freq of *y* / *n*)	*b* (freq of *x* * freq of not *y* / *n*)	*a* + *b* (freq of all *x*-initial bigrams)
x absent	*c* (freq of not *x* * freq of *y* / *n*)	*d* (freq of not *x* * freq of not *y* / *n*)	*c* + *d* (freq of all non-*x*-initial bigrams)
totals	*a* + *c* (freq of all *y*-terminal bigrams)	*b* + *d* (freq of all non-*y*-terminal bigrams)	*a* + *b* + *c* + *d* (size of the corpus *n* in bigrams)

mutual information, log likelihood, and the Dice coefficient.[3] Nonetheless, in most of the cognitive studies of formulaic language, measures of simple frequency (or transitional probability) have been used. At the same, corpus linguists have long recognized that lexical association measures, which are more mathematically sophisticated, tend to reveal more interesting multi-word sequences that better reflect 'real' formulaic language. However,

[3] Technically, the log likelihood equation represents a variant of mutual information; however, because in lexical association applications log likelihood and mutual information are conventionally treated as distinct, and correspond to two distinct equations, I will herein refer to them as separate measures.

little cognitive work has yet adopted these more sophisticated approaches (but see Stefanowitsch and Gries 2003).

While, in general, contingency-based measures appear better equipped to account for formulaic language, there is considerable variability in the performance of these different measures. Ironically, some of the most popular ones, such as transitional probability and mutual information, appear to report high strength values for many low frequency word combinations (e.g., Daudaravicius and Murcinkeviciene 2004) that do not represent good examples of formulaic sequences.

One lexical association measure that has yielded quite good results compared to some of these other measures is log likelihood (Dunning 1993). In particular, it does not fall victim to this problem of inflating the strength of low frequency co-occurrences. Part of the reason for its sound performance is that it takes into account information from all four frequency cells (A – D) from both the observed and expected contingency tables. Another reason for the strong performance of this measure is that it provides a close approximation to Fisher's Exact Test (Evert 2005), considered on mathematical grounds to be the best method for quantifying statistical association (yet its computational cost to implement makes it prohibitive for iterative applications like MERGE). The formula for the log likelihood measure is given in Equation 2.1.

While log likelihood is developed in Dunning (1993) as a lexical association measure, it is in fact a multiple of another measure known as the Kullback-Leibler Divergence from the field of Information Theory (Evert 2005). This field is a branch of mathematics and computer science that emerged in the middle of the last century, with the goal of quantifying the amount of possible information (in bits) conducted through particular

channels, given the properties of minimal encoding schemes (Shannon, 1948). Kullback-Leibler was not developed to quantify word co-occurrences, but rather to measure the difference between two discrete probability distributions that share the same domain. Nevertheless, in its guise as a lexical association measure, it has shown itself to be quite effective (e.g., Wahl 2015), and, for this reason, it is the measure that I use in the various implementations of the MERGE algorithm reported in this dissertation.

Log likelihood/Kullback-Leibler values essentially report how much information would be lost by representing the target bigram using the expected contingency table frequencies rather than the observed ones. Note that, based on Equation 2.1, a log likelihood value will always be positive, regardless of whether the observed frequency of the target bigram is greater than or less than the expected frequency of that bigram. Thus, once the log likelihood values are calculated, the values corresponding to bigrams in which the observed frequency is less than the expected frequency are multiplied by -1. In this way, the resulting values correspond to the convention discussed above whereby positive association measure values denote attraction and negative values denote repulsion.

Equation 2.1. Formula for Log Likelihood Measure

$$\text{log likelihood} = 2\sum_{i=a}^{d} i_{observed} * \log\left(\frac{i_{observed}}{i_{expected}}\right)$$

2.3.2.2 Extraction and Ranking Algorithms

Once a metric for quantifying the strength of word co-occurrences is chosen, a method for how to extract co-occurrences from a corpus must be selected. Throughout the history of corpus-linguistic work on collocations, by far the dominant approach to this problem has been the simple pre-specification of the type of *n*-grams to be studied. The most typical type examined is adjacent bigrams (e.g., Evert 2005), primarily because of the above-mentioned fact that lexical association measures are optimized for bigrams. Of course, many MWEs are longer than two words. Because of this, many high-ranking bigrams may be only a piece of the larger, 'true' unit. In addition, as we have seen, formulaic expressions may be discontinuous. Ultimately, then, extraction and ranking approaches that can identify MWEs of various lengths in a bottom-up fashion, with or without discontinuities, are ideal (similar to how parsing/grammar induction algorithms can identify lexicalized multi-word sequences of various lengths). There are a handful of such approaches from the literature, but I will focus on two that have been cited somewhat extensively.

First, Da Silva et al. (1999) develop an algorithm entitled LocalMax, which learns what they call *multi-word units*, which may or may not include discontinuities. At the first step, all *n*-grams (again, with or without discontinuities) are extracted from a corpus up to some maximum *n*-gram size. Then, co-occurrence strength is calculated for each *n*-gram. The authors develop two new statistical measures for quantifying co-occurrence strength—one for continuous *n*-grams (Symmetrical Conditional Probability [SCP]) and one for discontinuous n-grams (Mutual Expectation [ME]). Next, the strength of each *n*-gram is compared to the strength of each *n*+1-gram that contains the target *n*-gram, and each *n*-1-

gram that is contained by the target *n*-gram. If the strength of the target *n*-gram is the highest among these scores—that is, it is a local maximum—it is selected as a multi-word unit.

Da Silva et al. (1999) evaluate the continuous multi-word unit-finding and discontinuous multi-word unit-finding versions of LocalMax using a Portuguese news corpus. They compare the performance of each of these versions of the models implemented using the SCP or ME measures to implementations using other conventional association measures from the corpus linguistic literature. These include specific mutual information and log likelihood, among others.[4] The items output by the model are considered to be multi-word units if they are proper names, compound names, compound verbs, "frozen forms," and "other n-grams occurring relatively frequently and having strong 'glue' among the compound words" (p. 10). Ultimately, the models implemented using the statistical measures developed by the authors perform the best (precision scores in the 80% to 90% range), while the other measures perform in the upper 40% to upper 70% range. However, particularly in the case of these last two criteria, there is room for researcher interpretation/bias, given that the notion of frozen forms and other *n*-grams having strong glue are rather vague and broad. Thus, it is difficult to say how reliable these results are (see chapter 4 for further discussion).

Brook O'Donnell (2011) develops a purely frequency-based approach entitled the Adjusted Frequency List (AFL). The algorithm works by first extracting all *n*-grams (continuous only) up to some user-defined threshold. Next, only *n*-grams exceeding some frequency threshold are retained in the AFL along with their frequency. Then, for each *n*-

[4] Because these measures are optimized for use with bigrams, they had to be adjusted for use with larger *n*-grams via the Fair Dispersion Point Normalization developed by the authors.

gram, starting with those of threshold length and then descending by order of length, all component *n*-grams are derived. Finally, the number of tokens of each component *n*-gram making up the tokens of the target *n*-gram are subtracted from the frequency counts of the corresponding types in the AFL. Unfortunately, Brook O'Donnell's initial work is primarily descriptive; he does not evaluate the performance of his approach against a gold standard, or in comparison to other approaches, in order to provide the reader with a sense of the quality of the formulaic sequences that the algorithm extracts.

There are a number of interesting points of contrast between these two approaches. First, Brook O'Donnell's (2011) approach allows for continuous *n*-grams only. Second, while Da Silva et al.'s (2009) approach allows for *n*-grams within *n*-grams (as long as their difference in number of constituent words is greater than or equal to two), the AFL does not. Finally, while the AFL is based on simple frequency counts, LocalMax is implemented using different contingency table-based association measures. In chapter 5, I will discuss these algorithms further, and I will compare the performance of MERGE to that of the AFL.

2.4 MERGE as Novel Extraction and Ranking Algorithm

As I highlighted at the end of chapter 1, MERGE addresses important gaps in the current literature on distributional learning algorithms that are used to explore both cognitive and non-cognitive questions. On the one hand, the cognitive models reviewed above succeed in acquiring MWEs of various shapes and sizes in a bottom-up, cognitively realistic fashion, but they all embody the parsing/grammar induction architecture. As Swingley (2005) shows (and as I discussed in chapter 1), the extraction and ranking approach—which is the standard approach used in non-cognitive algorithms—can be fruitfully used generate results that are

compatible with human-like linguistic knowledge. Furthermore, the nature of the output representations relate to human-like representations in a fundamentally different way than the output representations of parsing/grammar induction algorithms do. For example, while parsing/grammar induction algorithms must commit to a winning parse for any given utterance, and then compare this winning parse against a gold standard, extraction and ranking approaches offer a different kind of evidence. They output a list of scored and ranked items, and then a correlation can be examined between the scores of these items and their varying closeness to a gold standard. Thus, evidence for the cognitive viability of the learned representations achieved by extraction and ranking approaches is provided by a gradient measure.

On the other hand, most of these non-cognitive, extraction and ranking algorithms require the researcher to specify beforehand the size(s) of the n-grams to be learned, as well as whether there are any discontinuities, which is not ideal for reasons discussed above. And while there is currently one algorithm that I am aware of that does learn both of these features in a bottom-up fashion (Da Silva et al. 1999), it has not emerged as a gold standard, and the authors have not attempted to make any arguments as to potential cognitive applications of the algorithm.[5]

[5] After completion of this project, I discovered an algorithm from 2006 conference proceedings that bears important similarities to MERGE and should be mentioned (Wible et al. 2006). Much like MERGE, it works by recursively combining bigrams to form longer and longer MWEs, whereby the input to the next iteration may include a multi-word structure output at the last iteration. Also like MERGE, it selects bigram candidates for combination on the basis of a lexical association measure. However, there are important points of contrast. First, while MERGE is envisioned for both cognitive and non-cognitive applications, Wible and colleagues' approach is designed purely as a non-cognitive tool for the identification of conventionalized word sequences in second language learning contexts. To operate it, the user inputs a particular node word and the algorithm returns a list of MWEs in which that word participates. Thus, unlike MERGE and the other distributional learning algorithms discussed in this chapter, it does not learn a large lexicon of MWEs containing various words from throughout the corpus. Consequently, it cannot be used to model one's knowledge of MWEs in any sort of global way, or one's knowledge of a text parsed in terms of MWEs. In addition, unlike MERGE, the

MERGE therefore embodies a novel extraction and ranking architecture that learns MWEs of varying lengths, potentially with discontinuities of varied lengths. Furthermore, it does so through a process of recursively merging word bigrams into potentially larger and larger MWEs. This process is related to the chunking mechanism, as well as to the binary-splitting tree approach employed in Bod (1999), and in this way is compatible with known aspects of human learning.

There are, however, a couple of limitations to the MERGE approach that must be noted. First, as I have stated, the batch approach of extraction/ranking approaches such as MERGE offers a greater degree of abstraction away from the online, incremental learning environment faced by the child and modeled in most parsing/grammar induction approaches. Second, while MERGE does acquire discontinuous MWEs such as *put _ down*, it cannot learn that *put _ down* and *put _ _ down* (with two intervening words) are equivalent. Furthermore, MERGE cannot distinguish between homonymous MWEs.

Ultimately, however, MERGE is able to acquire formulaic structures which, as we have seen in this chapter, are fundamental to and omnipresent in human language. In the next chapter, I describe in detail the MERGE algorithm.

---

algorithm runs on a part-of-speech-tagged corpus, so it works with considerably more built-in knowledge than algorithms used to explore cognitive questions typically do. Finally, while the algorithm does allow for discontinuities (up to a gap size of 5 words), once a discontinuous bigram is combined, future iterations of the model cannot consider candidates in which one element lies in the middle of the other element. For example, there would only be two pathways for the algorithm to acquire *in spite of*: the combination of *in* and *spite of* (where the latter gram is the output of a previous iteration), or *in spite* (again, output from before) and *of*. MERGE, on the other hand, could also arrive at *in spite of* via the combination of previously-merged *in _ spite* and the unigram *of*, since it searches for new candidates wherein one element can lie in the gap of the other element. In this way, MERGE is more flexible in the pathways it can take to arrive at MWEs, which may mean that it can find MWEs that Wible et al.'s (2006) algorithm would exclude.

3. The Operation of MERGE

In this chapter, I detail in prose how the MERGE algorithm operates. MERGE was developed in and runs from the Python programming language. I begin the chapter by discussing the preprocessing steps that need to be performed on a corpus in order to make it available as input to MERGE. Then, I discuss the selection of a metric for assessing co-occurrence strengths of word bigrams. In the present dissertation, I use the log likelihood metric. Finally, I turn to the step-by-step description of the algorithm itself. MERGE is an iterative algorithm that involves two major blocks of code: the initialization and choice of the first winning bigram, and an iterative loop that handles the choice of all subsequent winning bigrams. I discuss each of these blocks of code in turn.

3.1 Corpus Preprocessing

Before MERGE begins, a corpus must be selected. In MERGE's current implementation, this corpus must be preprocessed to remove all non-alphanumeric characters (though, in principle, this would not be obligatory under modified implementations). In corpora of spoken language such as the ones used in the simulations reported in chapters 4 and 5, such characters often denote features such as in-breaths, laughter, pauses, etc. Non-alphanumeric characters that are part of the orthography of particular words—such as hyphens—or which denote important morphosyntactic relations—such as apostrophes—can be replaced with alphanumeric representations that signal these features, and they can be capitalized to set them apart from the surrounding

characters that signal graphemes (or phonemes, allophones, etc.). For example, a word like *let's* can be represented as *letAPOSs*.

3.2 Selection of a metric of bigram strength

Another step that must take place prior to running the algorithm is the selection of some metric for quantifying the strength of word co-occurrences. As discussed in the previous chapter, numerous metrics for assessing co-occurrence strength have been employed in various lines of research, and MERGE could in principle be implemented with any one of these. Perhaps most typical among those in the literature is simple co-occurrence frequency. However, as some of this research has shown, metrics that assess the contingency of co-occurrence patterns—rather than just the frequency of co-occurrence—seem to better account for human learning of multi-word sequences (e.g., Aslin et al. 1998; Gregory et al. 1999; McDonald and Shillcock 2003). These contingency-based measures include transitional probability as well as numerous more sophisticated lexical association measures. These latter measures, developed by various corpus linguists, are numerous—Pecina (2009) reviews 80 different lexical association measures—and they include popular measures such as mutual information, the Dice coefficient, and log likelihood. For reasons extensively discussed in the previous chapter, I have selected log likelihood as the lexical association measure for assessing bigram strength for all of the current implementations of MERGE reported in this dissertation.

3.3 The algorithm

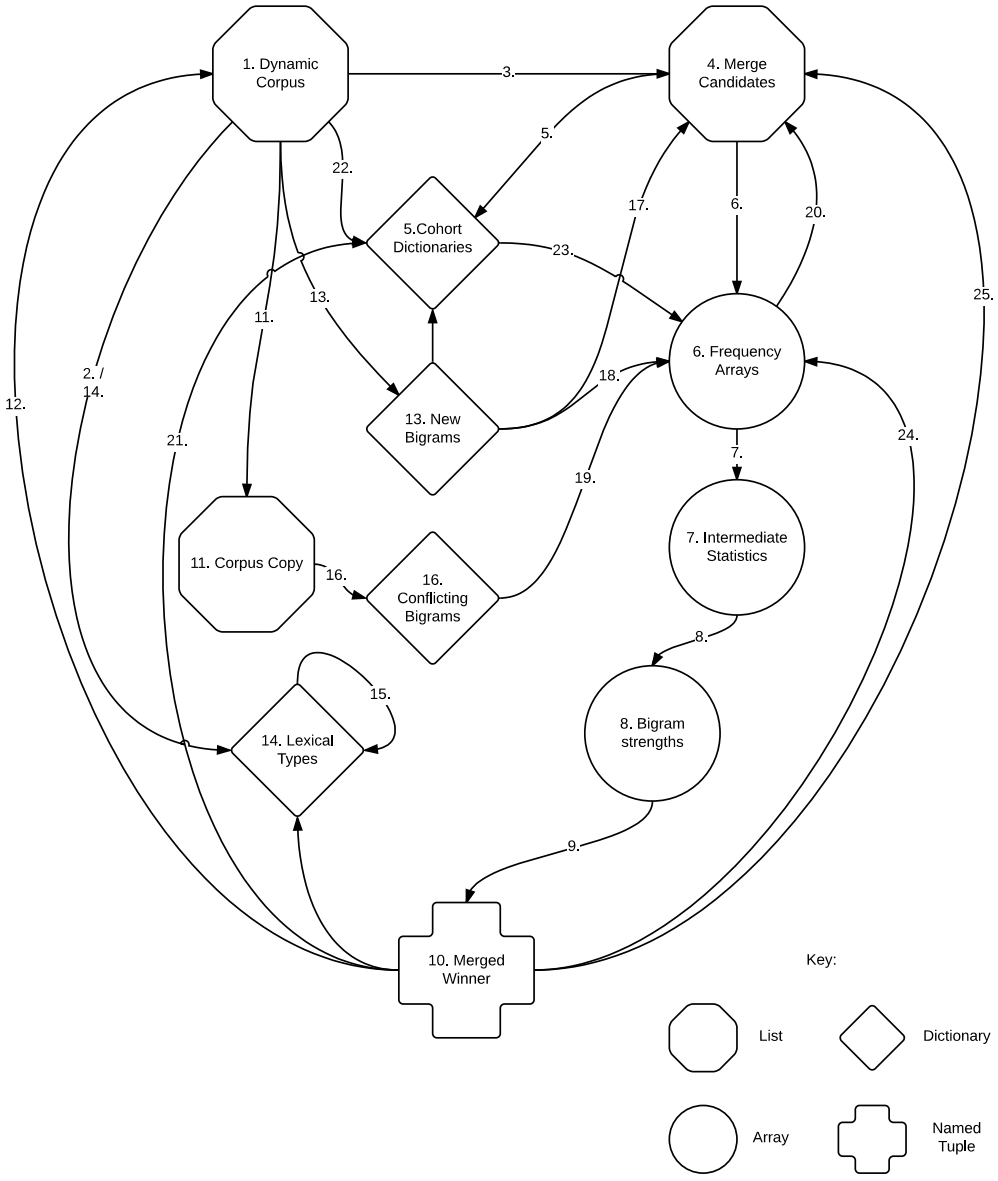
Once preprocessing is complete and a metric of bigram strength has been chosen, MERGE can begin. A general sketch of the procedure that the algorithm follows is depicted in Figure 3.1.

3.3.1 Initialization Block

As a first step, the preprocessed corpus is loaded into a class of object in Python called a list; this particular list is called *dynamic_corpus* **(1)**. This list has two levels of hierarchical subdivisions: the deepest may correspond to turns, paragraphs, or some other genre-internal subdivisions within the corpus. Here, I use turns. Note that the model does not recognize further subdivisions deeper than the turn/paragraph, such as utterance. The higher subdivisions are bins that are available for the calculation of, for example, dispersion measures. Dispersion refers to how spread out across a corpus particular tokens of a type are. In the current implementation, these upper bins are not utilized.

As the corpus is loaded into the *dynamic_corpus* list, each word is instantiated in a named tuple (another type of Python object) called a *lextoken* (not depicted in Figure 3.1). Each instance of a *lextoken* contains further named tuples embedded within it, which together contain the word (as a string) along with other important information about the word's position (e.g., within the turn). In addition, a *lexical_types* dictionary (a third object type) is generated, which contains each *lextoken* type as well as its corpus frequency and the order in which it was merged (if the *lextoken* happens to be the output of a merge; see below) **(2)**.

