STARTING BIG – THE ROLE OF MULTI-WORD PHRASES IN LANGUAGE LEARNING AND USE

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF LINGUISTICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Inbal Arnon

Dec 2009
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Eve Clark, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Daniel Jurafsky

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Michael Ramscar

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Meghan Sumner

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas Wasow

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
ABSTRACT

Why are children better language learners than adults? Intuitively, there is a difference between the unstructured (yet successful) way a child learns language and the effortful and often frustrating experience of trying to master a new language as an adult. On an empirical level, non-native speakers rarely reach native proficiency in pronunciation, morphological and syntactic processing, or in the use of formulaic language and idioms. How we understand these differences is related to our conception of first language learning, and of the object of learning – what does it mean to know language? What is it that children or second learners have to learn?

In this dissertation, I suggest that the answer lies, at least in part, in the linguistic units that children and adults learn from. I propose the Starting Big Hypothesis: children learn from units that are larger and less analyzed than the ones adults learn from. Because they do not know where word-boundaries are (or even that they exist), infants start out with a mix of words, multi-word fragments (e.g. give-it), and short multi-word utterances (e.g. I-don’t-know or It’s-my-turn), and from this early inventory extract linguistic knowledge. This process of using larger units to learn about smaller ones can lead to better learning of certain grammatical and distributional relations between units, especially when the relations are more semantically opaque, as in the case of grammatical gender, or idiom formation. Adults’ prior knowledge and experience with language leads them to ‘break language down’ in a way that might hinder the learning process.

The idea that early units can shape subsequent learning, and that larger units may play a crucial role in achieving native-like proficiency, prompted my investigation into the role of compositional multi-word phrases (units that cross word-boundaries and are semantically and syntactically transparent) in language learning and use. Before we can...
argue that L2 learners make insufficient use of larger units in learning, we have to first show that such units are an integral part of the native speaker’s inventory.

I first highlight the prominence of multi-word chunks (unanalyzed units) in children’s early language use. I then show that multi-word phrases play a facilitative role in children’s production even after segmentation has taken place: for example, 4;6 year-olds produce irregular plurals better after sentence-frames they often occur with (e.g. *Brush your – teeth*) that they do in response to questions (*What are those?*). These findings show that children attend to larger units, and that such units play a role in the process of lexical and morphological development.

Multi-word phrases, I argue, are also part of the adult inventory. In a series of studies, I demonstrate that adult speakers are faster to process higher frequency compositional four-word phrases like *don’t have to worry* compared to lower frequency ones like *don’t have to wait*. Since the phrases were matched for part frequency (the frequency of all the unigrams, bigrams and trigrams) and plausibility, this effect must reflect speakers’ knowledge of the frequencies of the whole-phrases. That is, on some level adult speakers have memory traces of phrases that could otherwise have been generated. This in turn undermines the claim that ‘atomic’ and ‘derive’ forms are represented in a qualitatively different way.

Children’s enhanced word production after familiar frames and adults’ sensitivity to phrase frequency both illustrate the importance of multi-word phrases in native language learning and use. If adult difficulty in learning a second language is related to the more segmented units they learn from, then their learning outcome should improve if they start with larger chunks of language. I test this prediction by looking at adult learning of grammatical gender in an artificial language. Their learning was facilitated when they learned segmentation, semantics (the labels for objects), and grammar (which article a noun appears with) at the same time. These findings demonstrate that (1) early units can influence subsequent learning, and (2) there is an advantage to learning grammar from segmenting larger units.
In examining the size and nature of early linguistic units, I’ve offered a novel perspective on the difficulty that adults experience in learning a second language while at the same time endorsing a usage-based and exemplar-driven approach to first language learning where early experience plays such an important role. My focus on compositional phrases also allowed me to contrast single-system and dual-system approaches to language and to present evidence in support of an emergentist view of language where all linguistic experience (be it atomic or complex) is processed by the same cognitive mechanism. Finally, the Starting Big Hypothesis has theoretical implications for theories of L1 and L2 learning, and practical implications for the teaching of second languages.
ACKNOWLEDGMENTS

When I landed in California, a bit over 5 years ago, I wasn’t sure I would stay for long. It seemed like a different world: far away from my home and my community, a sharp contrast with the communal and politically charged reality I grew up in. What drew me to Stanford in the first place, and what I’ve benefited so greatly from, is the stimulating mix of people and ideas, the non-hierarchical, non-conformist, and familial group that is the Stanford Linguistics department.