Figure 3.1. The MERGE Algorithm



At the next step **(3)**, a bigram extraction function is run across the corpus to extract all adjacent and nonadjacent bigrams up to a user-defined maximum gap size. Each bigram is represented within an instance of a named tuple called *bigram_locs* (not depicted in Example 3.1). Each instance of the named tuple contains two *lextokens* (the 1st and 2nd elements of the bigram) as well as information about the gap size of that bigram. The set of these named tuples are stored in a *merge_candidates* list **(4)**.

During this process of bigram extraction, two dictionaries are created, entitled *left_cohort* and *right_cohort* **(5)**. These dictionaries are used to track all of the different bigrams in which a particular *lextoken* participates as the right or left element, respectively. More specifically, the *left_cohort* dictionary reports, for a particular *lextoken*, the positions in the *merge_candidates* list of all bigrams in which that *lextoken* appears as the right element. Conversely, the *right_cohort* dictionary reports, for a particular *lextoken*, the positions in the *merge_candidates* list of all bigrams in which that *lextoken* appears as the left element. These initialized dictionaries play a vital role in updating frequency information at a later stage of the algorithm.

Next, information about the corpus frequency of *bigram_locs* types as well as the corpus frequencies of the individual *lextoken* types making up each *bigram_locs* type are stored in a series of arrays, a fourth Python object type **(6)**. These arrays maintain the same order as the *bigram_locs* in the *merge_candidates* list.

The frequency arrays are then used in array-wise calculations to generate the various intermediate statistics that are used in the determination of bigram strength **(7)**. These intermediate statistics are next used to calculate the strength for each *bigram_locs* type **(8)**. Here, for example, the intermediate statistics correspond to the frequencies in the different

cells of the expected and observed contingency tables (Tables 2.2 and 2.3), which are then used in the calculation of the log likelihood measure. These intermediate and final statistics are also instantiated in arrays that exhibit the same ordering as the *merge_candidates* list.

At the next step **(9)**, the bigram type exhibiting the highest strength is found, and this bigram is merged into a new, single lexical item that is instantiated in a *lextoken* called *winning_bigram* **(10)**. This demonstrates a crucial fact about *lextokens*: they are not simply equivalent to words, but rather to elements. Although initially every *lextoken* will contain but one word, after merges start taking place, some *lextokens* will start to contain more and more words.

Once the merged *lextoken* type is generated, two things take place. First, a copy of the corpus in its current state is created **(11)**. This copy is reserved for later (see below). Once this is complete, *lextoken* replacement can proceed **(12)**: a loop scans through the dynamic corpus (not the copy) and looks for *lextokens* that correspond to the first (left) element of the winning bigram type. These old *lextokens* are replaced with the new *lextoken* representing the merged lexical item. This particular *lextoken* is called the *merged_winner*. Simultaneously, the *lextoken* that corresponds to the right element of the winning bigram type is also replaced, but by a placeholder that points to the old left element's (i.e., new *merged_winner's*) position.

Let us take an example. Consider the turn-at-talk from a spoken English corpus represented in Example 3.1 as a list of *lextokens*. The word (or words) represented by a *lextoken* can be found within single quotes within that *lextoken*. Now let us say that the winning merge at a particular iteration is the discontinuous bigram *a_of* (where the underscore here represents a single intervening word). Note that, in the turn in Example 3.1,

the seventh element in the sequence is the *lextoken* representing the word *a*, while the ninth word in the turn is the *lextoken* representing the word *of*. (In Python, the initial element in a sequence is in position 0). Thus, together these items instantiate the winning bigram, and are subject to replacement with the new merged representation.

Example 3.1. Sample Turn before Merge

```
[NTlextoken(lexeme=(NTwordpos(word='then', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='i', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='went', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='off', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='worked', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='for', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='a', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='couple', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='of', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='companies', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1))]
```

Once the replacement loop just described passes over this turn, the representation of the turn will be changed so that it appears as in Example 3.2. That is, the leftmost element of the winning bigram (here, the *lextoken* representing *a*) is replaced with the new *merged_winner*, a single *lextoken* that combines both elements of the bigram. Note in all *lextokens* the *wordpos* field. In *lextokens* containing only one word, the value of this field is always 0. In *lextokens* containing more than one word, as in the current case, each word is paired with a *wordpos* field that tracks the relative position in the MWE of that word (again, starting with position 0). Thus, here, the word *of* for the MWE *a_of* has a *wordpos* value of 2 (since it is 2 positions to the right of *a*, the leftmost word of the MWE).

Example 3.2. Sample Turn after Merge

```
[NTlextoken(lexeme=(NTwordpos(word='then', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='i', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='went', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='off', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='worked', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='for', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='a', wordpos=0), NTwordpos(word='of', wordpos=2)),
present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='couple', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='of', wordpos=0),), present=(9, 7)),
 NTlextoken(lexeme=(NTwordpos(word='companies', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1))]
```

In addition, the other element of the winning bigram, the *lextoken* representing *of*, is replaced by a placeholder *lextoken*. Normal *lextokens* are assigned a value of *(-1,-1)* to the *present* field. Placeholder *lextokens* are assigned a value to the *present* field whereby the left number refers to the position in the turn of the placeholder while the right number refers to the position in the turn of the leftmost element (the *merged_winner*) of the MWE to which the placeholder belongs.

The relationship between leftmost elements where the *lextokens* instantiating the MWEs are located, and placeholders to their right, is fundamental to the functioning of MERGE. This becomes evident at subsequent steps of the algorithm. After the winning bigram type has been merged into the new *lextoken* and the corpus tokens of the bigram's left and right elements replaced by tokens of this *merged_winner* and placeholders, respectively, the set of potential future merges has changed.

On the one hand, remember that the new merged MWE behaves as a single unit (a *lextoken*). This means that it should likewise be able to act as one of the elements in a bigram, even if it actually contains more than one word. This is the mechanism by which the

algorithm is able to create longer and longer chains of words. Thus, following a merge, there are many new bigram merge candidate types that could not have existed before (i.e., those in which one [or both] of the elements are the new *merged_winner*), and these new candidates need to be generated.

To find new merge candidate tokens, a loop moves across the *dynamic_corpus* and generates bigrams in which one of the elements is either a newly merged *lextoken* or one of its placeholders. If the element is a placeholder, this element is replaced with the left-headed *merged_winner* and positional information is updated.

Consider the case in Example 3.2 again. First, new merge candidate tokens will be found in which one of the elements is the *merged_winner*. Assuming a maximum discontinuity threshold of 1, these new candidates will include *worked + a_of*, *for + a_of*, and *a_of + couple*. In addition, new merge candidate tokens will be found in which one of the elements is the placeholder. Again, with a maximum discontinuity of 1, these new candidates will include *couple + of*, *of + companies*, and *of + and*. Because the placeholder is replaced with the left-headed *merge_winner* for these latter new candidates, they become *a_of + couple*, *a_of + companies*, and *a_of + and*. Note that these processes would twice yield the false candidate *a_of + of*; this type of artifact is automatically removed.

Every new bigram extracted contains information about its absolute location in the corpus. Therefore, when the set of these bigrams is generated, redundancies among tokens are removed. Note in the previous paragraph that *a_of + couple* is generated twice; the removal of redundancies makes it so that this token is only attested once. The set now represents the tokens of the new bigram candidates **(13)**.

On the other hand, potential future merges have also changed because particular *lextokens* no longer exist. That is, now that the old winning bigram has been merged, this means that the corpus instances of the *lextokens* that instantiated the two elements of the now-replaced winning bigrams have been replaced as detailed above. Therefore, the number of winning bigram tokens in the corpus should be subtracted from the frequencies of each of the winning bigram's two constituent lexical items in the *lexical_types* dictionary **(14)**. Any items in the *lexical_types* dictionary whose frequency is now less than one are deleted **(15)**.

In addition, these now-discarded tokens of the two lexical item types likely participated in other merge candidate bigrams that were in the vicinity in the corpus of the tokens of the winning bigram types. These 'conflicting' bigram tokens must be found, and their array-based frequencies reduced by the number of winning bigram tokens. In order to find the conflicting bigram candidates, a loop moves not across the dynamic corpus but rather across the copy of the corpus that was created right before the winning bigram tokens were replaced (see above). Here, the loop generates bigrams in which one of the elements is also one of the elements of a winning bigram token.

Returning to the example in Example 3.1, bigram candidates that would conflict with the winning merged representation of *a_of* include *worked + a*, *for + a*, *a + couple*, *couple + of*, *of + companies*, *of + and*, and, of course, *a + of* (once again, this is all assuming a maximum discontinuity of 1). Again, information about absolute locations of bigrams in the corpus is maintained in order to avoid any redundancies. The set of these bigrams represents the tokens of the conflicting bigram candidates **(16)**.

At the next step, the merge candidates list and corresponding frequency arrays must be updated. First, the information about the new candidate bigrams is added to

merge_candidates and the co-indexed frequency arrays **(17 and 18)**. In addition, frequencies in the arrays are reduced for conflicting bigram candidate types (**19**). At this point, candidates whose frequencies are zeroed out as a result of this reduction can be removed from *merge_candidates* **(20)**.

There is another category of candidate bigrams that must have their frequency information adjusted. Remember that the calculation of the log likelihood measure includes information about the total corpus frequencies of the elements of a given bigram. Thus, all those bigrams in which one of the elements is of the same type of *lextoken* as one of the elements of the winning bigram must have the frequencies in the arrays corresponding to left and right element frequencies reduced by the number of winning bigram tokens. This is accomplished by looking up the right and left elements of the winning bigram in the cohort dictionaries **(21)**, as these dictionary values will report the other merge candidates that these right and left elements participate in. The algorithm then takes the number of winning bigram tokens in the *dynamic_corpus* **(22)**, and subtracts this value from the corresponding positions in the frequency arrays of the other looked-up merge candidates **(23)**.

Finally, the winning bigram type is removed from the *merge_candidates* list and from the concomitant frequency arrays **(24 – 25)**.

3.3.2 Subsequent Iterations Block

The code representing Block 2 is essentially identical to items 8 through 25 in Block 1, the main difference being that the code is embedded within a loop that continues iterating until some condition no longer holds. This structure is designed to allow the user to define different cutoff criteria: for example, a user may decide that s/he wants MERGE to stop

running after 1000 iterations, or after the bigram strength descends below some minimum threshold parameter, etc. In this way, the algorithm continues recursively calculating the bigram with the highest strength and merging it into a single representation, replacing the corpus tokens of the bigram with the merged representation, and extracting new bigrams and adjusting frequencies of old ones, until the cutoff criterion is met.

The fact that the new bigrams generated at each iteration can contain one or more *merged_winners* enables the algorithm to learn MWEs of theoretically any length, with or without discontinuities (assuming that these new candidates are selected as winners at later iterations). For example, returning to the discussion of new candidate generation in the previous section, one of the new candidates was the bigram *a_of + couple*. Let us assume that this bigram is selected as the winner at a later iteration. Given the same *merged_winner*/placeholder replacement procedure detailed above, the same turn depicted in Example 3.2 would be transformed to the representation in Example 3.3. Now, the MWE still resides at position 7 in the turn (i.e., the position of the leftmost word of the MWE), but the *lextoken* now contains 3 words: *a*, *couple*, and *of*. Furthermore, now the position 8 *lextoken*, corresponding to the word *couple*, has become a placeholder, with its *present* field denoting its position in the turn as well as pointing to the leftmost word of its parent MWE.

Ultimately, the end products of this process are the *dynamic_corpus* list, now parsed in terms of MWEs, as well as the *lexical_types* dictionary, containing all MWEs, their frequencies and order in which they were merged, and any and all remaining single-word items found in the corpus.

Example 3.3. Sample Turn after Second Merge

```
[NTlextoken(lexeme=(NTwordpos(word='then', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='i', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='went', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='off', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='worked', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='for', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='a', wordpos=0),
NTwordpos(word='couple', wordpos=1), NTwordpos(word='of', wordpos=2)),
present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='couple', wordpos=0),), present=(8, 7)),
 NTlextoken(lexeme=(NTwordpos(word='of', wordpos=0),), present=(9, 7)),
 NTlextoken(lexeme=(NTwordpos(word='companies', wordpos=0),), present=(-1, -1)),
 NTlextoken(lexeme=(NTwordpos(word='and', wordpos=0),), present=(-1, -1))]
```

3.4 Conclusion

In this chapter, I have described the step-by-step procedure that the MERGE algorithm uses in its distributional learning of MWEs. At this point, we are now ready to deploy the algorithm in empirical research in order to evaluate its performance as a viable model of distributional learning of MWEs. In the next chapters, I detail three experiments and one corpus study to this end.

4. The Experimental Validation of MERGE

This chapter is devoted to evaluating the performance of MERGE. In order to evaluate any computational algorithm for the distributional learning of MWEs, one must employ a gold standard against which model output can be compared. More specifically, some way of establishing what sequences count as good MWEs is necessary. In the various strands of research described in chapter 2, scholars draw on a number of techniques to this end.

For example, in the ample research on frequency effects and formulaic language, formulaicity is typically established on the basis of a variety of behavioral indicators that evince holistic processing and storage, and these are then correlated with different frequency-based measures for particular word sequences. To the extent that these correlate, evidence is provided that the sequences under examination are indeed formulaic.

In some of the other research reviewed, researchers define particular morphosyntactic, lexical, and semantic criteria for formulaicity in an a priori fashion. For example, Da Silva et al.'s (1999) computational approach to extraction/ranking of multi-word units in corpora specifies that those sequences identified by their model that belong to one of the following categories would be deemed 'correct' multi-word units: proper names, compound names, compound verbs, 'frozen forms,' and "other *n*-grams occurring relatively frequently and having strong 'glue' among the compound words" (p. 10). Clearly, items like proper names and compounds are fairly uncontroversially identifiable; however, the identification of frozen forms and other *n*-grams occurring relatively frequently and having

strong ‘glue’ among the compound words leave open the possibility for an enormous amount of unscientific researcher intuition.

Relatedly, in Erman and Warren’s (2000) study of the density of formulaic language in corpora, the authors identify prefabs on the basis of whether sequences of two or more words (with possible intervening slots) demonstrate the phenomenon of restricted exchangeability, whereby one or more words in the sequence determines the selection of one or more other words (roughly equivalent to Pawley and Sider’s [1983] concept of native-like selection). In other words, exchanging one or more words with a synonym would result in a loss of idiomaticity. This criterion allows the authors to identify numerous prefabs, such as *good friends* (≠ *nice friends*) and not bad (≠ *not lousy*).

However, this criterion fails to pick up prefabs in which only one word is lexically specified, as well as habitual collocations such as *dark night* in which neither collocate truly determines the other. Furthermore, the authors point out that there are cases where no appropriate synonym may be found for one or more words to test for restricted exchangeability, such as in the case of *very well*. As a result, they admit that there are a number of word combinations that they accept as prefabs but “for which [they have] no other criterion than [their] intuition” (p. 33).

The myriad dimensions along which formulaic language can vary makes comprehensive definitions based on a priori morphosyntactic, lexical, and semantic criteria such as those discussed above nearly impossible. This is likely why the authors just cited find it necessary to turn to intuition in certain cases.

At first glance, a third approach to establish whether a sequence is a MWE, which appears to circumvent these issues of researcher bias, would be the use of a collocation

dictionary: that is, a list of sequences that have already been established as formulaic. Like Erman and Warren (2000), Moon (1998) examines the extent of formulaic language in a corpus, but uses the *Collins Cobuild English Language Dictionary* to establish whether particular word combinations are formulaic. Unfortunately, such an approach merely obscures the potentially unscientific nature of determining formulaicity from the researcher: it is not clear what role intuition about formulaicity plays for the lexicographers who construct these dictionaries in the first place!

A fourth possibility for establishing a gold standard for MWEs is through a rating experiment. In such an approach, human raters assign scores to word sequences on the basis of how well the sequences reflect certain specified criteria of formulaicity, and then these ratings are compared to frequency-based metrics for the sequences.

For example, in Ellis et al. (2008), the authors ask (1) experienced instructors of English for academic purposes and (2) experienced language testers to rate formulas on a scale of 1 to 5 according to three criteria. First, they ask six raters whether they think a sequence represents "a formulaic expression, or fixed phrase, or chunk" (p. 381). Next, they ask eight raters whether they think a sequence exhibits "a cohesive meaning or function, as a phrase" (p. 381). Finally, they ask six raters whether they think the sequence "is worth teaching, as a bona fide phrase or expression" (p. 381). Interrater alpha is .77, .67, and .83 for these rating experiments, respectively. The authors then correlate these ratings with the mutual information scores and frequencies of the sequences, finding that, overall, the higher the frequency and mutual information of a sequence, the higher the participant rating. These results suggest the importance of frequency-based measures for cognitive storage of sequences as holistic entities. Interestingly, however, as mentioned in chapter 2, the authors

find that mutual information is a better correlate than frequency, providing further evidence for the primacy of contingency-based measures over raw frequency in modeling distributional learning.

While in this approach intuition is the primary diagnostic, it differs from the role of intuition in some of the above studies in two key ways. First, results are averaged across a number of raters who are not invested in a particular hypothesis. Thus, the possibility for single-rater bias and conflict of interest are controlled. Second, using a rating scale allows for a gradient assessment of how well particular word sequences instantiate formulaicity. In contrast, the researcher-defined and dictionary-based approaches represent a purely Boolean metric of whether something is a MWE.

Given the advantages of a rating methodology as I have just described, in this chapter I employ variants of such an approach to evaluate the output of MERGE across three different experiments. It should be noted, however, that one disadvantage of offline, participant rating-based experiments is the inaccessibility to the researcher of what truly is going on in the minds of the participants. Different participants may interpret rating instructions in wildly different ways, and they may use markedly different evaluation strategies in carrying out the rating procedure. Ultimately, the participant ratings in the experiments reported here show that the algorithm does indeed identify reasonable MWEs, and it is thus a viable approach for the distributional learning of MWEs. But, the nature of the internal mental processes deployed by participants that yield these significant results remains admittedly opaque.

In the next section, I report the first experiment, in which participants assign ratings to output that is learned early versus late in MERGE's iterative progression, under the

hypothesis that early-learned MWEs will be rated more highly. Then I report on two more experiments, which compare the ratings assigned by participants to output from the MERGE algorithm to ratings assigned to output from another algorithm from the literature, the Adjusted Frequency List (Brook O'Donnell 2011). The hypothesis of this comparison is that MERGE-sourced output ought to be rated more highly, indicating a better performance of this algorithm in extracting reasonable MWEs.

4.1 Experiment 1

The first experiment I report on is designed simply to establish that MERGE is capable of finding reasonable MWEs. In the materials section, I discuss the corpus used, the implementation of the algorithm, the preparation of stimuli and selection of participants, as well as the survey instructions and procedures of administration. Then, I turn to the results and the discussion of these results.

4.1.1 Materials

The corpus on which MERGE was run is actually a combination of two corpora: the Santa Barbara Corpus of Spoken American English (SBC; Du Bois, Chafe, Meyer, and Thompson 2000; Du Bois, Chafe, Meyer, Thompson, and Martey 2003; Du Bois and Englebretson 2004; 2005) and the spoken portion of the International Corpus of English—Canada (ICE—Canada; Newman and Columbus 2010). Together, these corpora contain about 700,000 words. This includes about 250,000 words in the SBC across 60 separate discourse events/corpus files, and 450,000 words in ICE—Canada across 300 separate discourse events.

The files in the two corpora span a variety of speech genres, including face-to-face and telephone conversations, academic course lectures/lessons, news programs, religious sermons, parliamentary debates, business meetings, radio programs, and many others. These various genres comprise talk that is both monologic and dialogic, scripted and unscripted. The speakers represent a variety social types, differing in terms of such demographic categories as region, gender, sexuality, race/ethnicity, and age.

It was decided to combine these two corpora because it was desirable to use a spoken corpus of North American English that approached 1 million words (researchers producing collocation-based work typically work with corpora of around 1 million words or more). Curiously, modern spoken-language corpora for North American English are rare, and while these two examples are on their own much smaller than 1 million words, together they approach this target. Furthermore, the reason for opting for a spoken corpus has to do with the fact that one of the primary goals of MERGE is to reflect how human beings might use distributional learning to acquire MWEs, and spoken language is the primary input that people use to acquire linguistic representations.

Finally, it was desirable that the spoken language be North American English because the study participants were to be U.S. university students. While the British National Corpus contains perhaps the largest collection of spoken texts (10 million words) of English, the rather significant differences between British and North American English may decrease the participant-assigned ratings, as British MWEs may not be recognized by North Americans as formulaic. This may especially be true given the fact that research has shown that an important axis along which varieties of the same language vary is in terms of their formulaic language (Gries and Mukherjee 2010). By contrast, while Canadian and

American English do exhibit dialectal differences, these differences are rather minimal, especially when compared to their differences with respect to Old World Englishes.

The two corpora, while transcribed according to different conventions, were preprocessed so that they are formatted equivalently for input to MERGE. This preprocessing included removal of all tags and transcription characters that were not part of the lexical representation of the words themselves. These included features such as markers of overlap, laughter, breathing, incomprehensible syllables, pauses, and other non-lexical vocalizations, among other features. For words that contained non-alphanumeric characters, such as an apostrophe or a hyphen, these characters were replaced by an alphanumeric representation of them. For example, *it's* became *itAPOSs* where the *APOS* component is short for *apostrophe*.

Additionally, while speaker tags were deleted, boundaries between turns-at-talk were retained. In lieu of utterance boundaries, such turn boundaries represented the lowest level type of separation used by MERGE in its representation of the corpus. A major reason for using turns as the lowest level division is because the two corpora employed here use different units below the turn to segment speech: while ICE—Canada is transcribed according to utterance boundaries, the SBC is transcribed according to intonation units. In any case, MWEs can in theory span both multiple utterances and intonation units, but it is unlikely that they would span turn boundaries. Therefore, such turn boundaries provide a more natural segmentation across which MWEs will not be tracked.

Following corpus pre-processing, the dataset was entered into MERGE and the algorithm was implemented using a maximum gap size threshold of 1. That is, the algorithm

can acquire MWEs with one or more gaps within them, provided that these gaps are no longer than one word long. The algorithm was run for 20,000 iterations.

Next, output MWEs were selected for use as stimuli in the rating experiment. These included the first 40 and last 40 merged items for each size of MWE in terms of the number of words that they contained, from MWEs of two words to MWEs of five words. While the model did extract sequences of 6 or more words, these were relatively few in number, so a maximum size threshold of 5 words was chosen. Thus, 320 total MWEs were selected, with half belonging to an *early* bin and half to a *late* bin.

Four different versions of the rating survey were then created, each containing 80 MWEs. These included 10 two-word MWEs from the early bin, 10 two-word MWEs from the late bin, 10 three-word MWEs from the early bin, and so forth. Each group of 10 words was selected at random from all the MWEs that exhibited the same bin identity and were of the same size. Twenty surveys were then created, including 5 of each version. Thus, each stimulus MWE was to be rated by five participants. The order of presentation of stimulus items for each survey was randomized. Each stimulus item was also accompanied by an utterance sourced from the corpus containing that stimulus item, so that study participants had a sense of the use of the candidate MWE in context.

One of the most important components of the survey—and one with crucial theoretical implications—is the instructions. In the Ellis et al. (2008) study cited above, study participants were asked to assign three different ratings according to notions such as formulaicity, cohesiveness of meaning/function, and pedagogical merit of the stimulus items. While these notions get at important facets of formulaic language, the first and last ones likely require a degree of sophisticated linguistic knowledge that most undergraduates

in introductory linguistics courses do not possess. Remember that Ellis et al. (2008)'s participants are trained pedagogues, and thus they already had an idea of what formulaic language is, and whether it is useful in language instruction.

Instead, the goal here is to tap into intuition about formulaic language that non-specialists may have, even if they do not have formal knowledge of issues pertaining to formulaicity. The taxonomy of criteria that Gries (2008) identifies as typically employed in different definitions of phraseologisms (discussed in chapter 2) suggests a number of typical features of formulaic language about which non-specialists may indeed have intuitions. For example, it is plausible that undergraduates without formal training might have some intuitions about the frequency of word sequences, whether or not they represent a meaning that is not predictable based on their parts, or whether they represent a complete phrase.

As we have seen, though, not all MWEs are frequent (such as idioms), not all MWEs are semantically non-compositional, and some MWEs may represent incomplete syntactic phrases, or they may cross syntactic phrase boundaries. Thus, explicitly targeting any one of these specific attributes may exclude perfectly reasonable MWEs. Perhaps because of the multi-dimensional nature of formulas, Ellis et al. (2008) chose to have participants assign three separate ratings, each on the basis of different criteria. Again, however, such criterial distinctions may overwhelm non-specialists; moreover, the sheer number of stimulus items to be used here might result in fatigue if participants were asked to assign multiple ratings for each one.

Remember that uniting all definitions of phraseology/formulaicity is the notion that these word sequences are retrieved whole from memory. For this reason, an instructional design was chosen to try to tap into participant intuition about the degree to which they have

target sequences memorized. To this end, participants were instructed to assign ratings, on a scale of 1 to 7, indicating the degree to which a particular stimulus item represented a *common, reusable chunk*. An item assigned a 1 would be considered a poor representation of a common, reusable chunk, while an item assigned a 7 would be considered an excellent representation. The instructions were supplemented with both good and bad examples of such complete, reusable chunks, based on the opinion of the researcher. These examples were sourced from the MERGE output, and were not included as stimulus items. The full instructions provided to participants are reproduced in Appendix A.

The notion of *reusability* is used here as a proxy for memorization. It was felt that such an explicitly cognitive notion as memorization might be less tractable to intuition than the notion of reusability. Furthermore, the notion of *commonness* was selected in order to gesture toward the idea of frequency. While frequency may sometimes exclude perfectly good low frequency idioms, it is nonetheless an important correlate of formulaicity. It is not known whether instructions to assign high ratings to common chunks will cause low frequency yet high salience/contingency items, which may be good MWEs, to be rated poorly.

At the following stage, twenty participants were recruited from introductory linguistics courses at the University of California, Santa Barbara. Each participant was placed in a quiet room by themselves and given as much time as they needed to complete the survey. Only data from native English speakers were used in the final analysis.

4.1.2 Results

All statistical tests were run and plots produced using the programming language R. The participant ratings were first evaluated by way of a linear mixed effects model using the lmertest package in R. The dependent, or predicted, variable was the scores assigned by the study participants. These scores, which again range from 1 through 7, in principle represent an ordinal variable, but the linear model treats them as ratio-scaled.

There were a number of independent variables, or predictors. These include two fixed effects: binned rank order and gram size. As described above, the binned rank order is a categorical variable with two levels—early and late; the former level refers to the first 40 merged items of each gram size and the latter level refers to the last 40 merged items of each gram size (out of the 20,000 total merges for which the model was run). Gram size is an ordinal variable that refers to the number of words contained in the test items, with a range of 2 through 5 (since, as discussed above, these were the gram sizes of the chosen test items). In addition, a polynomial curve to the 2^{nd} power was fitted to this variable, since visual inspection revealed an apparent curvature in the data in the gram size-by-score plane. The interaction of the two predictors was also included in the model.

Besides the fixed effects, random effect predictors were included as well. These include both random slopes and intercepts for gram type as well as participant. That is, separate regression lines were fit for each gram type and participant, and then the overall regression was adjusted so that it reflected the central tendency of these individual regression lines (assuming that a normal distribution describes these individual regression lines). The model summary statistics are provided in Appendix B.

For the fixed effects, all of the main effects and interactions are significant, with p-values < 0.05. Note that, for the polynomial curve, the model tested all powers up to and including the power that was specified in the model—here, a polynomial of the second power. Thus, the model shows significance for both first and second power polynomials, fit over the gram size variable, both as main effects and in interactions with binned rank order.

Because mixed effects modeling does not return traditional statistics of model fit such as R^2, this value must be calculated separately. The r.squaredGLMM() function was used (Nakagawa and Schielzeth 2013; Johnson 2014; Barton 2015), which is part of the MuMIn package. The present model exhibits a marginal R^2 of 0.64 and a conditional R^2 of 0.84. The former value is most relevant, since it suggests the degree of fit that might be expected in the general population (i.e., not specific to this data set). The value of 0.64 represents a high degree of fit, indicating that the regression models the data quite well.

Figure 4.1 is a plot of the fixed effects, and it contains both the regression lines and individual data points. The line and data points in light gray correspond to the early bin and the line and data points in black correspond to the late bin. The horizontal axis depicts the gram sizes (1 through 5), while the vertical axis depicts participant-assigned scores (1 through 7). First, note that the regression line for the early bin is located entirely above the regression line for the late bin, with no overlap in the range of the lightly colored confidence intervals surrounding the two lines. This indicates that early-merged items are rated more highly overall than late-merged items. This represents the hypothesis and is the most important finding of this experiment.

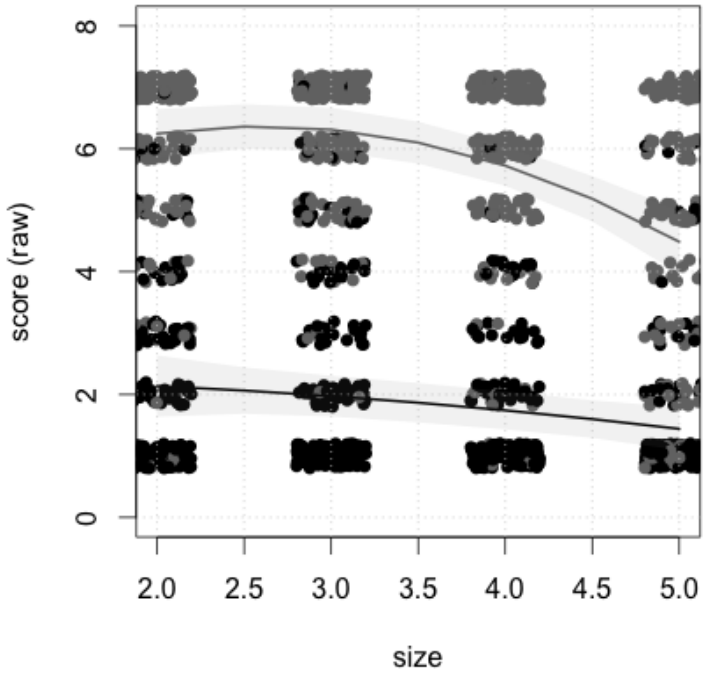
In addition, however, there is an interesting interaction. Note, for the early (light gray) bin observations, the greater skewing of points towards lower scores for the larger

gram sizes compared to the smaller gram sizes (where the points are concentrated towards higher scores). Put simply, smaller grams are ranked higher than larger grams. Moreover, the light gray polynomial curve allows us to see that the rate at which grams are rated worse significantly increases in the negative direction as gram size increases.

In contrast, there is no such effect within the late bin. That is, the black observations do not show a clear difference in vertical skewing across different gram sizes. While the black regression line superimposed on these points does appear to exhibit a negative slope, the fact that a possible regression line with a slope of zero falls within the shaded confidence interval indicates that we cannot be confident in the apparent negative slope. Thus, for grams merged late, they are overall all rated as being equally poor MWEs, irrespective of size.

It is not clear why this interaction exists, but one possibility has to do with the different distributions of rank orders across gram sizes. Among the earlier grams that the algorithm merges, most of the output items are bigrams, followed by trigrams, and so forth. Thus, the top 40 bigrams have a lower average absolute rank order than the top 40 5-grams, since one must search lower in the absolute rank order to amass 40 5-grams than to amass 40 bigrams. Thus, in theory, the top 40 bigrams are 'better' MWEs than the top 40 5-grams. In contrast, for the bottom 40 grams for each gram size, size does not matter—they are all 'bad' MWEs.

Figure 4.1. Effects Plot[6]

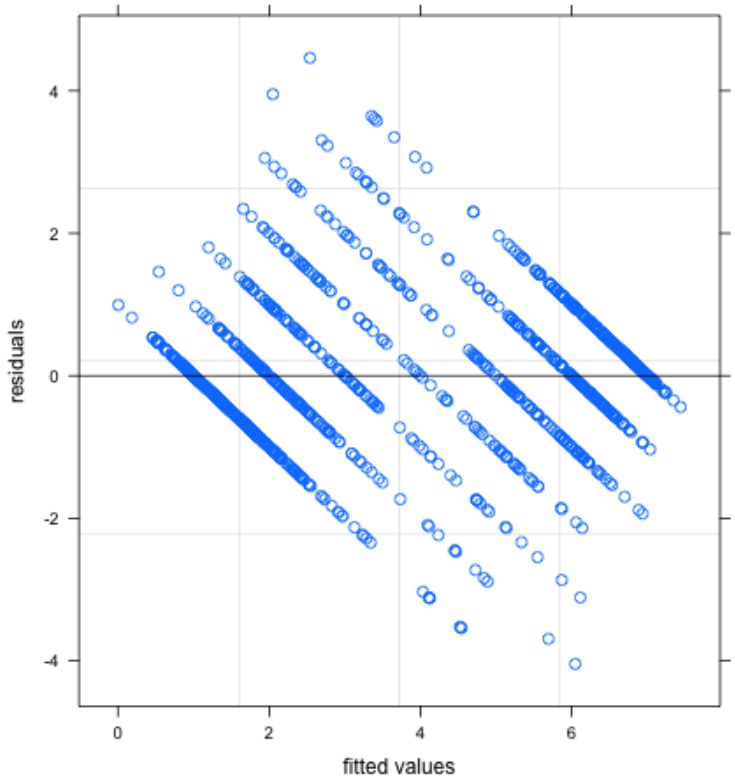


In Figure 4.2, the residuals are plotted against the fitted values. The important observation to make about this plot is that there are numerous data points where participants assigned either high (6's and 7's) or low (1's and 2's) scores, but relatively few scores from the middle of the score range. As a result of this uneven distribution, the model regressed a line that provides a closer fit to these extreme scores than to the mid-range scores; this is why the residuals at these extremes are closer to 0.

This uneven distribution of scores is likely a result of the fact that the test stimuli were selected from extremes of the order in which items were merged by the algorithm.

[6] The R code for this plot was authored by Stefan Th. Gries and provided to me via personal communication.

Figure 4.2. Plot of Residuals as Function of Fitted Values



Items were either from the top 40 or bottom 40 merges of a particular gram size (out of the first 20,000 merges). In other words, there were no grams selected from mid-way through the 20,000 merges. Had grams been selected evenly from across the 20,000 merges, a more uniform distribution of residuals might be observed in Figure 4.2.

However, such an evenly-distributed selection of test items was not pursued because, in pilot testing, it seemed that there was a tremendous amount of variance in the scores that participants assigned to test items that were close in rank order. It was felt, therefore, that R^2 might very well end up being quite low, if there was a correlation at all. To compensate for this, test items were selected that would likely be rated to be quite good (the very earliest

merge items) or quite bad (items from up to 20,000 merges later). That is, if the hypothesized effect were to be present, this design would surely find it.

In summary, we know that for the bin rank variable, ‘early’ corresponds to higher scores, as hypothesized. Furthermore, this is true across all gram sizes. At the same time, because the model is set up in such a way that higher scores for the early bin are only seen in the context of an interaction with the gram size variable, we do not know how strong the effect of early score > late score is on its own. To ascertain this, we need to remove the effect of gram size and look at the independent effect of binned rank order on participant score.

This may be accomplished through a process known as residualization. First, a linear model predicting participant score as a function of gram size was fit. As above, this model included participant and gram type as random slopes and intercepts. Note, however, that this model did not include binned rank order.

Next, the model was used to generate the predicted scores given the observed gram sizes. Crucially, the random effects adjustments were not included in the predictions. (That is, while the random effects were used in the original fitting of the overall regression line to the data, the predicted values that were used lie along this overall regression line—they were not adjusted upward or downward along the dependent axis so that they fell along the individual participant and gram type regression lines.)

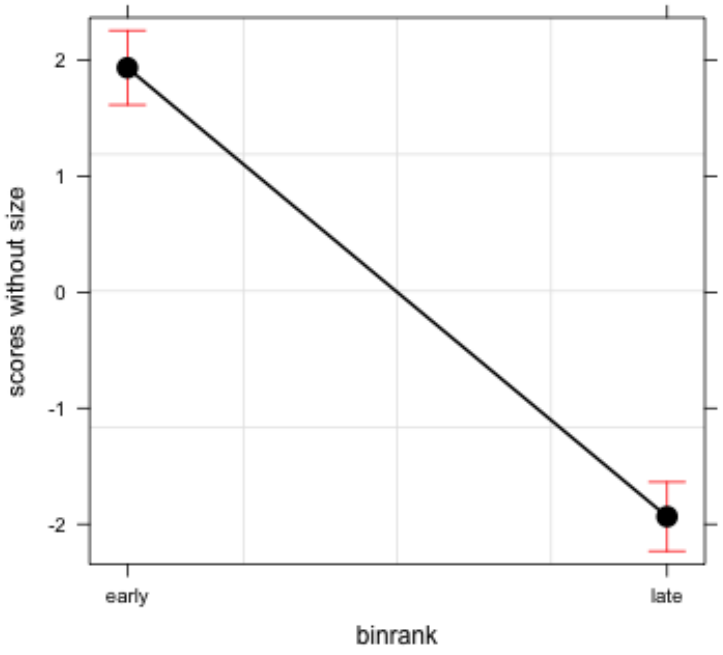
At the next step, the residuals were calculated by taking the difference between the predicted and the observed scores. Importantly, the residuals represent variance not accounted for by the above model of score as a function of gram size. We therefore want to see if there is structure leftover in this data that can be predicted by binned rank order. The

amount of leftover structure that can be predicted by binned rank order is equal to the independent effect of binned rank order on scores. Accordingly, another linear model was then fit to these residuals, with binned rank order as the predictor and participant and gram type again as random intercepts and slopes.

The summary statistics for this model are provided in Appendix C. With a p-value of less than 2×10^{-16}, the binned rank order fixed effect is highly significant. Furthermore, calculating R^2 as described above, we see a marginal R^2 of 0.62 and a conditional R^2 of 0.83. The marginal R^2, which tells us the correlation that we might expect independent of the current data set, is quite good.

Figure 4.3 depicts the effects plot for the residualized model. Clearly, the residuals for the early bin are higher than the residuals for the late bin, and the fact that the confidence intervals are nowhere near exhibiting any overlap suggests the significance of this difference. This directionality of effect is what we expect to see. In the structure left over after the effect of gram size has been removed, the fact that the residuals for the early bin are positive indicates that the leftover structure overall lies above the regression line. Conversely, the negative residuals for the late bin indicate that the leftover structure overall lies below the regression line. Thus, independent of gram size, early MWEs are rated more highly than late MWEs.