Like many of the worthwhile things in life, this dissertation is very much the product of a community. First and foremost my thanks go to my family, my parents Ruth and Arie and my brother Adam whose presence is a constant anchor and inspiration in my life. Your calls and visits made the distance more bearable. Your love and confidence gave me wings to fly. A special thanks to my mom who read multiple drafts and spent many hours helping me find structure and maintain faith in the project.

This dissertation grew out of the Stanford community of linguistics, which provided me a wonderfully stimulating environment to learn in. I came into the program (as Eve often reminds me…) thinking I know what I want to work on. I ended up with a much more complex project that was made possible by the people I learned from and with. Eve Clark has been an inspiring mentor, in all matters academic and beyond. I thank her for numerous discussions, for challenging my ideas when they were not clear, for endless editing, for her encouragement and kindness in every step, and for showing how life and academia can be combined. Dan Jurafsky has been an infinite source of stimulating insight and challenging questions: I could always count on Dan to expose ‘holes’ and inconsistencies in my arguments. His compassion (and brilliance) have made him a wonderful teacher, advisor, and friend. Michael Ramscar has challenged my views on language and learning constantly, exposing me to new ideas and making me ferociously defend the old ones I wanted to keep. Meghan Sumner has been supportive about
everything from designing stimuli to setting up a lab and submitting papers: I love your critical comments! Tom Wasow has patiently read drafts and asked hard questions.

The work in this dissertation benefited enormously from my advisors and intellectual cohort. In particular, the work reported in Chapter 4 could not have been completed without Neal Snider who was an instrumental collaborator from beginning to end. The research in Chapter 3 was done thanks to access to the lovely Bing nursery at Stanford. My thinking about language has been influenced by long debates on grammar with Ivan Sag (as has my time at Stanford, Thank you Ivan for also being a real friend), and my ability to study it was facilitated by the tools taught in Joan Bresnan’s classes. I’ve been blessed to study alongside a close group of students, my intellectual cohort – Neal, Hal, Florian, Lis – I’ve learned as much from you as from anyone else. Here’s to house-sharing across the globe…

These years in California have led me to wonderful people who have touched my life. Itamar, Naama, and Alma who felt like family from the very first day. Jakov who taught me more than I was ready to learn, Fabian who picked up the pieces. Lis and Hal, my train-mates and house-mates, my lovely lovely friends: you made daily life so much better. I would live with you again any time. N+N – I would live with you too (and dance to Hedwig…). Uriel for his insight (in Hebrew…). Daphi, my oldest friend, whose words (and car and phone) were invaluable. Zoe – my co-deliberator, my co-conspirator, whose soul and mind are so full of warmth and light.

While far away, my friends and extended family in Israel have remained a constant source of support, reminding me that I have a home to return to (and participating in all the belated birthday parties…). My love goes to all of you. Adami – for knowing me so well and laughing it all away; Einati - for opening a way back into activism whenever I was back; Nadavi – for music and talks, and last but definitely not least Mayush – for late night calls and perpetual understanding, for being who you are. Some of you, like me, have moved away, creating pockets of belonging across the world (New-York, Chicago). Special thanks go to Daniella who was involved in every stage of this dissertation: who
heard ideas and read abstracts, who listened when I was down, and offered me a unique combination of emotional and intellectual companionship. I couldn’t have asked for a more generous friend.

Finally, my thanks and love go to David, who saw it all: you are my home.
To my parents, Arie and Ruth, whose love, support, and inspirational human spirit have made me who I am.
# Table of Contents

Abstract iv  
Acknowledgments vii  

**Chapter 1: Introduction** 1  
1.1 The Initial Puzzle .................................................................................................................. 1  
1.2 The Starting Big Hypothesis .................................................................................................. 2  
1.2.1 Motivating the Starting Big Hypothesis ........................................................................... 2  
1.3 Defining Multi-word Phrases ............................................................................................... 4  
1.4 The Representational Status of Multi-word Phrases ............................................................ 4  
1.5 The Starting Big Hypothesis - Tenets and Predictions ....................................................... 6  
1.6 Thesis Overview .................................................................................................................. 7  