Figure 4.3. Effects Plot for Residualized Model



4.1.3 Interim Discussion

The preceding experiment validated the effectiveness of MERGE as a computational model of statistical learning of MWEs. That is, it showed that the model was capable of learning reasonable MWEs, according to human ratings of such reasonableness. More specifically, 'better' MWEs were identified by MERGE early in its iterative cycles, while 'worse' MWEs were identified later, where merge order was determined by log likelihood-based association strength (discussed in chapter 2). In addition, the sizes of the MWEs identified by the model exhibited different patterns of reasonableness, according to raters. That is, larger early MWEs were considered to be not as good as smaller early MWEs, while, for MWEs identified later in MERGE's iterative cycle, all MWEs were considered bad, regardless of size.

4.2 Experiment 2a

While experiment 1 established that MERGE is a viable model of distributional learning of MWEs, as reviewed in chapter 2, it is not the only computational approach to this problem (e.g., Bannard et al. 2009). The most directly comparable approaches to MERGE are the extraction and ranking approaches described in section 2.3.2.2. As discussed, these approaches operate in a batch manner and generate a ranked list of MWEs found in a corpus, whose lengths are not pre-specified but rather determined in a bottom-up fashion. One way in which MERGE differs from these approaches is that it is envisioned to be usable for both cognitive and non-cognitive research goals. However, this is primarily a question of theoretical framing. While the corpus linguists who have developed LocalMax (Da Silva et al. 1999) and the Adjusted Frequency List (AFL; Brook O'Donnell 2011), for example, did not have cognitive goals in mind, the learned multi-word representations of these approaches—as well as the distributional mechanisms the learning engenders—could end up being compatible with human-like cognitive representations.

Thus, the question remains as to whether MERGE offers something additional compared to these existing approaches. More to the point, is the performance of MERGE better than that of LocalMax and/or the Adjusted Frequency List? The second experiment explores this question using another rating experiment similar in design to the first one.[7]

[7] As noted in a footnote at the end of chapter 2, there are certain important similarities in architectures between MERGE and the algorithm proposed in Wible et al. (2006): namely, they both grow MWEs of various lengths in a bottom-up fashion by recursively combining bigrams. However, as I also noted, Wible and colleagues' algorithm generates MWEs for a single user-specified node word, while MERGE, LocalMax, and the AFL generate global lists of MWEs for a particular corpus. Because of this crucial difference, I do not include Wible and colleagues' algorithm in this comparison.

4.2.1 Materials

The corpora used in experiment 1 were also used here, with the same preprocessing procedures. Next, LocalMax (Da Silva et al. 1999) and the AFL (Brook O'Donnell 2011) were implemented in Python. These implementations were of my own design, based on the descriptions of the algorithms from the original research papers. I then ran the two algorithms as well as MERGE on the corpus (that is, the combination of the SBC and ICE—Canada). In all three algorithms, only continuous word sequences were permitted.

In addition, in the case of LocalMax, the researcher must select beforehand the largest gram size to be considered. This parameter was set to 5. Moreover, this algorithm may be implemented with different metrics of word sequence strength. The authors propose an approach—Symmetrical Conditional Probability (SCP)—that they say is the best performer of several different measures that they try (Da Silva et al. 1999). For this reason, I implement the algorithm using the SCP measure.

In the case of the AFL, the user must indicate a minimum frequency that a word sequence must exhibit in order to be considered an output sequence. Following Brook O'Donnell (2011), this parameter was set to 3.

Finally, in the case of MERGE, the algorithm was run for 1000 iterations.

Next, model output was visually inspected. Note that Da Silva et al. (1999) claim that 81% of the items extracted by LocalMax from the Portuguese news corpus that they use represent true formulaic sequences. (See introduction to this chapter for discussion of criteria for formulaicity used by the authors). However, initial inspection of the output of the LocalMax algorithm as implemented here revealed this output to be extremely poor—that is, the model does not appear effective at all at acquiring MWEs. Why might there be this

discrepancy between the findings of the original authors and the quality of the output generated here? Unfortunately, Da Silva et al. (1999) only provide a small sample of output items in an appendix, and it is not clear whether this list is the result of random sampling. Thus, it is impossible to use the original output to investigate how the authors calculated their precision value of 81%.

Despite the authors' claims regarding the efficacy of LocalMax, it seems that the way in which the SCP measure assigns scores may be responsible for the poor performance of the model, as observed here. High scores are assigned to word sequences whose number of occurrences represents a high proportion of the overall occurrences of the sequences' individual words. Thus, if a sequence occurs 100 times, and each component word likewise occurs 100 times, then the sequence will be assigned a maximal SCP score. However, this same proportionality can occur for extremely low frequency sequences and words as well. If the component words of a sequence are all hapaxes, the sequence itself must also be a hapax. In this case, the sequence represents 100% of the occurrences of the component words, and it will receive a maximal SCP score. Clearly, however, such hapax sequences are often very poor candidates for formulaic sequences, yet these kinds of sequences appear to constitute a large amount of the output of my implementation of LocalMax.

In contrast to this poor performance, the performance of both MERGE and the AFL did appear potentially strong, so they were included in the analysis while LocalMax was not.

The next step was the generation of output stimuli. The top 1000-ranked items from the AFL were compared to the output from the 1000 MERGE iterations. Two groups of items were then created: the first group comprised those items found in the AFL output but

not in the MERGE output; the second group comprised those items found in the MERGE output but not in the AFL output.

Thus, it was decided to focus on the MWEs that the two algorithms did not agree on rather than the MWEs that they had in common in their output. This allowed a highly tractable examination of how the respective performances of the two algorithms contrasted, as stimulus items fell into one of two categories. Conversely, there would have been difficulties in comparing the performance of the algorithms on the basis of the output that they had in common (i.e., by seeing which algorithm's ranking of output best correlated with participant-assigned scores to this output). Since the strength metrics used to rank output were different for each model, the algorithm-assigned strength values would have to have been rank ordered to make them comparable across algorithms. But the fact that the AFL is based on integer frequency means that there are numerous ties, whereas the log likelihood decimal values used by MERGE make for virtually no ties (at least at higher scores). Thus, the rank order distributions of the two model outputs were intractably different.

From the two groups of disjunctive output, 180 items were then randomly sampled. An even distribution of sampling from across the range of items was achieved by partitioning the two rank-ordered item groups into 10 bins and then randomly sampling 18 items from each bin. Next, groups of stimuli for the surveys were created, with each group containing 45 items sampled randomly without replacement from each of the two groups of 180 items above. Thus, each survey contained 90 items—45 generated by MERGE and 45 generated by the AFL.

In Table 4.1, I provide a random sampling of 20 stimuli sourced from the 180 AFL items and 20 sourced from the 180 MERGE items. One can immediately appreciate the

qualitative difference between many of the items in these two lists. While the high-frequency sequences represented in the AFL output comprise many combinations of function words, the MERGE output comprises many sequences combining function and content words. The combinations include structures such as noun phrases (*a good idea*), compound nouns (*square root*), compound prepositions (*in the middle of*), whole utterances (*thanks very much*), and phrasal verbs (*to make sure*), among others. Furthermore, while these combinations may be lower overall in frequency, their component words are mutually contingent. This type of relationship of mutual contingency is precisely the statistical pattern that lexical association measures like log likelihood are designed to capture.

Table 4.1. Random Sampling of Output from AFL and MERGE

AFL	MERGE
he is	auto reverse
and just	in the middle of
but if you	we need
because the	to make sure
and this	square root
and I think	I want you
well it	you think
they all	kind of thing
and how	let us
to their	major depression
of it	good afternoon
a real	Melissa Soligo
it the	they weren't
get a	must have been
before the	next week
what kind of	a good idea
says the	I wanted to
with that	we'll see
there and	thanks very much
so this	a great

Note that the sequences comprising hapax words that proved so problematic for the SCP measure used in the LocalMax approach likewise exemplify the case of high mutual contingency of component words. Log likelihood scores do not share this problem, however, because they are also modulated by overall sequence frequency, and are not strictly a metric of component word contingency.

At the next step, as in experiment 1, 20 surveys were created, including 5 of each version, each to be rated by a single participant. Again, the order of presentation of stimulus items for each survey was randomized, and each stimulus item was accompanied by an utterance sourced from the corpus containing that stimulus item, so that study participants had a sense of the use of the candidate MWEs in context.

Again as in experiment 1, 20 participants were recruited from an introductory linguistics course at the University of California, Santa Barbara. Each participant was placed in a quiet room by themselves and given as much time as they needed to complete the survey. Data only from native English speakers was used in the final analysis.

4.2.2 Results

Again, the statistical analysis and production of plots were conducted using R. First, I provide a visualization of the distributions of scores assigned by study participants to the output from the two algorithms (Figures 4.4 and 4.5). Note that the AFL output exhibits a rather flat distribution. In contrast, MERGE exhibits a higher number of scores of 7 and 1, while the middle-range scores are slightly lower than the middle-range scores of the AFL-sourced items. At the same time, the shapes of the distributions of the middle range scores are nearly identical.

Figure 4.4. Score Distribution for MERGE

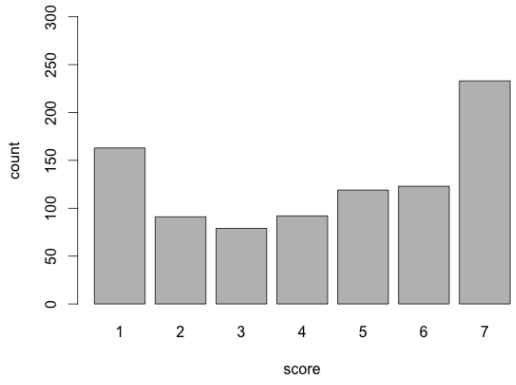
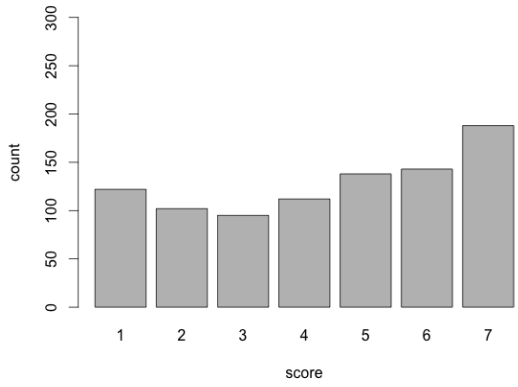


Figure 4.5. Score Distribution for AFL



The question we wish to answer is whether there is a significant difference between the scores of the outputs from the two algorithms. Specifically, the prediction is that the scores for the MERGE-sourced items should be higher overall than those for the AFL-sourced items, given the hypothesis that MERGE does a better job of learning MWEs.

To examine this, I first used the Kolmogorov-Smirnov (KS) statistical test, which reports whether two distributions of data points are significantly different from each other. More specifically, the KS test compares the empirical cumulative distribution functions (ECDFs) of two distributions, checking whether one ECDF is significantly higher than the other (see Gries 2013, pp. 173 – 178). Here, the prediction was that the ECDF of the scores of AFL-sourced items should be higher than that of the MERGE-sourced items, which would indicated that the AFL scores are more skewed to the left (lower) side of the domain. One must note, however, that the one assumption of this test is that the data be continuous; because the data here are ordinal (1 – 7), the use of this test is not strictly sanctioned and

thus its results must only be interpreted as a first-pass heuristic look at distributional differences between scores of the AFL- and MERGE-sourced items.

The KS test revealed that the ECDF function of the AFL items is non-significantly higher than that of the MERGE items, with a p-value of 0.11 and a critical D-value of 0.05, which indicates the maximal vertical distance between the two ECDF functions. Thus, the distribution of scores for AFL-sourced items is not more skewed to the lower score range than the scores from the MERGE-sourced items, at least according to this heuristic test (for the purposes of comparison, a KS test in the other direction revealed that the ECDF of the MERGE items is non-significantly higher than that of the AFL items [critical D = 0.05, p-value = 0.15]).

Another statistical approach I applied to the data from experiment 2a was linear mixed effects regression, which was also used in experiment 1. The scores assigned by participants represented the predicted variable, while the origin of the test items—either the output of the AFL or MERGE algorithms—represented the (categorical) fixed effect predictor variable. Furthermore, random slopes and intercepts were included for participants, while random intercepts were included for gram type.

Given the null results under the KS test, one might anticipate that this regression model would not perform well. And indeed it did not: the p-value was highly non-significant at 0.954, and the marginal R^2 value was 5.31×10^{-6}. However, the conditional R^2 value was .576, which is actually quite good. The difference in these two R^2 values suggests that, while overall model fit was poor (indexed by marginal R^2), the random effects structure was doing a lot of the predictive work (indexed by conditional R^2). Thus, my primary focus in using linear modeling here was to explore the random effects structure, and in particular the

random intercepts for gram type (because of this narrower focus, I do not reproduce the full summary statistics of the model, as in the cases of the linear models elsewhere in this chapter).

In Figures 4.6 and 4.7, I provide histograms of the random intercepts for the individual gram types. Each intercept itself represents a central tendency among the different scores assigned to a particular gram type by the five different participants who rated it. The vertical bars in the histogram represent a bin of intercept values in increments of 0.5, and the vertical axis indicates the number of intercepts that fall within each bin. Note that the peak of the MERGE distribution falls on a higher score range (5.0-5.5) than the peak of the AFL distribution (4.5 – 5.0). This represents a skewing of scores in the predicted direction.

Figure 4.6. Histogram of Random Intercepts (Scores) for MERGE Items

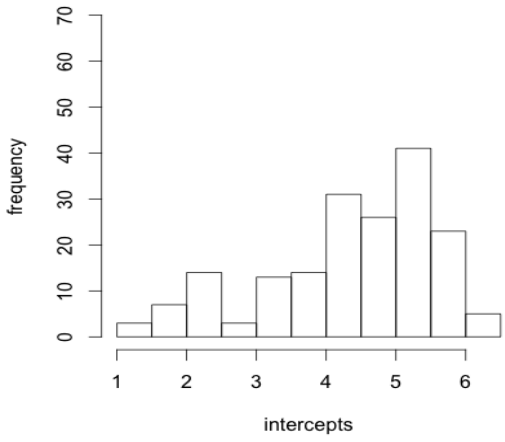
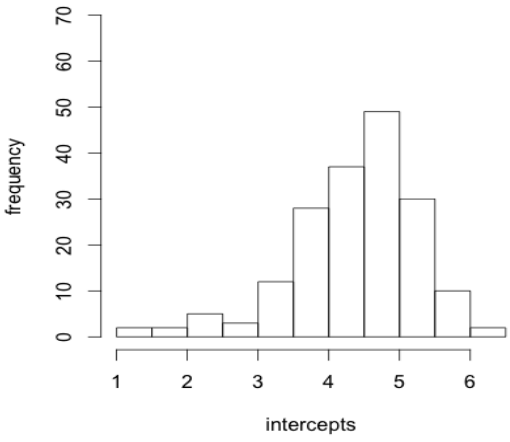


Figure 4.7. Histogram of Random Intercepts (Scores) for AFL Items



However, also note that there is a noticeably large tail of low score-range intercepts in the MERGE distribution that is lacking in the AFL distribution. In order to investigate this tail further, I have listed in Table 4.2 the gram types from MERGE whose random intercepts were less than 4. These items are eclectic, but there are some patterns in what the table contains: there are a number of proper names (e.g., *Clara Brett, Holy Spirit, New York*), compound nouns (e.g., *square root, coat check, world war*), and phrasal verbs (e.g., *know how, ensure that, worry about*), among other items.

Table 4.2. MERGE-Sourced Gram Types with Random Intercepts < 4

"Mr. Grey-tootoo"	"auto reverse"	"two years"
"drill splash"	"in across the line"	"three or four"
"centre ice"	"referendum law"	"know how"
"Clara Brett"	"Golden Bears"	"San Francisco"
"crown must prove"	"down the ice"	"his name"
"Daniel Johnson"	"British Columbia"	"pick it up"
"Melissa Soligo"	"Kurt Browning"	"on the floor"
"you're listening to the Chevron Ecofile"	"my client"	"ensure that"
	"square root"	"major depression"
"Young Offenders Act"	"it comes"	"do that"
"the Honourable Member"	"coat check"	"let us"
"power play"	"Holy Spirit"	"next week"
"Annie Gillis"	"hundred and"	"put it"
"Aaron Zarowny"	"you're listening to the"	"New York"
"annual allowable"	"two thousand"	"worry about"
"Madame Speaker"	"world war"	
"Raoul Eto"	"you're listening"	
"that's good said tiger"	"six months"	
"the Honourable Member for"	"three months"	
	"square root of"	

4.2.3 Interim discussion

This experiment did not provide significant statistical evidence for the alternative hypothesis: that the scores of items uniquely identified by MERGE, taken from the top 1000 MERGE items, were higher than the scores of items uniquely identified by the AFL (taken from the top 1000 AFL items). One thus may wish to conclude from these results that there is no significant difference in the quality of the MWEs that the two approaches identify. However, such may not be the case. The visual inspection in Table 4.2 of the MERGE items that received the lowest overall scores, measured in terms of their random intercepts, shows that many of these items represent lexical phenomena that are plausible MWEs (such as proper names, compound nouns, and phrasal verbs). Moreover, also note that they are precisely the types of sequences that made the sampled MERGE-sourced items in Table 4.1 qualitatively seem so different from AFL-sourced items. Thus, it appears that many of the kinds of sequences that give the MERGE output its distinctive character were evaluated as poor examples of MWEs. Had these grams received higher scores, the output of MERGE may have overall been rated significantly higher than that of the AFL, as hypothesized.

Why, then, did this output not receive higher scores? A likely culprit is the instructions. Naturally, the goal of the instructions is to tap into participant intuitions of formulaicity, where a formulaic sequence is conceived of as a holistic item retrieved from memory, rather than constructed online. In order to draw on these intuitions, I have used the wording *common, usable chunk* up until this point in the instructions. However, it is possible that the term *common* may have compelled participants to overly prioritize evaluations of sequence frequency in their assignment of scores. This would explain why many of the items in Table 4.2 are ostensibly good MWEs and yet were scored low: such sequences are

low frequency, but at the same time their component words are rather highly mutually contingent. (Remember that this is a pattern of distributional behavior that lexical association measures are designed to capture, and which raw frequency is not).

Another possible failure of the instructions includes the use of the term *reusable*, which was designed to get participants to tap into intuitions about formulaicity/memorization of sequences. However, the connections between these concepts and the term *reusable* may have been too opaque.

Finally, the use of the term *chunk* was intended to guide participants to prefer sequences that represented complete MWEs. Again, however, the connection between the term *chunk* and the notion of completeness may not have been sufficiently clear.

4.3 Experiment 2b

In the previous experiment, the notion of a *common, reusable chunk* was used as the central component of the instructions to tap into participant intuition about formulaicity (primarily, whether a sequence is retrieved from memory). In this experiment, I instead use the central notion of a *complete unit of vocabulary* in an attempt to better tap into this same targeted intuition. In the next section, I detail the revisions to the instructions. Then, I provide and discuss results for experiment 2b. Finally, I turn to a general discussion and conclusions covering the three experiments in this chapter.

4.3.1 Materials

With the exception of the changes to the instructions, the materials used in experiment 2b are identical to those in experiment 2a. Thus, in this section, I focus on these

instructional changes. The new instructions are provided in Appendix D. In order to access participants' intuitions about word sequences they have memorized, it was decided to focus on the notion of one's *vocabulary*. This is a non-specialist term that average native speakers of English ought to be quite familiar with, and it can serve as a proxy for the specialist notion of a lexicon. Indeed, anyone who has gone through English language-based schooling will have experience with the idea of a vocabulary as a list of items subject to memorization.

However, it may typically be the case that these non-specialists will conceive of a vocabulary as comprising individual words. For this reason, the instructions explicitly direct readers to consider a list of multi-word sequences that are obviously formulaic, with the intention being to enable readers to discover for themselves that memorized vocabularies can include items larger than single words.

In addition, in the earlier instructions, the idea that items under consideration must represent complete units in order to be scored highly remained a rather hidden point; here, it is included as part of the underlined, primary focal point of the instructions, *complete unit of vocabulary*. This greater emphasis was given because, under the current theory, MWEs represent holistic items in memory, and they are retrieved as such. Accordingly, test items that lack one or more elements necessary to make them complete do not fit this criterion.

Also note, as in the previous instructions, both good and bad example items are given, and the sequences used as such examples are not included as test items in the surveys themselves.

Finally, the notion of frequency as a criterion for rating a sequence highly (suggested via the word *common* in the earlier instructions) has been removed. While its inclusion in these earlier instructions was justified on the basis of the fact that frequency is a strong

correlate of formulaicity, it is not a requirement for formulaicity; as I have stated elsewhere, low contingency, high frequency items can be highly formulaic. Thus, the inclusion of a requirement that items be frequent to be scored highly was deemed a tangential distraction from the true intuition being targeted: whether items are memorized wholes. (One should note, though, that this removal of a reference to frequency could negatively affect the scores assigned to AFL output while boosting the scores assigned to MERGE output, since the former contains only high frequency sequences while the latter may contain the very kinds of low frequency, high contingency items that would benefit from this change to the instructions).

4.3.2 Results

For experiment 2b, the barplots for the distributions of scores for the AFL-sourced items and the MERGE-sourced items are represented in Figures 4.8 and 4.9, respectively. Immediately, one can notice a difference compared to the barplots from experiment 2a in Figures 4.4 and 4.5. Now, the scores for the MERGE items do indeed appear skewed towards the higher end of the domain, while those for the AFL items appear rather evenly distributed (perhaps with a slight skewing towards the lower end of the domain). That is, it appears that AFL-sourced items are roughly equally likely to be assigned any score, suggesting that AFL is not finding formulaic structure in terms of the new instructions, at least for those items not shared with the MERGE output.

Figure 4.8. Score Distribution for MERGE

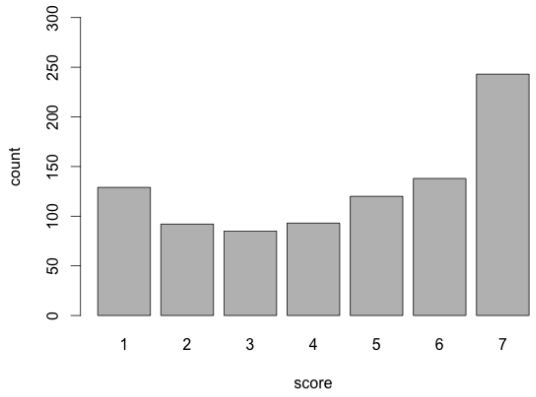
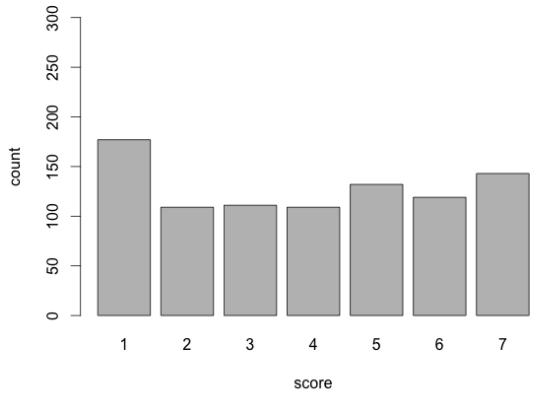


Figure 4.9. Score Distribution for AFL



To test these appearances empirically, we turn again to the KS test. Remember that in the previous experiment this test revealed that the empirical cumulative distribution function (ECDF) of the AFL scores was non-significantly higher than the ECDF of the MERGE scores (whereby the ECDF is used to evaluate differences in the score distributions). Now, however, the KS test did reveal a highly significant difference, with a critical D value of 0.13 and a p-value of 1.46×10^{-7}. What this suggests is that the scores for the AFL items are significantly more skewed to the left (=lower scores) than those for the MERGE items. Again, however, the fact that the data are ordinal rather than continuous means that an assumption of the KS test has been violated, and thus these results can only be taken as a suggestive heuristic.

As in the previous two experiments, a linear mixed effects model was also fit to the data, with participant scores as the predicted variable and the origin of the grams as the fixed effect predictor. Both participant and gram type were included as random effects. Again, as

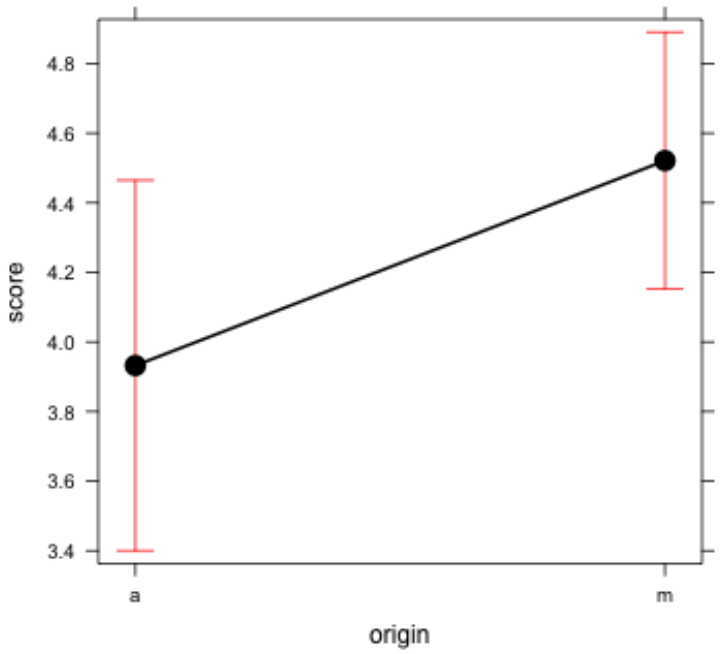
in experiment 2a, both random slopes and intercepts were included for the participant variable. Only random intercepts were used for the gram type variable.

The summary statistics for the model are provided in Appendix E. With a p-value of 0.03, the origin variable is significant, indicating that an effect has been found. However, the overall model fit (calculated as above) is again very poor, with marginal R^2 equaling 0.02. Again, however, conditional R^2 is good, at 0.37. Thus, although model fit is significant, there is a tremendous amount of variance unexplained by the overall regression, although the random effects structure does fit the data fairly well.

Nonetheless, the fact that the predicted score of the MERGE output is higher than the predicted value of the AFL output is the important finding. Despite rather wide confidence intervals, this is the relationship that is depicted in the model effect plot, provided in Figure 4.10.

The fact that this experimental design is so coarse—predicting scores only on the basis of the algorithm from which the output was generated—may be responsible for the overall poor fit. One might conjecture that, had a predictor been included that took into account the order in which the grams were generated by each algorithm, a better model fit would have been achieved. However, as I previously mentioned, the fact that the AFL ranks output on the basis of integer frequency while MERGE ranks output on the basis of decimal log likelihood scores means that the distributions of these two rank orders would be incomparably different, as the AFL rank order would exhibit numerous ties while the MERGE rank order would not.

Figure 4.10. Effect Plot



To examine the random effects again in more detail, I provide in Figures 4.11 and 4.12 histograms of the random intercepts for each gram type, with the units of the random intercepts being participant-assigned scores. Note that the peak of the MERGE distribution lies over a higher score range (4.0 – 4.5) than the peak of the AFL distribution (3.5 – 4.0). Most importantly, note that the sizable tail of low-end intercepts that was present in the histogram of the MERGE intercepts in experiment 2a is no longer present.

Figure 4.11. Histogram of Random Intercepts (Scores) for MERGE Items

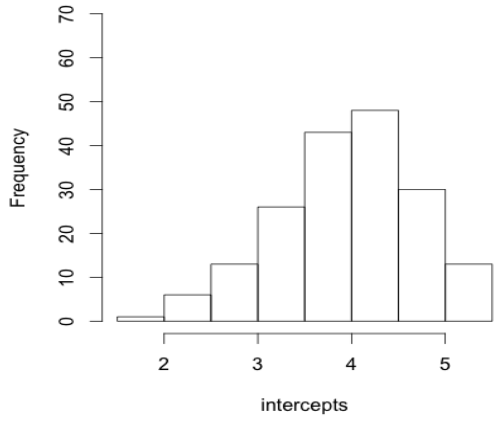
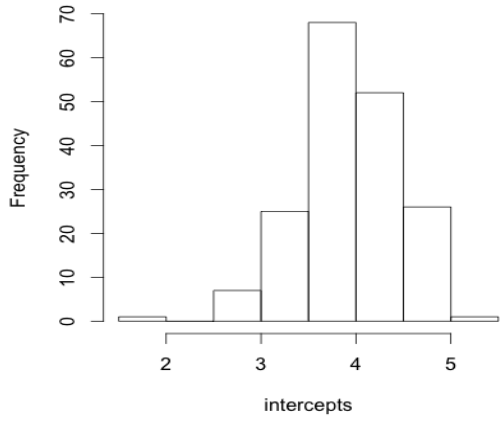


Figure 4.12. Histogram of Random Intercepts (Scores) for AFL Items



An important question, then, is what has happened to intercepts of these gram types that previously corresponded to such low intercept values? Table 4.3 includes a list of the 20 gram types that exhibit the largest increase in their intercept values from experiment 2a to experiment 2b. All of the gram types in the list exhibited among the lowest intercept values in experiment 2a (i.e., less than 4). This should not be highly surprising, since in order for a gram type to exhibit a large increase in its intercept value from one experiment to another, the intercept value must have started out low (as the domain is bounded by 1 and 7).

However, note that the MERGE-sourced items with large intercept increases are precisely the types of sequences that stood out as ostensibly good MWEs despite receiving low intercepts in the experiment 2a regression. This suggests that the modifications made to the instructions in experiment 2b had the intended effects: to boost the scores that participants assigned to the low frequency, high contingency sequences that exhibit formulaic properties in spite of their not being *common*.

Table 4.3. Gram Types Exhibiting Largest Increases in Intercepts

Gram Type	Origin	Intercept Increase
centre ice	m	2.14
square root	m	2.00
British Columbia	m	1.99
square root of	m	1.61
world war	m	1.58
it it's	a	1.55
power play	m	1.51
are are	a	1.48
it it	a	1.48
about what	a	1.42
the Honourable Member for	m	1.31
auto reverse	m	1.15
well you	a	1.13
Clara Brett	m	1.09
Melissa Soligo	m	1.09
Golden Bears	m	1.09
New York	m	1.06
referendum law	m	1.04
Holy Spirit	m	1.03
ah a	a	1.01

4.3.3. Interim discussion

Given the modified instructions in experiment 2b, the results now strongly show that the unique items from the top 1000 output of MERGE are rated as better MWEs than the unique items from the top 1000 output of the AFL. This was demonstrated through a Kolmogorov-Smirnov test, comparing the score distributions for the outputs from the two

algorithms, as well as through a linear mixed effects regression. And although the poor fit of the model showed that it could not predict a score given the source algorithm with high accuracy, the fact that the model was significant showed that there is nonetheless a difference in the scores predicted given the source algorithm. Again, specifically, the predicted MERGE score was higher than the predicted AFL score. Thus, experiment 2b provided strong evidence that MERGE did a better job of finding and assigning a high ranking to true MWEs, compared to the Adjusted Frequency List, which is a current standard-bearer among extraction/ranking-based distributional learning algorithms for MWEs.

It is important to remember that experiments 2a and 2b only compared items that were uniquely found by MERGE to items that were uniquely found by the AFL. That is, items that were found by both algorithms among their respective top 1000-ranked items were not examined. This is because the way in which the two algorithms rank output is fundamentally different. First, they use different units (AFL uses frequency and MERGE log likelihood scores). Second, because frequency is integer-based and log likelihood decimal-based, one cannot simply compare their rank orders, since frequency rank order will exhibit numerous ties and log likelihood rank order will not. Ultimately, then, it is currently impossible to say whether the intersection of the outputs of the two algorithms would be scored higher or lower than the unique items from each of the algorithms.

The fact that experiments 2a and 2b yielded distinct results points to the effects of the alterations made to the instructions. Since several major and minor changes were made, it is impossible to know which specific alteration or alterations were responsible for the different results. It seems likely at least that the two major changes of (1) placing greater

emphasis on the notion of memorization through the use of the term *vocabulary*, and (2) the de-emphasis of the importance of frequency through removal of the requirement that the sequences be *common*, were principally responsible for the modulated results.

One question that remains unanswered is why, if the original instructions were flawed, the results in experiment 1 were so robust. It seems that, despite their non-optimal formulation, they were still effective enough to tap into participant intuition regarding formulaicity for the purposes of the experiment 1 stimuli, which were divided into two groups starkly different in terms of quality of MWEs. That is, because the early bin test items (taken from the very beginning of the MERGE process) were very good MWEs, and the late bin test items (taken from very late in the MERGE process) were very bad MWEs, the initial instructions were capable of directing participants to distinguish between these groups, despite non-ideal wording. By contrast, the MERGE-sourced items and the AFL-sourced items from experiments 2a and 2b were both from the top 1000 ranking of their respective algorithms. Thus, even though these groups of word sequences differed in typology (as described above), they were likely more similar in validity as formulaic language compared to the stimuli groups in experiment 1. As a result, the original instructions were not fine-tuned enough to direct participants to distinguish between the stimulus groups in experiment 2a.

4.4 Conclusion

In this chapter, I set out to accomplish two goals. First, I wished to demonstrate that MERGE is a viable computational approach to the distributional learning of MWEs. The results of experiment 1 provided evidence for this. Specifically, they showed that the output

items that MERGE learns early in its iterative process (i.e., those that it ranks highly as MWEs) are accordingly scored higher by human raters as viable MWEs compared to those output items that it learns later/ranks lower.

Second, I wished to demonstrate that the distributional learning architecture that MERGE embodies offers something more powerful than the other extraction and ranking algorithms available in the literature. Experiments 2a and 2b were designed to provide evidence for this. Initially, I set out to compare the results of MERGE to the results of two other popular algorithms: LocalMax (Da Silva et al. 1999) and the Adjusted Frequency List (Brook O'Donnell 2011). However, after early inspections comparing model outputs revealed that the LocalMax algorithm was actually quite a poor performer, it was decided to only compare MERGE to the AFL.

And while the results of experiment 2a failed to distinguish between the performance of these two algorithms, refinements to the instructions given to study participants led to a clear distinction in algorithm performance in experiment 2b. Specifically, out of the top 1000 items found by MERGE and the top 1000 items found by the AFL, a sampling of those items unique to each algorithm's output was compared. It was found that MERGE's unique output is scored more highly than the AFL's unique output. In fact, the distribution of scores of unique items from the AFL output was rather uniform, suggesting that this algorithm was not finding much formulaic language at all. Note that this is at least true for those items that it found that did not overlap with MERGE's output, for overlapping output between the two algorithms was not examined.

Thus, the strong performance of MERGE warrants its use in the various applications for distributional learning algorithms of formulaic language, as discussed in chapters 1 and

2. These include applications such as lexicography (Sinclair 1987), genre/dialectology/variety studies (Gries and Mukherjee 2010), and second language acquisition research (Simpson-Vlach and Ellis 2010).

5. Modeling Children's Acquisition of MWEs

The previous chapter established that MERGE learns MWEs, and it performs better at this task than other extraction and ranking algorithms from the literature. From a cognitive perspective, it acquires representations that are compatible with human-like representations. Thus, it provides evidence that design features embodied by the model, including its approach to chunking bigrams and basing co-occurrence decisions on contingency-based lexical associations, may be useful in building and maintaining mental representations of MWEs.

In addition to adult mental representation and processing, a central area for cognitive-oriented research in general is child language acquisition. In this chapter, I use MERGE to explore issues relating to children's acquisition. There are a couple of reasons for such an approach. Childhood is when learning language begins, and it is a very learning-intensive time for language. Relatedly, and as discussed in chapter 2, MWEs play a particularly integral role in child language. Specifically, they serve as a means to more productive grammatical knowledge: children begin with stored MWEs and, over time, generalize across them to acquire a mature grammar. (At the same time, remember that this is not to say that representations of MWEs acquired during childhood do not endure into adulthood, nor that new MWEs are not acquired beyond childhood).

Furthermore, studying child language allows us to look at the 'real' input for learning, especially if we are examining longitudinal data for a particular child. In the experiments reported in the previous chapter, the corpora used comprised diverse texts

sourced from a variety of speech genres. And while such corpora are theoretically taken to be representative of the language that native speakers may be exposed to, in no way do these corpora represent the actual data a particular human hears. With a longitudinal child language corpus, however, a researcher has a record of a sample of the real caregiver input on which a particular child is basing his or her acquisition of language. What is more, the corpus provides a record of the child's linguistic productions, which must reflect this process of acquisition.

One possible approach, then, is to attempt to show that particular structures in the child's output can be derived from particular structures in the input. This is essentially the approach taken in one of the case studies in Bod (2009), detailed in chapter 2. In that study, the author develops a parsing/grammar induction algorithm called UDOP (for *Unsupervised Data-Oriented Parsing*; see 2.3.1). He then evaluates it in various case studies; in one, he partitions a longitudinal child language corpus into two sections, and then trains UDOP on the adult utterances in the earlier partition (in separate trials, he also trains the algorithm on the child utterances, and a combination of the child and adult utterances). [8] Next, he evaluates the algorithm by seeing how well it can parse the child utterances in the later partition, based on the grammar it had acquired from the earlier adult utterances. The parses assigned are compared to hand-annotated, gold standard parses of the data.

Ultimately, then, the goal of that approach is the acquisition of a phrase structure grammar—while the algorithm does acquire lexicalized subtrees (viz., MWEs), these are

[8] A related approach is taken in Bannard et al.'s (2009) study using a Bayesian-based distributional learning algorithm that the authors had developed, as well as in Lieven et al.'s (2009) corpus-based discourse-analytic study. However, a crucial difference is that these studies use child utterances for both training and test; thus, there is no attempt to link the children's acquired structures to adult input, but rather just to account for the children's advancing linguistic development across different stages of the child's own usage.

merely a means to a phrase structural end. In the case of MERGE, however, the specific MWEs that a child learns (and, conversely, those that they do not learn) based on adult input are the primary focus, rather than a grammar yielded. Indeed, MERGE does not acquire a grammar. Thus, the UDOP approach is not directly applicable.

Recall that the primary output of MERGE is an extracted and ranked list of structures, rather than a most probable parse. Thus, the cognitive approach whose output is most like that of MERGE in its format is the algorithm by Swingley (2005), the only other cognitive-oriented extraction/ranking algorithm I am aware of, which is reviewed in chapter 2. In fact, Swingley's study is on child language, although he examines the distributional learning of word boundaries from syllable co-occurrences, rather than the acquisition of MWEs. Furthermore, the corpus he uses is not longitudinal, but rather a collection of caregiver utterances (phonologically transcribed) to a collection of different children.

Nonetheless, Swingley's basic design is instructive. He extracts all syllable bigrams and trigrams, scores them on the basis of the mutual information association measure and frequency, and ranks them. He then correlates this ranked list with how well the *n*-grams instantiate words. In other words, he examines the question of how well association strength and frequency among syllable co-occurrences can predict the word boundaries that children learn. However, his definition of what children learn is mature, adult-like gold standard boundaries. An interesting and perhaps more informative approach would be to examine how well the ranked *n*-grams predict the word boundaries that a child learns at the particular developmental stage of the corpus (which would be possible with a longitudinal corpus).