**Chapter 2: The Units of First Language Learning** 9  
2.1 The Process of Learning..................................................................................................... 9  
2.2 Gestalt Processes in Learning ........................................................................................... 11  
2.3 Multi-word Phrases in Children's Language ...................................................................... 13  
2.3.1 Evidence from perception .......................................................................................... 16  
2.3.2 Evidence from production ......................................................................................... 17  
2.4 Evaluating the Evidence .................................................................................................... 18  

**Chapter 3: The Facilitative Role of Multi-word Phrases in Child Language** 20  
3.1 Learning to Produce Words............................................................................................... 20  
3.1.1 What is in a word? ....................................................................................................... 20  
3.1.2 Re-thinking the single-word focus in production ...................................................... 21  
3.1.3 Learning to produce irregular forms .......................................................................... 23  
3.1.4 The current study ....................................................................................................... 25  
3.2 Experiment 1: Producing irregular plurals in sentence-frames ..................................... 26  
3.2.1 Participants ................................................................................................................ 26  
3.2.2 Materials ................................................................................................................... 26
Chapter 4: Multi-word Phrases in the Adult Lexicon

4.1 The Building Blocks of Language
   4.1.1 Dual-system and single-system approaches
   4.1.2 The representational status of compositional phrases
   4.1.3 Frequency effects and mental representation
   4.1.4 Lessons from morphology
   4.1.5 Whole-form frequency effects for phrases

4.2 The current study
   4.2.1 Previous findings
   4.2.2 The information speakers are sensitive to
   4.2.3 Study goals

4.3 Experiment 3: High and Low Frequency Phrases
   4.3.1 Participants
   4.3.2 Materials
4.3.3 Procedure .................................................................................................................. 63
4.3.4 Results ....................................................................................................................... 64
4.3.5 Discussion .................................................................................................................. 69

4.4 Experiment 4: Mid Frequency Phrases ........................................................................ 70
4.4.1 Participants .............................................................................................................. 70
4.4.2 Materials .................................................................................................................. 70
4.4.3 Procedure ................................................................................................................. 70
4.4.4 Results ....................................................................................................................... 71
4.4.5 Discussion ................................................................................................................. 73

4.5 Meta-Analysis .............................................................................................................. 75
4.5.1 Data and Materials ............................................................................................... 75
4.5.2 Results ....................................................................................................................... 75
4.5.3 Discussion ................................................................................................................. 78

4.6 Discussion .................................................................................................................... 78
4.6.1 Summary of findings ............................................................................................. 78
4.6.2 Phrases in the lexicon ............................................................................................. 79
4.6.2.1 The distinction between 'stored' and 'computed' forms ................................. 81
4.6.3 The information speakers are sensitive to ............................................................ 82
4.6.4 A possible lexical model ....................................................................................... 83
4.6.5 Wholes and parts ................................................................................................. 84
4.6.6 Processing phrases ............................................................................................... 85
4.6.7 Limitations .............................................................................................................. 86
4.6.8 Phrases in parsing and production ...................................................................... 87

4.7 Summary ..................................................................................................................... 87

Chapter 5: Multi-word Phrases in Adult Language Learning ...................................... 89
5.1 Learning Language in Adulthood .............................................................................. 89
5.1.1 Why is it do hard? ................................................................................................. 89
5.1.2 What adults learn from ....................................................................................... 91
5.2 Learning Grammatical Gender - A Case Study ....................................................... 93
5.2.1 Learning patterns in native and non-native speakers...............................93
5.2.2 The units of learning..........................................................94
5.2.3 Insights from learning theory .................................................95
5.2.4 Previous studies ..........................................................................97

5.3 Experiment 5: Training Study ..........................................................98
5.3.1 Participants ............................................................................99
5.3.2 Materials ..............................................................................99
5.3.3 Procedure ...........................................................................100
5.3.4 Results and Discussion .........................................................102

5.4 Discussion ................................................................................104
5.4.1 Summary of findings ..........................................................104
5.4.2 Comparison to previous studies ...........................................106
5.4.3 The effect of early units on learning .......................................107
5.4.4 Other linguistic domains ..........................................................108
5.4.5 Unit size and statistical learning ..........................................110
5.4.6 Order-of-Acquisition effects in language learning ...............111
5.4.7 Starting Big or Starting Small? .............................................112

5.5 Summary .................................................................................115

Chapter 6: Multi-word Phrases in Language Learning and Use 117
6.1 Summary of Findings .................................................................117
6.2 First Language Learning ...............................................................119
6.2.1 The nature of early units ......................................................119
6.2.2 Multi-word chunks ...............................................................122
6.2.3 Units of production or perception? ........................................125
6.2.4 Identifying multi-word chunks in production .......................126
6.2.5 Segmenting early units ........................................................128
6.2.6 Individual differences ..........................................................129
6.2.7 The advantage of Starting Big ..............................................131
6.2.8 Processes in learning ...........................................................132

xiv
6.3 Adult Language Learning

6.3.1 The units adults learn from

6.3.2 The role of multi-word units in learning

6.3.3 Different kinds of L2 learners

6.3.4 Critical period effects in language learning

6.4 The Units of Language

6.4.1 Multi-word phrases in the adult lexicon

6.4.2 Multiple-access models of the lexicon

Chapter 7: Starting Big Revisited

7.1 Tenets and Predictions

7.2 What is Language? Looking Back for Answers

7.3 Future Research

A Experimental Stimuli

A.1 Experiment 3

A.2 Experiment 4

A.3 Experiment 5

References
# List of Tables

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Lexically-Specific frames</td>
<td>28</td>
</tr>
<tr>
<td>3.2</td>
<td>Percentage of each response-type elicited for irregular nouns (computed over all trials)</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>Correct and erroneous plurals for irregular nouns in spontaneous speech</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Mean frequency (per million words) and example items in the High and Low frequency bins</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>Item properties in the three frequency bins (in words per million)</td>
<td>74</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Mean accuracy of irregular plurals in the three prompt conditions</td>
<td>31</td>
</tr>
<tr>
<td>3.2</td>
<td>Mean accuracy of irregular plurals for each item in the three prompt conditions</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Response times averaged over items for High and Low frequency variants in the High frequency bin</td>
<td>66</td>
</tr>
<tr>
<td>4.2</td>
<td>Response times averaged over items for High and Low frequency variants in the Low frequency bin</td>
<td>68</td>
</tr>
<tr>
<td>4.3</td>
<td>Response times averaged over items for High and Low frequency variants in the Mid frequency bin</td>
<td>72</td>
</tr>
<tr>
<td>4.4</td>
<td>Model fit for reaction times to all phrases</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Proportion correct by trial type (article or noun) and learning condition (sequence-first vs. label-first)</td>
<td>102</td>
</tr>
<tr>
<td>5.2</td>
<td>Correct article-noun production in the two learning conditions</td>
<td>104</td>
</tr>
</tbody>
</table>
 CHAPTER 1: INTRODUCTION

1.1 The initial puzzle

Any good research starts with a puzzle, and the one that prompted this dissertation is: Why are children better language learners than adults, despite being worse at a range of other cognitive tasks? Intuitively, there is a difference between the unstructured (yet successful) way a child learns language and the effortful and often frustrating experience of trying to master a new language as an adult. On an empirical level, non-native speakers rarely reach native proficiency in pronunciation, morphological and syntactic processing, or the use of formulaic language and idioms (Birdsong, 1999; De Keysar, 2007; Wray, 2002, and more). The path and end-state of language learning differ in child and adult learners.

An explanation for these differences has been at the heart of many theoretical debates - it bears on how we think about the nature of both language and language learning. To understand what makes language learning more difficult in adulthood, we first need to understand what is involved in first language learning. For this, we need a theory of what it means to know language – what is it that children, or second language learners have to learn? In many ways, the debate over adult difficulty with language learning is also a debate about how first language is learned.

In my dissertation, I explore an answer that is connected to several other, seemingly unrelated, issues. Do children produce words more easily when they are preceded by sentence-frames they often appear with (e.g., teeth in *brush your teeth* or mice in *three blind mice*)? Are compositional multi-word phrases like *don’t have to worry* part of the lexicon? Why is grammatical gender hard for second language learners to master? Does treating articles and nouns as a single unit facilitate adult learning of grammatical gender? What all these questions have in common (and a central component of my
argument) is a focus on the role of multi-word units in first language learning and use, and the different role such units may play in second language learning.