In the current chapter, my approach brings together techniques developed in evaluation methods from the child language studies in Bod (2009) and Swingley (2005). As

in both approaches, I train the algorithm (MERGE) on a set of adult utterances. And like the Bod (2009) study—but unlike Swingley (2005)—I use longitudinal corpora, focusing on the input to/output from individual children. I compare the multi-word representations based on earlier adult utterances that are generated by the model against the actual output of these children, as registered in later child utterances. And like Swingley—but unlike Bod—I work with a list of output candidates scored and ranked on the basis of association strength, rather than best grammatical parses for whole utterances. The hypothesis is that higher-scoring MWEs, extracted from the adult utterances, will go on to be learned/used by the child, while MWEs that scored lower will not.

In the next section, I discuss the corpora that I use as well as their pre-processing, and I discuss the technique for generating the stimuli items from the corpora using MERGE. After that, I turn to the results of the study. Finally, I discuss these findings and conclusions.

5.1 Materials

In this study, I use two longitudinal child language corpora, both of which were sourced from the CHILDES database (MacWhinney 2000). They are the Lara corpus (Rowland and Fletcher 2006) and the Thomas corpus (Lieven et al. 2009). Both children were growing up in the United Kingdom (and were thus being raised as native speakers of varieties of British English), and the recordings for both corpora were made in the children's respective homes.

These corpora were selected for a few different reasons. First, they both span the early multi-word speech stage of development: Lara was between the ages of 1;9.13 and 3;3.25 when her recordings were made, and Thomas was between the ages of 2;00.12 and

4;11.20 when his recordings were made. Thus, this time period is ideal for investigating early MWE acquisition. Second, both corpora include extensive speech from the children as well as the caregivers with whom they were interacting (and, in the case of the Thomas corpus, researcher speech as well). As discussed in section 5.0, both sources of utterances are necessary for the current study. Finally, the corpora are relatively large: while Lara comprises 120 hours of transcribed audio, Thomas totals 379 hours of transcribed audio.

The Thomas recordings/transcriptions are divided into three subcorpora. The first subcorpus spans the ages of 2;00.12 to 3;02.12, and recordings were made for one hour, four times per week. The second and third subcorpora span the remainder of the time, and recordings were made for one hour, once per week.[9] Because the first subcorpus overlaps in time most closely with the Lara corpus, only recordings from it were used. Even with this limitation, the first subcorpus still comprises 279 hours of transcripts: more than double the size of the Lara corpus. In order to make the corpora more comparable in size, the first Thomas subcorpus was downsampled by including only every other corpus file. This resulted in a more comparable 140 hours of transcripts.

Both corpora were transcribed according to the CHAT format (MacWhinney 2000), and so the same preprocessing procedure was used. This included the removal of metadata, transcriber commentary, punctuation, time stamps, non-speech vocalizations, and incomprehensible syllables. In addition, transcription tags were removed, which marked phenomena such as missing words, grammatically correct forms when an incorrect form appeared, and invented forms, among other things. Note that, while incomprehensible forms

[9] The Thomas corpus additionally included video data, but this was not used in the present study.

were removed, grammatically/phonologically incorrect and invented forms were themselves included.

Speaker tags were also removed, but not before they were used to separate each corpus into child and caregiver/adult utterances. Additionally, the two corpora were divided into two partitions, whereby the first two-thirds of each corpus represented partition A and the final third represented partition B.

MERGE was then run on the adult utterances of partition A. No gaps in the MWEs acquired were permitted, and the algorithm was allowed to run until the log likelihood score of the top-scoring merge candidate reached 0. (Remember that positive log likelihood values signify statistical attraction between bigram elements while negative values signify statistical repulsion. By this standard, all bigrams exhibiting a positive log likelihood score are technically MWEs.). From the final output, all MWEs of lengths 2 through 5 were retained.

Next, all *n*-grams from lengths 2 through 5 were extracted from the child utterances in partition B. From this group, any *n*-grams that also appeared among the child utterances in partition A were discarded in order to ensure that the group comprised only *n*-grams that were new attestations in the child's speech. Finally, the MWEs from the MERGE output were compared to the *n*-grams from the partition B child utterances, and two lists were created. The first list comprised those MERGE output MWEs that also appeared as *n*-grams in the child utterances. These are MWEs that the child plausibly went on to learn in partition B from the input they received from the adult utterances in partition A. The second list comprised those MERGE output MWEs that did not appear as *n*-grams in the child utterances. These are items that, despite being MWEs in the adult utterances from partition

A, did not later go on to be learned by the child. The hypothesis is that the log likelihood scores on the basis of which the MWEs were merged ought to be higher for the first, 'learned,' group than for the second, 'nonlearned,' group. In other words, the 'best' MWEs according to MERGE should be learned while the 'worst' MWEs should not.

5.2 Results

As in the experiments in chapter 4, all statistical tests were run and plots were produced using the R programming language. In order to analyze the results of the study, first the scores of both learned and nonlearned MWEs were combined and rank ordered for each child. The resulting two rank-ordered lists were then partitioned into (nearly) equal sized bins. The list containing the scores from the Lara trial were partitioned into 75 bins, with 99 scores in each. The list containing the scores from the Thomas trial were partitioned into 213 bins, with 97 scores per bin. Due to the limited number of factors of the total number of scores in each list, the final bins (i.e., the one corresponding to the highest log likelihood scores), have slightly less than 99 and 97 items in them, respectively. At the same time, these bin counts / number of items per bin were chosen in order to ensure that the final bins came as close to possible in number of items contained as the other bins in the series, and to ensure that the bins for the Lara scores versus the Thomas scores were roughly equal in size. This did mean, however, that the number of bins for the Thomas scores is much greater, since the Thomas corpus is larger—and thus there are more learned and nonlearned MWEs that MERGE extracted from the adult utterances.

At the next stage, the proportion of MWEs (corresponding to the scores) that were learned is calculated for each bin. These proportions are plotted against the bin numbers for

Lara in Figure 5.1 and for Thomas in Figure 5.2. Note the consistent pattern across the two. On the right half of the plot, as one moves from mid-range log likelihood scores to high log likelihood scores, there is an increase in the proportion of MWEs per bin that are learned by each child. This is precisely what we predicted. However, the plots also display something unexpected. Moving from the low log likelihood scores on the left of the plots, to the mid-range scores, there is a *decrease* in the proportion of learned MWEs. This pattern goes against intuition—why might this be?

Figure 5.1. Lara LL Bin X Proportion MWEs Learned

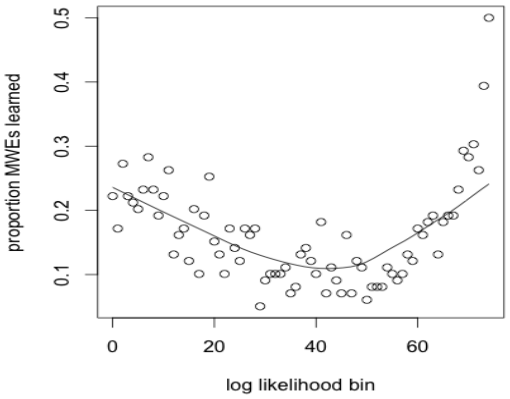
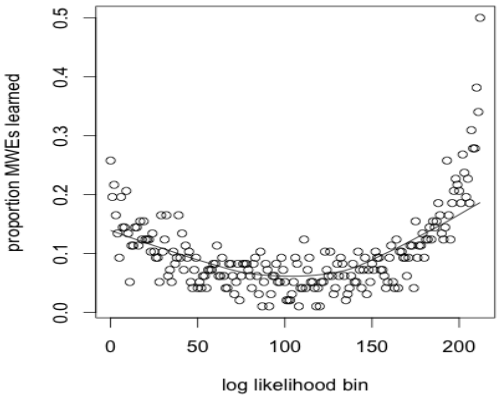


Figure 5.2. Thomas LL Bin X Proportion MWEs Learned



Given that children are highly particular about the lengths of sequences that they can produce at a young age, and given that we are considering MWEs acquired by MERGE of up to five words, length is a logically possible confound. Thus, in Figures 5.3 and 5.4, I have plotted the average lengths of the MWEs in each bin against the bin numbers. Strikingly, the pattern is a virtual mirror image of the scatterplots depicted in Figures 5.1 and 5.2, despite

Figure 5.3. Lara LL Bin X Average MWE Length

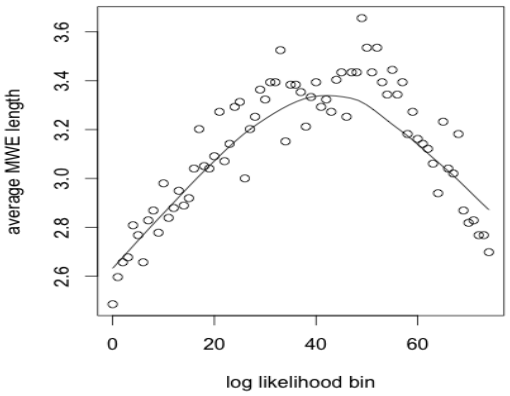
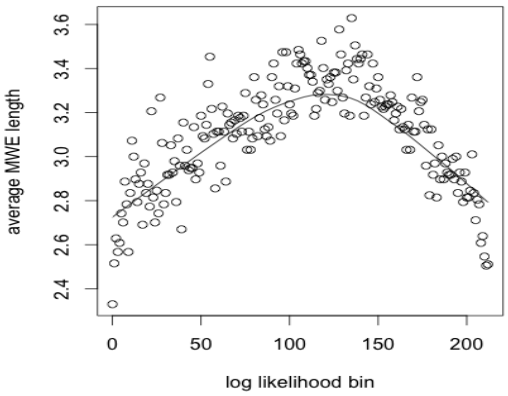


Figure 5.4. Thomas LL Bin X Average MWE Length



the fact that the *y*-axis measures a different unit. In the present context, the pattern signifies that, for both children, the average length of very low and very high scoring MWEs is very short; however, MWEs that were merged on the basis of a mid-range score are, on average, considerably longer.

The isomorphy between the plots in Figure 5.1/5.2 versus 5.3/5.4 suggests that perhaps the variable which holds all the predictive power for the proportion of MWEs learned is average length, not log likelihood bin. Indeed, in Figures 5.5 and 5.6, I have plotted average MWE lengths against the proportion of MWEs learned for each child, and the apparent correlation between these two suggests that average length may be strongly predictive of the dependent variable. This appears particularly true for higher average lengths, where all data points correspond to a low proportion of MWEs learned.

Figure 5.5. Lara Average MWE Length X Proportion MWEs Learned

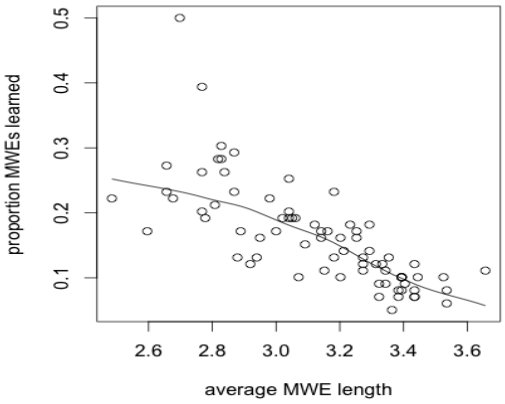
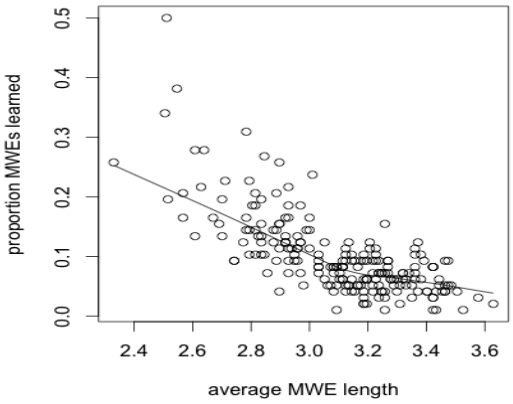


Figure 5.6. Thomas Average MWE Length X Proportion MWEs Learned



Note, however, that for shorter average lengths, there are data points which correspond to both rather high and rather low proportions of MWEs learned. Now consider this in light of what we have just seen in Figures 5.3 and 5.4: shorter average lengths are split between early and late log likelihood bins, and, as we can see in Figures 5.1 and 5.2, these early and late bins exhibit differences in the proportions of MWEs learned. Thus, while it might be the case that average length predicts proportions learned quite well for high length values, the considerable variance in proportions learned for low length values may be explained by the log likelihood bin variable. In other words, it might not be the case that log likelihood is superfluous after all.

To investigate this hypothesis empirically, I fitted linear models to both the Lara and Thomas data sets. For both models, proportions of MWEs learned served as the dependent variable, and log likelihood bin and average MWE length served as predictors. All variables were treated as numeric. The summary statistics are provided in Appendices F and G. In both models, both main effects (Lara bin p-value = 5.33×10^{-4}; Lara average MWE length p-

value = 1.28 x 10^{-3}; Thomas bin p-value = 4.54 x 10^{-4}; Thomas average MWE length p-value = 2.31 x 10^{-11}) as well as their interaction (Lara predictor interaction p-value = 2.7 x 10^{-3}; Thomas predictor interaction p-value = 0.01) are significant. Furthermore, adjusted as well as multiple R^2 for both children's linear models are around 0.7, indicating a high degree of model fit.

In Figures 5.7 and 5.8, I provide plots depicting the regression surfaces themselves.[10] The *x*- and *y*-axes correspond to the two predictor variables, log likelihood bin and average MWE length, respectively. The integer values within the plots can be thought of as 'relative elevations' corresponding to the different predicted proportions of MWEs learned, given the intersecting values of the two predictors. In other words, for each model, the difference between the lowest and highest predicted proportion was rescaled to fall between 0 and 9 on an ordinal scale. For the Lara model, the lowest predicted proportion was 0.02 (=0) and the highest was 0.38 (=9). For the Thomas model, the lowest predicted proportion was an artifactual -0.02 (=0) and the highest was 0.27 (=9).

As expected, the model fit the regression surface such that low proportions of MWEs learned are predicted for high average MWE length, regardless of log likelihood bin. In contrast, for a shorter average MWE expression length, the model predicts a moderate proportion of MWEs learned for lower log likelihood scores yet a high proportion of MWEs learned for higher log likelihood scores. This contrasting effect appears in the summary statistics above as the significant interaction.

[10] The R code for this plot was authored by Stefan Th. Gries and provided to me via personal communication.

Figure 5.7. Effects Plot for Lara Regression

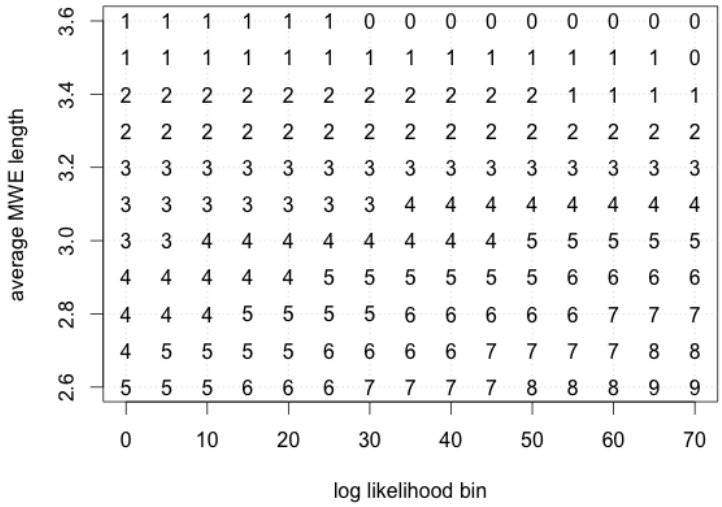
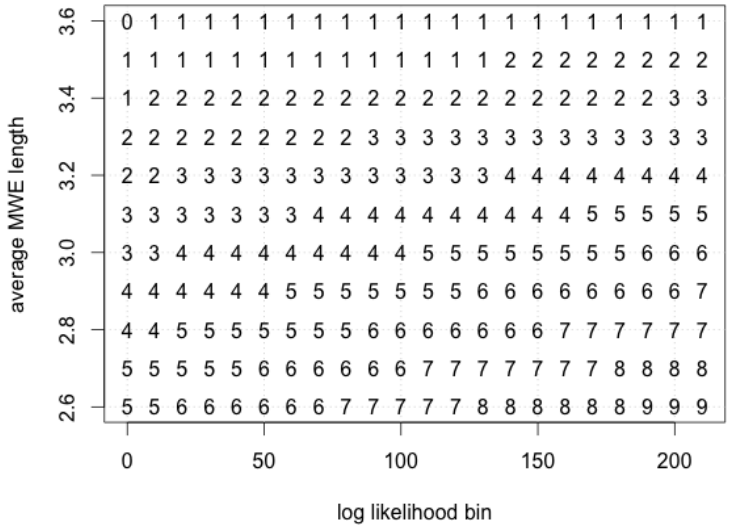


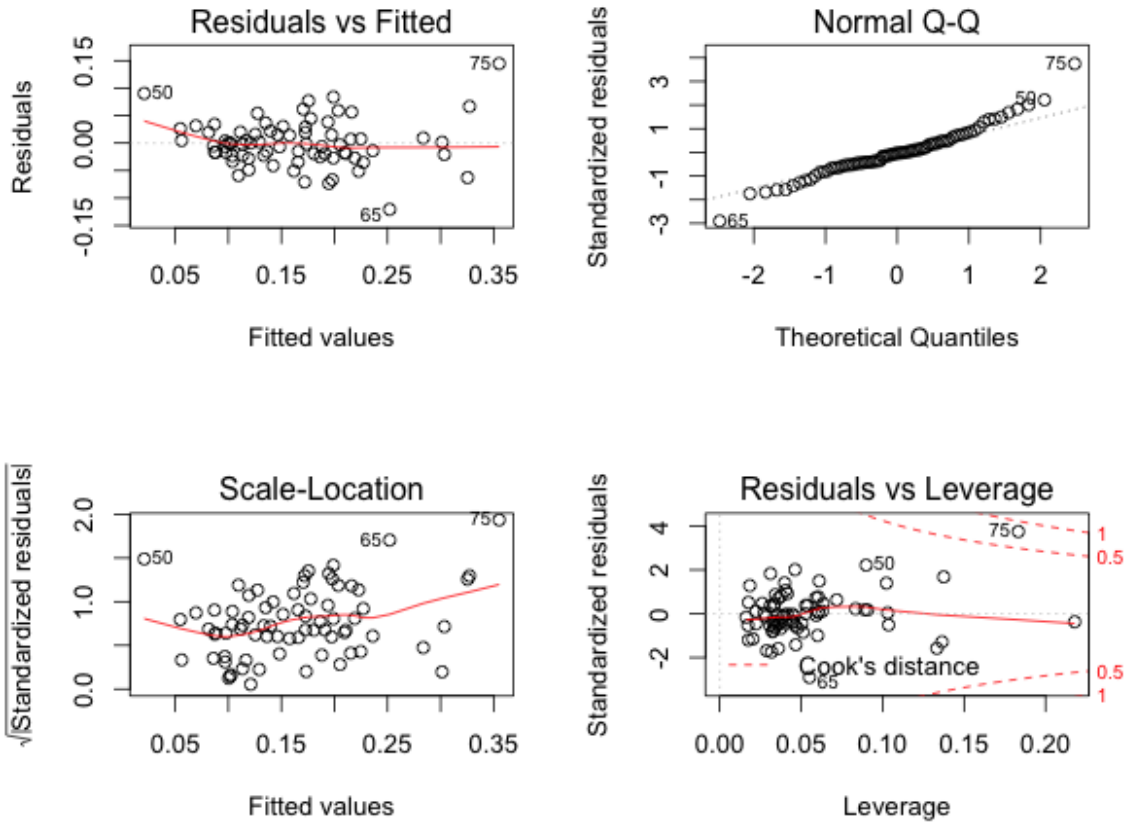
Figure 5.8. Effects Plot for Thomas Regression



Note that the axes are the same as in Figures 5.3 and 5.4. Thus, the real data comprises the parabolic distribution of points depicted in those earlier plots, yet the model has predicted values for the entire plane. Most of these predicted values represent impossible combinations of predictor values. For example, the average length of MWEs from the middle log likelihood bins is rather high. Thus, even though the model has assigned predicted values to hypothetical mid-range log likelihood bins that contain MWEs of a short average length, this combination of variables does not in reality happen. Nonetheless, for the attested variable combinations, the model fit is quite good.