1.2 The Starting Big Hypothesis

Specifically, I propose the Starting Big Hypothesis (SBH). The idea is that children start off with lexical units that are larger and less analyzed than the ones that adults learn from. Children’s early units include ones that cross word boundaries (multi-word phrases) while adults primarily learn from units where word boundaries are already marked. Learning from more segmented units (words, say), I argue, may lead to worse learning of the grammatical relations between them, especially when those relations are not semantically transparent (as in gender agreement between articles and nouns). To put this another way, learning about grammar by analyzing and segmenting larger chunks (e.g., learning article-noun agreement by analyzing noun-phrases) may lead to a more successful outcome than learning by concatenating already analyzed parts (e.g., pelota, ball in Spanish, is a feminine noun and must therefore appear with a feminine article like la or una).

1.2.1 Motivating the Starting Big Hypothesis

This hypothesis stems from several observations about the different ways that children and adults talk and learn. Children sometimes attempt to produce multi-word utterances early on - a child that is mostly producing single words may on occasion produce a longer sequence like ‘how-are-you’ (Peters, 1977; Tomasello, 1992). Many of the more memorable errors they make involve under-segmented or mis-segmented speech, like lead us into Penn station (instead of lead us not into temptation) or I wish I want cookies (where I wish is used as a general attention getter, Peters, 1983). More importantly, children do not know how to segment speech into linguistically meaningful units like words. Infants do not know where lexical word boundaries are, or even that they exist. As a result, their initial lexical units may often include more than one word (Peters, 1983). Children’s early item-based productions, and the gradual development of
more abstract knowledge over time, suggests that grammatical knowledge develops by segmenting and analyzing multi-word phrases (Abbott-Smith & Tomasello, 2006; Tomasello, 2003). This process should create strong links between segmented elements that were part of the same unit early on, like articles and nouns.

Adults, in contrast, probably approach the task of learning language in a more fragmented fashion. They already know that language is segmented into units, and are familiar with kinds of elements (words, morphemes) in their native tongue. Many adult learners (especially those tested in L2 studies) are literate, and have access to another form of input where words are also clearly segmented. In addition, they are often presented with words in isolation in classroom settings, and frequently learn from written input (where word boundaries are visually salient). Even when they hear un-segmented speech, they may listen for parts. One difference between child and adult learners may lie in the adult tendency to ‘look for parts’.

This tendency is evident in the language that L2 learners produce. Their speech is often fraught with blended idioms and inappropriate collocations (e.g., Pawley & Syder, 1980); they treat language as less formulaic than it actually is (Wray, 2002). They seem to lack native-like selection - the ability to use expressions that are not only grammatical but also native-sounding (e.g., non-native speakers may say obey me! Instead of do what I say!, Pawley & Syder, 1980: 197). Reflecting on my own experience as a learner of Spanish and English, I noticed I had persistent difficulty with grammatical gender and idioms, even though they are always the same. A noun always has the same gender in Spanish, and an idiom always has a fixed set of lexical items. And still I would say things like el mano or We hit it off on the wrong note. This difficulty with consistent aspects of language, involving relations between more than one word, is shared by many L2 learners (DeKeysar, 2007), and might reflect overly-active segmentation of the novel language.

Based on the language that child and adult learners produce, the background knowledge they bring to the task of language learning, and the areas in which they have particular difficulty, I suggest that there is a fundamental difference in the units that
children and adults learn from, and that this difference may lead to poorer learning outcomes in adult learners.

But before I can argue that L2 learners make less use of multi-word phrases in learning, I first have to show (beyond anecdotal evidence) that such units are an integral part of the native speakers’ inventory. The Starting Big Hypothesis is also a hypothesis about the range of units that language users attend to and use. For children to learn from units larger than words, they need to represent them, so multi-word phrases have to be part of the inventory of units that children acquire. If children draw on such units in learning, what prevents adults doing the same? The question can now be cast in terms of the linguistic material that language users represent and make use of. I approach this question by looking at the status of compositional multi-word phrases for speakers.

1.3 Defining Multi-word Phrases

I use the term multi-word phrase to refer to a linguistic sequence made up of at least two orthographic words, that is not idiomatic in form or meaning, and that is a syntactic constituent. Examples include article noun combinations (the dog, those kids), verb phrases (catch the ball), and complete utterances (It’s my turn). These units should not be confused with multi-word expressions as studied in the syntactic and computational literature (e.g. Moon, 1998; Sag et al., 2002; Sinclair, 1991) where the term is used to refer to “idiosyncratic interpretations that cross word boundaries or spaces” (Jackendoff, 1997, adopted by Sag et al., 2001). In contrast, the units I’m interested in are compositional in nature - their syntactic form and their semantic meaning can be derived in a relatively straightforward fashion from their parts.

1.4 The Representational Status of Multi-word Phrases

Many theories of language posit a distinction between the mental lexicon and the mental grammar (e.g., Chomsky, 1981; Pinker, 1991, 1999), the mental lexicon contains linguistic units that cannot be generated (atomic elements) like non-inflected words, and irregular nouns and verbs. The mental grammar is used to derive complex forms from
these atomic elements. Multi-word phrases are only stored when they are non-compositional; when their form or meaning cannot be transparently deduced from their parts (as in the case of certain idioms, Pinker, 1999). Compositional multi-word phrases (like article-noun combinations, or verb-phrases) are good examples of rule-generated units, but are not in themselves seen as building blocks of any kind.

Over the past years, this distinction, between ‘the mental lexicon’ and ‘the mental grammar’, has come under attack from several directions: theories of grammar that don’t differentiate between words and more abstract entities like constructions (Goldberg, 2006); models of processing where all linguistic material is processed by the same cognitive mechanism (e.g., Rumelhart & McClelland, 1986); empirical evidence that linguistic knowledge incorporates fine-grained distributional information on many levels of linguistic analysis; and the formulation of learning models where multi-word chunks play an important role.

A growing number of language models assume that the mental lexicon contains patterns of varying sizes and levels of abstraction (words, but also multi-word phrases and constructions). I label this class of models ‘emergentist’ and group together under this label connectionist, usage-based and exemplar models (Bybee, 1998, 2001; Goldberg, 2006; Langacker, 1987; Bod, 1998, 2006; Johnson, 1997; Goldinger, 1996; Pierrehumbert, 2001; McClelland & Rumelhart, 1986). In lieu of ‘atomic elements’ and ‘rules’, such models propose that all linguistic experience is processed in a uniform fashion. Multi-word phrases are just like any other linguistic unit, affected by the same factors that affect the processing of smaller elements like words.

Numerous studies show that language users are sensitive to distributional information at many grain-sizes – from sound combinations (Hay and Baayen, 2005), to words and two-word sequences (Fidelholtz, 1975; Jurafsky et al. 2001), through larger constructions (Bock, 1986), and discourse structures (Elman, 2009). These findings highlight the psychological reality of patterns of varying sizes. Language use isn’t affected only by abstract categories (though those also play a role) but also by fine-grained details of occurrence. If multi-word phrases are processed like any other linguistic pattern, then we should find a similar sensitivity to their frequency. Finally, multi-word phrases play a
central role in recent usage-based and exemplar models of language development (Tomasello, 2003; Abbot-Smith & Tomasello, 2006; Bod, 1998; 2009). In such models, lexical and grammatical knowledge is learned by abstracting over stored utterances, including multi-word ones.

There are now multiple theories and models predicting that multi-word phrases are psychologically real linguistic units for speakers. But there is relatively little research that targets such units. At the adult level, much of the research on processing effects focuses on words, the units that makeup words (phonemes and morphemes), and the larger constructions they appear in. Less work has looked to see whether processing is also affected by the properties of multi-word phrases. At the child level, very few studies have asked whether children store such units (e.g., Bannard & Matthews, 2008) or use them in learning. So there is little evidence to argue for, or against, the psychological reality of multi-word phrases. Whether language users represent such units though, has implications for the process and product of language learning. Do language users represent units they could otherwise generate? Multi-word phrases, like the ones I investigate in this dissertation, serve as a way to contrast and test different views of what it means to learn and know language.

1.5 The Starting Big Hypothesis: Tenets and Predictions

Within this context, I can formulate the SBH as following:

1. Children learn by segmenting and analyzing stored utterances, including multi-word ones.
2. Such units play a facilitative role in children’s language.
3. Adult speakers store information about compositional forms, including multi-word ones.
4. Adults learn from already analyzed units, and this negatively influences the learning process.