Finally, I turn to the issue of model diagnostics. The four plots produced from the R effects function are rendered for Lara and Thomas in Figures 5.9 and 5.10 below, respectively. The two left-hand plots depict the fitted values against the raw (top left) and square root of standardized (bottom left) residuals. Ideally, one should see little if any change in the vertical dispersion of points as one moves horizontally across the graph. The non-constant variance score test in R checks for such a change. Unfortunately, this test reveals a significant change in such variance along the horizontal dimension for both the Lara ($p = 1.61 \times 10^{-5}$, chi-square = 18.6) and the Thomas ($p = 6.25 \times 10^{-13}$, chi-square = 51.76) linear models. And while this finding is undesirable because it means that there is additional structure in the data not being accounted for by the models, it is not fatal to the results.

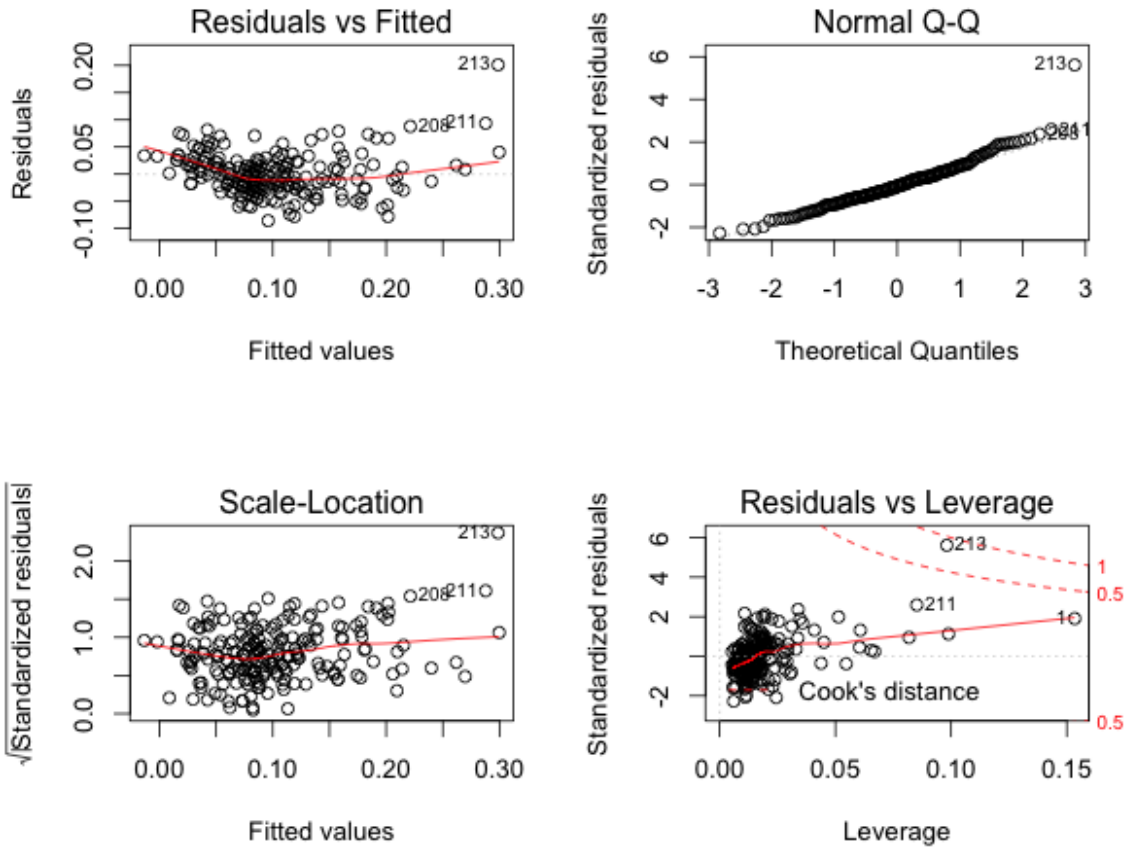
Figure 5.9. Model Diagnostics for Lara Regression



In the top right-hand plots, the fact that the points cluster close to the dotted line indicates that the residuals are more or less normally distributed, which is an assumption of the linear modeling technique.

The bottom right-hand plots indicate that at least one data point in each model exhibits a considerable amount of leverage—in other words, this point exercises a disproportionate amount of influence over the slope of the regression. In the Lara model, this is the 75th log likelihood bin. In the Thomas model, this is the 213th log likelihood bin.

Figure 5.10. Model Diagnostics for Thomas Regression



In both cases, these are the final bins, corresponding to the highest log likelihood values. It is perhaps unsurprising that these bins exhibit such high leverage: returning to Figures 5.1 and 5.2, the proportion of MWEs learned for these two bins is considerably higher than for the bins corresponding to the second-highest proportion of MWEs learned.

Excessively high leverage data points are considered potentially undesirable in linear modeling because they can distort the true fit to the data. One technique for handling such cases is to simply remove these points and re-run the models. After doing this, there is little overall change to the results of the two models, with R^2 values and significant predictors

remaining largely the same. The one major difference is that, after removing bin 213 in the Thomas model, the interaction between the average MWE length and log likelihood bin variables is no longer significant (see Appendices H and I for summary statistics for these cropped models).

At the same time, these are not typical high leverage data points. First, they are consistent with the underlying shape of the data. Second, remember that these are not individual observations in a traditional sense. Nearly 100 separate observations went into the calculation of the proportion of learned MWEs and average length for each bin. Thus, the high leverage of these points results from the combined influence of all of these separate observations, rather than of just one outlier observation. Ultimately, although the cropped models are largely isomorphic with the original models, it is not clear that the process of cropping is even necessary in the current case.

5.3 Discussion

As hypothesized, the two children whose early multi-word productions are examined in this study, Lara and Thomas, acquire a higher proportion of MWEs exhibiting high association strength than MWEs exhibiting a low or mid-range association strength. Such strength is measured in terms of the log likelihood scores used by the MERGE algorithm to extract MWEs from adult utterances preceding the child's productions—thus these adult utterances are plausible as input to child acquisition. These results once again assert the importance co-occurrence regularities between linguistic units in children's acquisition of structure.

As we have seen, however, this relationship is not the whole story. The average length of MWEs for particular ranges of log likelihood strengths (the bins) also bear on MWE acquisition. Perhaps unsurprisingly, these children, in the age range of 2-3, are strongly averse to learning long sequences, regardless of the association strength. Thus, the effect of log likelihood on learning, just described above, holds true only for short sequences.

5.4 Conclusion

Overall, these results provide further evidence for the viability of MERGE in investigating human distributional learning of MWEs. This study goes beyond the experiments in the previous chapter. As before, those experiments were based on corpus data that were certainly not the actual input on which the study participants had based their MWE acquisition. Furthermore, they focused on adult language, when child language is arguably a richer domain for the study of the processes of distributional learning of MWEs. Through the use of these two longitudinal child language corpora, it was possible to plausibly link specific child output to co-occurrence regularities in the adult input likely responsible for the newly learned (as well as absent nonlearned) structures. Ultimately, these results suggest the usefulness of contingency-based lexical association and chunking to children's acquisition of MWEs.

In the future, it would be interesting to deploy the same basic approach but using corpora of the speech of slightly older children—that is, from an age range when longer word sequences are more common. Under these conditions, the lack of an effect for log likelihood bin in the case of longer MWEs may disappear. Indeed, it would be desirable to

see such an effect of bin rank for all MWE lengths studied, since one of MERGE's most valuable components—and what allows it to master MWE knowledge that reflects human knowledge—is its ability to learn MWEs of all sizes. In the current studies, the same effect of bin rank being restricted to short MWEs could have been found with a simple bigram extraction approach; thus, to truly validate the unique contribution of MERGE, older child data would be helpful.

6. Final Summary and Conclusions

In this concluding chapter, I provide a summary of each of chapters 1 through 5, and then I end by discussing directions for future work.

6.1 Interim Summary

I begin this dissertation in chapter 1 by introducing the idea of computational modeling of distributional learning, and I outline several different dimensions along which types of distributional learning algorithms can vary. These include the types of linguistic units tracked (here, word co-occurrences are tracked in order to identify MWEs), whether the algorithm is designed to address cognitive or non-cognitive questions, how much built-in linguistic knowledge the algorithm has access to, and whether the algorithm learns knowledge that is syntagmatic, paradigmatic, or both.

Variation along this last dimension corresponds to three different basic algorithmic architectures that are relevant to computational research on distributional learning: extraction and ranking algorithms (which acquire syntagmatic knowledge), category acquisition/unsupervised part-of-speech tagging algorithms (which acquire paradigmatic knowledge), and parsing/grammar induction algorithms (which acquire both syntagmatic and paradigmatic knowledge). I discuss these three different types, and how they intersect with the cognitive/non-cognitive dimension, in order to identify gaps in the current research and thereby motivate the development of the MERGE algorithm. Specifically, in the non-cognitive domain, the extraction/ranking architecture is standard, yet few current algorithms

can identify MWEs that contain gaps and can learn MWEs of differing lengths in a bottom-up fashion. Conversely, in the cognitive domain, there are currently no extraction/ranking algorithms used to investigate learning of MWEs, despite the unique kind of cognitive evidence that I argue extraction extraction/ranking algorithms can provide. MERGE fills both of these gaps, through its design as a cognitive and non-cognitive extraction/ranking algorithm that can flexibly acquire MWEs of varying lengths, with or without gaps.

I continue in chapter 2 by providing theoretical context centered on the notions of distributional learning and formulaic language. I begin by introducing an approach that held great sway in the last century known as generative linguistics, and two of the central tenets of the generative perspective: universal grammar and its relationship to the poverty of the stimulus, and a strict separation of productive syntactic rules from stored lexical entries. I then review recent research on two phenomena that have been instrumental in challenging universal grammar and the poverty of the stimulus, as well as separate syntactic and lexical systems: distributional learning and formulaic language, respectively.

Next, I suggest that, while these two research trends have typically been treated separately because of an assumed incompatibility between them, this incompatibility is fallacious: distributional learning research on child language, which specifies the role of chunking in identifying the units of language, is not incompatible with findings in formulaic language research showing that children learn whole, multi-word sequences. Rather, computational research has shown that learning via whole multi-word sequences can be achieved through chunking; just because children produce such chunks whole does not mean they have no knowledge of lexical structure below the chunk (and in fact we know that they do).

Thus, while distributional learning research and formulaic language research have in different ways been seminal in dispelling outdated generative ideas, computational research has been indispensable in unifying these two important trends. At the end of chapter 2, I review in detail both cognitive and non-cognitive computational algorithms for the distributional learning of MWEs, and conclude by revisiting the contributions that MERGE makes to this important domain of inquiry.

In chapter 3, I give a thorough description of the MERGE algorithm itself. I begin by explaining the corpus preprocessing. Next, I present the choice of statistical metric for evaluating candidate merges. And finally, I give a step-by-step account of the iterative functioning of the algorithm.

Chapters 4 and 5 represent empirical approaches designed to validate the efficacy of the model in acquiring MWEs. In chapter 4, I report three experiments in which human participants rate model output items according to how well they represent MWEs, and then I correlate these ratings with the scores assigned by the model(s) to the output items. In the first experiment, I look purely at the output of MERGE, finding that the algorithm does indeed start by identifying high-quality MWEs, while later output items that it finds are not rated highly. In other words, not only does MERGE find MWEs, it ranks them appropriately by quality.

In the second and third experiments, I compare the output of MERGE to the output of the Adjusted Frequency List, another popular algorithm from the literature that is designed to learn MWEs. Ultimately, I find that MERGE performs better than this other algorithm, suggesting that the MERGE approach represents the vanguard of extraction/ranking-based distributional learning algorithms for multi-word expressions.

Finally, in chapter 5, I take two longitudinal child language corpora and run MERGE on a set of caregiver utterances from each. I then check a set of child utterances that temporally follow the adult utterances for instantiations of the output of the run of MERGE. Ultimately, I find that more high-scoring than low-scoring MWEs go on to be learned by the children, although this appears to be true only for MWEs that are short in length (all longer MWEs are learned poorly).

Overall, then, MERGE offers an innovative approach to the distributional learning of MWEs, an area of computational and corpus linguistics that is only growing as linguistic theory increasingly turns towards the importance of formulaic language and distributional learning. The algorithm is effective and computationally tractable, making it viable both as a non-cognitive tool for corpus-based research, and for the exploration of cognitive questions.

6.2 Future Directions

In the future, a number of open research questions remain. One area for further development has to do with the gram size considered for merging at each iteration. Currently, only bigrams are considered. While computationally this creates a smaller collection of merge candidates to consider, cognitively it is less realistic than a model which would consider grams of different sizes simultaneously. Indeed, it is likely that a child learning a MWE does not always combine words in groups of two. Thus, a future implementation of MERGE in which bigrams are considered alongside trigrams and perhaps tetragrams would be an interesting development.

Another area for further development has to do with the fact that MWEs are very often not fully lexicalized sequences. Often, positions within them can be filled with

different members of some lexical or phrasal category; knowledge of this is in theory part of the knowledge of the formulaic sequences. This is the insight behind the inclusion of discontinuous slots in the MWEs that MERGE can acquire. However, MERGE cannot tell us what can go in these slots. In the future, it would be ideal to develop an implementation of MERGE that learns not only that there are variable slots, but what different paradigmatic items can fill those slots. In this way, MERGE would become a model of paradigmatic as well as syntagmatic learning, and would be more similar to the parsing/grammar induction algorithms discussed in chapter 2 and elsewhere.

Relatedly, although MERGE can learn MWEs with gaps of different sizes, it cannot recognize that two MWEs with different gap sizes may in fact instantiate the same lexical type. For example, MERGE would treat *sit _ up* and *sit _ _ up* as different items since they have a different number of intervening words. Obviously, in many cases, they are not actually different lemmas. In the future, it would be desirable to develop the algorithm further to account for this equivalency.

Furthermore, MERGE does not distinguish between homonyms, which is not ideal. Thus, being able to distinguish homonymous tokens of a merge winner from true tokens is another future goal.

For these preceding three points, the technique that holds promise for acquiring the relevant knowledge is unsupervised part-of-speech tagging, whereby words or sequences that are surrounded by similar distributions of neighboring words/sequences are grouped together as members of the same paradigm. The value of such an approach to the issues under discussion here is that unsupervised part-of-speech tagging would allow one to distinguish between true and false instances of a category. For example, just because there is

a token of *sit _ _ up* does not mean that it is the phrasal verb: it could be, as in *sit the baby up so that he can eat* or *I need to sit myself up*, but it also could be a false match as in *sit with me up in the bleachers*. The different distributions of surrounding words could in theory be used to recognize that the noun phrases *the baby* and *myself* belong together—and thus the instances of *sit _ _ up* surrounding them instantiate the phrasal verb—while the prepositional phrase *with me* is distinct. Thus, the third token of *sit _ _ up* should not be counted as the phrasal verb.

In this way, unsupervised part-of-speech tagging could be used to group together MWE candidates with different-sized gaps (and exclude false matches), to find true paradigm members for slots (*myself* and *the baby* but not *with me*), and even to distinguish homonyms. Ultimately, it would allow MERGE to acquire a richer set of knowledge about formulaic language, which would enhance the already strong parallel between what humans know when they deploy MWEs in their day-to-day speech, and what the MERGE algorithm acquires.

Appendix A. Survey Instructions

In this experiment, you are going to rank different word sequences on a scale of 1 to 7. It is your job to determine how well the words tend to go together as a common, reusable "chunk."

If something is really familiar to you as a common, reusable chunk, you should give it a high score. If something is less familiar to you, you should give it a low score.

Note that the word sequences can contain gaps. These are marked by three underscores (___). Please also note that if the sequence is only *part of* the complete chunk, you should rate it lower.

Here are some examples of words that tend to go together as a chunk:

"you know what I mean"	He loves her but he's not in love with her, **you know what I mean**?
"or something like that"	Do they go for like a week, or two weeks, **or something like that**?
"a lot of"	**A lot of** people live there.
"you know"	Well, **you know**, I had some really really big problems.
"as ___ as"	A person who is a party to an offense is just **as** guilty **as** the principal offender.

Here are some examples of words that WOULD NOT form a common chunk:

"holds ___ eclipse awards which is"	Billy **holds** eight **eclipse awards, which is** more than any other horse.
"I already lost a dollar"	I said, "no way. I bet a dollar, **I already lost a dollar**."
Alan Eagleson	A discipline hearing against **Alan Eagleson** begins today.
"beacon ___ light"	We tell the world we are this **beacon** of **light**, of freedom, of liberty, justice, and all these different things.
"She had a brother"	**She had a brother** that died from asthma.

Appendix B. Summary Statistics for Mixed Effects Regression

Linear mixed model fit by REML t-tests use Satterthwaite approximations to degrees of freedom [merModLmerTest]
Formula: SCORE ~ poly(SIZE, 2) * BINRANK + (1 + BINRANK + SIZE | GRAM) + (1 + BINRANK + SIZE | PARTICIPANT)

REML criterion at convergence: 5065.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.03	-0.50	-0.01	0.44	4.44

Random effects:

Groups	Name	Variance	Std. Dev.	Corr	
GRAM	(Intercept)	1.40	1.18		
	BINRANKlate	1.47	1.21	0.96	
	SIZE	0.31	0.56	-1.00	-0.96
PARTICIPANT	(Intercept)	1.38	1.17		
	BINRANKlate	0.56	0.75	-0.59	
	SIZE	0.03	0.18	-0.93	0.49
Residual		1.01	1.00		

Number of obs: 1600, groups: GRAM, 320; PARTICIPANT, 20

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	5.69	0.16	29.6	34.86	< 2e-16 ***
poly(SIZE, 2)1	-26.26	4.13	129.6	-6.36	3.13e-09 ***
poly(SIZE, 2)2	-13.04	2.85	162.4	-4.57	9.73e-06 ***
BINRANKlate	-3.87	0.20	31.0	-19.17	< 2e-16 ***
poly(SIZE,2)1:BINRANKlate	15.88	4.93	178.6	3.22	1.52e-3 **
poly(SIZE,2)2:BINRANKlate	11.66	3.92	322.2	2.98	3.14e-3**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)	pl(SIZE,2)1	pl(SIZE,2)2	BINRAN	p(SIZE,2)1:
pl(SIZE,2)1	0.07				
pl(SIZE,2)2	0.17	0.54			
BINRANKlate	-0.66	-0.10	-0.13		
p(SIZE,2)1:	-0.29	-0.68	-0.46	0.11	
p(SIZE,2)2:	-0.12	-0.40	-0.73	0.20	0.17

Appendix C. Summary Statistics for Residualized Model

Linear mixed model fit by REML t-tests use Satterthwaite approximations to degrees of freedom [merModLmerTest]
Formula: SCORE.without.SIZE ~ BINRANK + (1 + BINRANK | PARTICIPANT) + (1 + BINRANK | GRAM)
Data: datatable

REML criterion at convergence: 5225.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.88	-0.50	-0.02	0.44	4.12

Random effects:

Groups	Name	Variance	Std. Dev.	Corr
GRAM	(Intercept)	1.43	1.20	
	BINRANKlate	0.49	0.70	-0.88
PARTICIPANT	(Intercept)	0.33	0.57	
	BINRANKlate	0.51	0.71	-0.56
Residual		1.05	1.02	

Number of obs: 1600, groups: GRAM, 320; PARTICIPANT, 20

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.9344	0.1632	38.8100	11.85	1.82e-14 ***
BINRANKlate	-3.8688	0.1992	33.7500	-19.42	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)
BINRANKlate	-0.664

Appendix D. Survey instructions (Second Version)

We make up new sentences whenever we open our mouths, and we do this by piecing together items that we have memorized as part of our vocabulary. When we think about the kinds of things in our vocabulary, we tend to think of individual words. However, consider the following multi-word sequences, and some sentences that they appear in:

"you know what I mean"	He loves her but he's not in love with her, **you know what I mean**?
"or something like that"	Do they go for like a week, or two weeks, **or something like that**?
"a lot of"	**A lot of** people live there.
"you know"	Well, **you know**, I had some really really big problems.
"Santa Claus"	Then **Santa Claus** might not bring you presents.
"carbon dioxide"	And you exhale **carbon dioxide**.