This hypothesis is somewhat controversial from several perspectives. First, it assumes that children can and do store units that cross word boundaries. Second, it assumes that
children also learn from larger units, contra the more common assumption that acquisition progresses from smaller units to larger ones (e.g., syllables to words to sentences). Third, it implies that multi-word phrases, despite being easy to generate, are units that language users represent. Finally, it suggests that learning about smaller units from larger ones is beneficial, and is in fact one of the difference in how children and adults learn language. This claim seems to contrast with previous proposals about how early unit size affects children’s learning as in Newport’s Less is More (1990) or Elman’s Starting Small (1993).

The dissertation accomplishes three goals. First, I provide evidence for the psychological reality of multi-word phrases in child and adult native-speakers. I do this by asking whether children (chapter 3) and adults (chapter 4) show evidence of representing such units. Second, I show that multi-word phrases have a facilitative role in children’s language use. I do this by showing that children’s production of irregular plurals is enhanced in familiar sentence-frames (chapter 3). Third, my most ambitious goal, I test the novel prediction that some adult difficulty in L2 learning is related to learning from smaller already-analyzed units (chapter 5). I test this prediction by asking whether adult learning of grammatical gender improves if adults are first exposed to ‘less analyzed’ chunks of language in the novel language.

1.6 Thesis Overview

The dissertation combines findings from child language (naturalistic and experimental), adult online processing, and adult artificial language learning. It is structured as follows. In the next chapter I review existing evidence for two of the assumptions of the Starting Big Hypothesis - that children can store multi-word phrases and that they draw on larger units in learning about smaller ones. Chapter 3 starts the empirical part of the dissertation. In it, I test the prediction that children represent multi-word phrases and that such units play a facilitative role in children’s language. I focus on the acquisition of irregular nouns in English. In chapter 4, I investigate the status of multi-word phrases in the adult lexicon by asking whether adult recognition times are affected by four-word phrase frequency. In
chapter 5, I ask whether L2 learning is facilitated when learners are first exposed to larger chunks of language. To do this I look at learning of grammatical gender in an artificial language. All these questions were carefully chosen; all have theoretical significance beyond the Starting Big hypothesis, specifically, they relate to the acquisition of irregular forms (chapter 3), the structure of the lexicon (chapter 4), and learning grammatical gender (chapter 5). In chapters 6 and 7 of the dissertation, I revisit the Starting Big Hypothesis and present a unified research program that highlights the role of multi-word phrases in child learning, adult processing, and L1/L2 differences.
CHAPTER 2: THE UNITS OF FIRST LANGUAGE LEARNING

The Starting Big hypothesis as formulated in the introduction makes two predictions about first language learning. It predicts that children can and do store multi-word phrases, and it predicts that such units are utilized in the process of learning - that is, children use larger units to learn about smaller ones. Neither prediction is commonly made in the literature. Whether multi-word phrases play a role in learning is closely tied to theoretical positions on the process and product of learning. In the next section, I discuss support for both predictions: (a) the idea that the process of learning involves learning from larger units, and (b) the idea that children can and do store multi-word phrases. I conclude with a review of the previous literature on the role of multi-word phrases in child language.

2.1 The Process of Learning

A text-book description of how production develops in first language acquisition (e.g., Bloom, 1994; Clark, 2003; Goodluck, 1991) often goes from babbling (producing syllables), through first word formation, to the production of multi-word utterances and finally complex sentences. Children’s progression is described as a move between several consecutive stages, each involving larger and more structured linguistic units: from the syllable, to the word, to the construction. This description emphasizes the combinatorial, or small-to-big, direction of learning, and hence the ability to create larger and more complex units from smaller ones: combining syllables into words, deriving morphologically complex words from words and morphemes, and generating larger phrases and sentences from single words.

This description captures the ‘generative’ aspect of language. But it masks what may be an equally important process: the move from larger unanalyzed units to smaller more structured units. I will refer to this class of processes as Gestalt processes (after Peters,
1977), or big-to-small. I use this expression for processes where the child learns from an unanalyzed whole about the parts, for example, using words to identify phonetic contrasts of a language, or learning from a formulaic chunk like how-are-you about the words it is made up from. Such processes are similar to combinatorial ones in that the product of learning is a more complex and structured representation. But they are different in the units they operate over. Children can learn from a range of units, not only atomic ones. Their early units do not necessarily correspond to adult ones, or to the linguists’ units of analysis.