These multi-word sequences are examples of complete, memorized units that are part of our vocabulary, even though they are bigger than individual words. In this survey, you are going to rank different word sequences on a scale of 1 to 7. It is your job to determine how well the words together form a <u>complete unit of vocabulary.</u>

If something seems like a complete unit of vocabulary, you should give it a high score (like the common examples above). If something does not seem like a complete unit of vocabulary, you should give it a low score. Here are some examples of multi-word sequences that are not complete units of vocabulary:

"eclipse awards which is"	Billy holds eight **eclipse awards, which is** more than any other horse.
"I already lost a dollar"	I said, 'no way. I bet a dollar, **I already lost a dollar**."
Alan Eagleson	A discipline hearing against **Alan Eagleson** begins today.
"She had a brother"	**She had a brother** that died from asthma.

Please note: if the sequence is only *part of* the complete unit, you should rate it lower.

Appendix E. Summary Statistics for Second Mixed Effects Regression

Linear mixed model fit by REML t-tests use Satterthwaite approximations to degrees of freedom [merModLmerTest]
Formula: SCORE ~ ORIGIN + (ORIGIN | PARTICIPANT) + (1 | GRAM)

REML criterion at convergence: 7450.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.57	-0.70	-0.05	0.72	2.55

Random effects:

Groups	Name	Variance	Std. Dev.	Corr
GRAM	(Intercept)	0.73	0.85	
PARTICIPANT	(Intercept)	1.33	1.15	
	ORIGINm	1.00	1.00	-0.77
Residual		2.97	1.72	

Number of obs: 1800, groups: GRAM, 360; PARTICIPANT, 20

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.93	0.27	20.71	14.49	2.61e-12 ***
ORIGINm	0.59	0.25	22.83	2.32	0.03 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)
ORIGINm	-0.75

Appendix F. Summary Statistics for Lara Regression

Call:
lm(formula = props ~ bin * ave.lengths)

Residuals:

Min	1Q	Median	3Q	Max
-0.12	-0.02	-1.69e-03	0.02	0.14

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.59	0.12	4.81	8.37e-06 ***
bin	0.01	2.92e-03	3.63	5.33e-04 ***
ave.lengths	-0.1406918	0.0419313	-3.355	1.28e-03 **
bin:ave.lengths	-3.18e-03	9.94e-04	-3.20	2.07e-03 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04 on 71 degrees of freedom
Multiple R-squared: 0.72, Adjusted R-squared: 0.71
F-statistic: 61.03 on 3 and 71 DF, p-value: < 2.2e-16

Appendix G. Summary Statistics for Thomas Regression

Call:
lm(formula = props ~ bin * ave.lengths)

Residuals:

Min	1Q	Median	3Q	Max
-0.09	-0.02	-2.67e-03	0.02	0.20

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.57	0.07	8.42	5.86e-15 ***
bin	1.95e-03	5.49e-04	3.56	4.54e-04 ***
ave.lengths	-0.16	0.02	-7.07	2.31e-11 ***
bin:ave.lengths	-5.21e-04	1.88e-04	-2.77	0.01 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04 on 209 degrees of freedom
Multiple R-squared: 0.70, Adjusted R-squared: 0.69
F-statistic: 161 on 3 and 209 DF, p-value: < 2.2e-16

Appendix H. Summary Statistics for Lara Regression with Highest Leverage Point Removed

Call:
lm(formula = formula(model_A), data = child_lg_data[-75,])

Residuals:

Min	1Q	Median	3Q	Max
-0.11	-0.02	-1.51e-03	0.02	0.09

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.63	0.11	5.70	2.61e-07 ***
bin	0.01	2.74e-03	2.71	0.01 **
ave.lengths	-0.16	0.04	-4.09	1.15e-04 ***
bin:ave.lengths	-2.19e-03	9.27e-04	-2.36	0.02 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04 on 70 degrees of freedom
Multiple R-squared: 0.70, Adjusted R-squared: 0.69
F-statistic: 54.91 on 3 and 70 DF, p-value: < 2.2e-16

Appendix I. Summary Statistics for Thomas Regression with Highest Leverage Point Removed

Call:
lm(formula = formula(model_A), data = child_lg_data[-213,])

Residuals:

Min	1Q	Median	3Q	Max
-0.08	-0.02	-2.89e-03	0.02	0.11

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.62	0.06	9.72	< 2e-16 ***
bin	1.23e-03	5.21e-04	2.37	0.02 *
ave.lengths	-0.18	0.02	-8.26	1.71e-14 ***
bin:ave.lengths	-2.86e-04	1.78e-04	-1.61	0.11

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03 on 208 degrees of freedom
Multiple R-squared: 0.69, Adjusted R-squared: 0.69
F-statistic: 157 on 3 and 208 DF, p-value: < 2.2e-16

References

Alishahi, Afra and Grzegorz Chrupala (2009). Lexical category acquisition as an incremental process. In *Proceedings of the CogSci2009 workshop on psychocomputational models of human language acquisition*, Amsterdam, Netherlands.

Altenberg, Bengt (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 101 – 122). Oxford: Oxford University Press.

Arnon, Inbal and Neal Snider (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62: 67 – 82.

Aslin, Richard, Jenny Saffran, and Elissa Newport (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4): 321 – 324.

Baker, Mark (2001). *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

Bannard, Colin and Elena Lieven (2012). Formulaic language in L1 acquisition. *Annual Review of Applied Linguistics*, 32: 3 – 16.

Bannard, Colin and Danielle Matthews (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19: 241 – 248.

Bannard, Colin and Danielle Matthews (2010). Children's Production of Unfamiliar Word Sequences is Predicted by Positional Variability and Latent Classes in a Large Sample of Child-Directed Speech. *Cognitive Science*, 34(2): 465 – 488.

Bannard, Colin, Elena Lieven, and Michael Tomasello(2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41), 17284 – 17289.

Barton, Kamil (2015). MuMin: Multi-model inference. R package version 1.13.4. http://cran.r-project.org/web/packages/MuMIn/index.html. Downloaded March 30, 2015.

Biber, Douglas, Susan Conrad, and Viviana Cortes (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3): 371 – 405.

Biemann, Chris (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of COLING ACL 2006* (pp. 7 – 12). Morristown, NJ.

Bod, Rens (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science* 33: 752 – 793.

Bowerman, Melissa (1990). Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? *Linguistics* 28: 1253 – 1289.

Butler, Christopher S. (1997). Repeated word combinations in spoken and written text: Some implications for functional grammar. In C.S. Butler, J.H. Connolly, R.A. Gatward, and R.M. Vismans (Eds.), *A fund of ideas: Recent developments in functional grammar* (pp. 60 – 77). Amsterdam: Institute for Functional Research into Language and Language Use (IFOTT).

Bybee, Joan and Joanne Scheibman (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*, 37(4): 575 – 596.

Cheng, Winnie, Chris Greaves, John M. Sinclair, and Martin Warren (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, 30(2): 236 – 252.

Chomsky, Noam (1957). *Syntactic structures*. The Hague/Paris: Mouton.

Chomsky, Noam (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam (1995). *The minimalist program*. Cambridge: MIT Press.

Conklin, Kathy and Norbert Schmitt (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32: 45 – 61.

Becker, A.L. (1983). Toward a post-structuralist view of language learning: A short essay. *Language Learning*, 33: 217 – 220.

Brook O'Donnell, Matthew (2011). The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal*, 35: 135 – 169.

Brown, Peter F., Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4): 467 – 479.

Chrupala, Grzegorz, and Afra Alishahi (2010). Online entropy-based model of lexical category acquisition. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 182 – 191). Uppsala, Sweden.

Clark, Alexander (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL 2003* (pp. 59 – 66). Morristown, NJ.

Curtin, Suzanne, Tobin H. Mintz, and Morten H. Christiansen (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96: 233 – 262.

Du Bois, John W., Wallace L. Chafe, Charles Meyers, Sandra A. Thompson (2000). *Santa Barbara corpus of spoken American English, part 1*. Philadelphia: Linguistic Data Consortium.

Du Bois, John W., Wallace L. Chafe, Charles Meyers, Sandra A. Thompson, and Nii Martey (2003). *Santa Barbara corpus of spoken American English, part 2*. Philadelphia: Linguistic Data Consortium.

Du Bois, John W. and Robert Englebretson (2004). *Santa Barbara corpus of spoken American English, part 3*. Philadelphia: Linguistic Data Consortium.

Du Bois, John W. and Robert Englebretson (2005). *Santa Barbara corpus of spoken American English, part 4*. Philadelphia: Linguistic Data Consortium.

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

Durrant, Philip and Alice Doherty (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2), 125 – 155.

Everett, Daniel (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46: 621 – 646.

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112: 21 – 54.

Grimshaw, Jane (1981). Form, function, and the language acquisition device. In C.L. Baker and J. McCarthy (Eds.), *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Ellis, Nick C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18: 91 – 126.

Ellis, Nick C., Erik Frey and Isaac Jalkanen (2009). The psychological reality of collocation and semantic prosody. In U. Romer and R. Schulze (Eds.), *Exploring the lexis-grammar interface* (pp. 89 – 114). Philadelphia: John Benjamins.

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.

Erman, Britt and Beatrice Warren (2000). The idiom principle and the open choice principle. *Text*, 20(1): 29 – 62.

Evert, Stefan (2005). The statistics of word co-occurrences: Word pairs and collocations. Dissertation. Universität Stuttgart.

Ferreira da Silva, Joaquim, Gaël Dias, Sylvie Guilloré and José Gabriel Pereira Lopes (1999). Using LocalMax algorithm for the extraction of continguous and non-continguous multiword lexical units. *Lecture notes in artificial intelligence: Progress in artificial intelligence*, 1996: 113 – 132.

Foster, Pauline (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, and M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 75 – 93). Harlow: Longman.

Fodor, Janet Dean (2003). Evaluating models of parameter setting. Handout at LSA institute workshop on UG principles and input. Michigan State University.

Goldberg, Adele E. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Gomez, Rebecca and Louann Gerken (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5): 178 – 186.

Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier and Dan Jurafsky (1999). The effects of collocational strength and contextual predictability in lexical production. In *Proceedings of the Chicago Linguistic Society*. Chicago, IL.

Gries, Stefan Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3 – 25). Amsterdam: John Benjamins.

Gries, Stefan Th. (2013). *Statistics for linguistics with R: A practical introduction*. Second edition. Boston: Walter de Gruyter.

Gries, Stefan and Joybrato Mukherjee (2010). Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics*, 15(4): 520 – 548.

Haiman, John (1994). Ritualization and the Development of Language. In W. Pagliuca (Ed.), *Perspectives on grammaticalization* (pp. 3 – 28). Amsterdam: John Benjamins.

Hockema, S. A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development*, 2: 119 – 146.

Hyams, Nina (2011). Missing subjects in early child language. Studies in Theoretical Psycholinguistics 41: 13 – 52.

Johnson, Elizabeth K. and Peter W. Jusczyk (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44(4): 548 – 567.

Johnson, Mark, Thomas L. Griffiths and Sharon Goldwater (2007a). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Proceedings of Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Johnson, Mark, Thomas L. Griffiths and Sharon Goldwater (2007b). Bayesian inference for PCFGs via Markov Chain Monte Carlo. In *Proceedings of the North American Conference on Computational Linguistics*. Rochester, NY.

Johnson, Paul C.D. (2014). Extension of Nakagawa and Schielzeth's $R2_{GLMM}$ to random slopes models. *Methods in Ecology and Evolution*, 5: 944 – 946.

Kirjavainen, Minna, Anna Theakston and Elena Lieven (2009). Can input explain children's *me*-for-*I* errors? *Journal of Child Language*, 36: 1091 – 1114.

Klein, Dan and Chris Manning (2002). A generative constituent-context model for improved grammar induction. In *Proceedings of the Association for Computational Linguistics*.

Krug, Manfred (1998). String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics*, 26(4): 286 – 320.

Lareau, François, Mark Dras, Benjamin Börschinger, and Robert Dale (2011). Collocations in multilingual natural language generation: Lexical functions meet lexical functional grammar. In *Proceedings of ALTA'11* (pp. 95 – 104).

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

MacWhinney, Brian (2000). *The CHILDES project. Tools for analyzing talk. Third edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, Michael (1981). Problems in categorical evolution: Can formal categories arise from semantic ones? In W. Deutsch (Ed.), *The child's construction of language* (pp. 245 – 261). New York: Academic Press.

Maratsos, Michael and Chalkey, Mary Anne (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language, volume 2* (pp. 127 – 214). New York: Gardner Press.

Mattys, Sven L., Laurence White and James F. Melhorn (2005). Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*: 477 – 500.

McDonald, Scott and Shillcock, Richard C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14(6): 648 – 652.

Mondini, Sara, Gonia Jarema, Claudio Luzzatti, Cristina Burani and Carlo Semenza (2002). Why is "Red Cross" different from "yellow cross"?: A Neuropsychological study of noun-adjective agreement within Italian compounds. *Brain and Language*, 81: 621 – 634.

Moon, Rosamund (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.

Nakagawa, Shinichi and Holger Schielzeth (2010). Repeatibility for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, 85(4): 935 – 956.

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

Newman, John and Georgie Columbus (2010). *The International Corpus of English – Canada*. Edmonton, Alberta: University of Alberta.

O'Donnell, Timothy J., Jesse Snedecker, Joshua B. Tenenbaum and Noah D. Goodman (2011). Productivity and reuse in language. In *Proceedings of the Cognitive Science Society Annual Conference* (pp. 1613 – 1618). Boston, MA.

Pawley, Andrew and Francis H. Sider (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (Eds.): *Language and Communication* (pp. 191 – 225). London: Longman.

Parisien, Christopher, Afsaneh Fazly and Suzanne Stevenson (2008). An incremental bayesian model for learning syntactic categories. In *Proceedings of the 12th Conference on Computational Natural Language Learning* (pp. 89 – 96). Manchester, UK.

Pecina, Pavel (2009). *Lexical association measures: Collocation extraction*. Prague: Charles University.

Pelucchi, Bruna, Jessica F. Hay and Jenny R. Saffran (2009a). Statistical learning in natural language by 8-month-old infants. *Child Development*, 80(3): 674 – 685.

Pelucchi, Bruna, Jessica F. Hay and Jenny R. Saffran (2009b). Learning in reverse: 8-month-old infants track backward transitional probabilities. *Cognition*, 113: 244 – 247.

Pinker, Steven (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, Steven (1989). *Learnability and cognition: The acquisition of verb-argument structure*. Cambridge, MA: Harvard University Press.

Pinker, Steven (1994). *The language instinct: How the mind creates language*. New York: William Morrow & Company.

Pinker, Steven (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Rayson, Paul (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519 – 549.

Reali, Florencia and Morten H. Christiansen (2006). Word chunk frequencies affect the processing of pronominal object-relative clauses. *The Quarterly Journal of Experimental Psycholinguistics*, 60(2): 161 – 170.

Reddington, Martin and Nick Chater (1998). *Connectionist and statistical approaches to language acquisition: A distributional perspective*. Language and Cognitive Processes, 13(2/3): 129 – 191.

Reddington, Martin, Nick Chater and Steven Finch (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4): 425 – 469.

Reeder, Patricia A., Elissa L. Newport and Richard N. Aslin (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66: 30 – 54.

Rowland, Caroline F. and S.L. Fletcher (2006). The Effect of sampling on estimates of lexical specificity and error rates. *The Journal of Child Language*, 33(4): 859 – 877.

Saffran, Jenny R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44: 493 – 515.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996a). Statistical learning by 8-month-old infants. *Science*, 274(5294): 1926 – 1928.

Saffran, Jenny R., Elissa L. Newport and Richard N. Aslin (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4): 606 – 621.

Schlesigner, Izchak M. (1981). *Semantic assimilation in the acquisition of relational categories*. In W. Deutsch (Ed.), *The child's construction of language* (pp. 223 – 243). New York: Academic Press.

Schlesigner, Izchak M. (1988). The origin of relational categories. In Y. Levy, I.M. Schlesigner, and M.D.S. Braine (Eds.), *Categories and processes in language acquisition* (pp.121 – 178). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schone, Patrick and Daniel Jurafsky (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? *Proceedings of Empirical Methods in Natural Language Processing*, Pittsburgh, PA.

Shannon, Claude E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379 – 423.

Simpson-Vlach, Rita and Nick Ellis (2010). An Academic Formulas List. *Applied Linguistics*, 31: 487 – 512.

Sinclair, John (1987). Collins COBUILD English language dictionary. Ann Arbor, MI: Collins.

Siyanova-Chanturia, Anna, Kathy Conklin and Walter J.B. Van Heuven (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology*, 37(3): 776 – 784.

Solan, Zach, David Horn, Eytan Ruppin, and Shimon Edelman(2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33): 11629 – 11634.

Sosa, Anna V. and James MacFarlane (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, 83: 227 – 236.

Stefanowitsch, Anatol and Stefan Th. Gries, Stefan (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2): 209 – 243.

Swingley, Daniel (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50: 86 – 132.

Tabossi, Patrizia, Rachele Fanari, and Kinou Wolf (2009). Why are idioms recognized fast? *Memory and Cognition*, 37(4): 529 – 540.

Thiessen, Erik D. and Jenny R. Saffran (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39(4): 706 – 716.

Thompson, Susan P. and Elissa L. Newport (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3(1): 1 – 42.

Tomasello, Michael (2005). Beyond formalities: The case of language acquisition. *The Linguistic Review*, 22: 183 – 197.

Tremblay, Antoine and Harald R. Baayen (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151 – 173). London: The Continuum International Publishing Group.

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

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

Wahl, Alexander (2015). Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics*, 13(1): 191 – 219.

Wible, David, Chin-Hwa Kuo, Meng-Chang Chen, Nai-Lung Tsao, and Tsung-Fu Hung (2006). A computational approach to the discovery and representation of lexical chunks. Paper presented at *TALN 2006*. Leuven, Belgium.

Wray, Alison (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.