The significance of Gestalt processes in language learning is highlighted by the continuous nature of speech. The infant doesn’t hear linguistic input neatly separated into phonemes, morphemes, and words. To discover such units, children first need to break into the speech stream and identify (some of) the relevant units in their language (e.g., Goldinger & Azuma, 2003; Kuhl, 2004). This process necessarily involves decomposing larger units (e.g., fragments of un-segmented speech) into smaller more structured ones (e.g., words, morphemes). Early speech segmentation, probably the first task an infant is faced with, involves a gestalt process. Yet gestalt processes in language learning have gone largely unstudied.

One reason is that such processes fall by the wayside in theories that emphasize the rule-governed aspects of language. When linguistic knowledge is viewed as consisting of atomic elements and rules to combine them (Chomsky, 1965; Pinker, 1999; Jackendoff, 2002), then these are also the elements that researchers look for in children’s language. A words-and-rules characterization of adult language tends to focus on (a) combinatorial properties of child speech, and (b) adult-like units - phonemes, morphemes, words - in children’s speech. Theories of grammar and development that emphasized the centrality of rules in language (e.g., Borer & Wexler, 1987; Pinker, 1991) led to a focus on the combinatorial aspect of children’s language learning.

However, more theories of grammar acknowledge the role of larger patterns, like constructions, in language (Croft, 2001; Fillmore, 1988; Goldberg, 1995, 2006; Kay & Fillmore, 1999; Pollard & Sag, 1994). Linguists now view a larger set of units as potential building blocks in a language (e.g., Bybee, 1998, 2001). Within developmental
research, there is increasing support for usage-based theories of learning where input plays a crucial role (Tomasello, 2003), and where abstract categories and rules are derived from larger stored chunks of language. These views of language structure and learning provide more of a theoretical anchor for understanding Gestalt processes, and more of an incentive to seek and study them.

2.2 Gestalt Processes in Learning

There is evidence for Gestalt processes on many levels of linguistic analysis. As mentioned earlier, segmenting words from speech necessarily involves a big-to-small process. How infants achieve this task has been the focus of much research (e.g., Saffran et al. 1996; Jusczyk, 1999). One way to solve it is by using the transitional probabilities between segments and syllables as a cue to word segmentation. The probabilities within words are higher than those between words. While transitional probabilities are not described in Gestalt terms, their use in language does reflect a Gestalt process. For transitional probabilities to be meaningful they must be calculated over strings of consecutive words. This is applied implicitly in studies that test human ability to make use of this information - infants, children, and adults (see Saffran 2003) are always exposed to unsegmented sequences of words.

Peters (1977) described one child’s use of what I will call the whole-utterance path to identifying words. Instead of going from single words to longer utterances, the child’s early communicative attempts were comprised of adult-like prosodic contours that were only later segmented into recognizable words. For instance, at 1;7 the child produced the sequence [A lər ri go mu nay] (interpreted by Peters as ‘I like read good moon night’) while holding the book ‘Good Night Moon’. Only later did the individual words appear in other utterances.

The idea of whole-word phonology, first formulated by Ferguson & Farwell, 1975 (see also Macken, 1979; Waterson, 1971), is another instance of a Gestalt approach to learning. Here, lexical words form the basis for the early learning of the segmental inventory of a language. Children start out with whole word forms from which they
extract information about the relevant phonological contrasts. As children’s grasp of the phonological system grows, their production of early words may become less accurate. As an example, Ferguson & Farwell look at the trajectory of the word ‘pretty’ in one child. The child first produced the word at ten months in “almost perfect phonetic form” (p. 432), at a stage where most of his other words were monosyllabic and lacked consonant clusters. At 1;9, as he builds up phonological classes and patterns, the phonetic form of ‘pretty’ changes to [piti], less accurate but a better reflection of the state of his phonological system at the time (e.g., with CV syllables dominant). This pattern of early phonetic accuracy followed by less accurate but system-internally consistent forms for children’s words has been observed in other studies (e.g., McCune & Vihman, 2001).

Vihman and her colleagues also emphasize of the big-to-small direction of phonological acquisition with phonological templates (Velleman & Vihman, 2002; Vihman, 1996). These templates are extracted from words acquired early. They then serve as a ‘filter’ through which new words are acquired. Children gravitate towards words that match their existing templates. Both Ferguson & Farwell and Vihman argue for a whole-word path to segments.

In a recent study, Swingley (2005) suggested that children can learn about the dominant stress pattern in their language from early word forms. By 8 months, English-speaking infants exhibit a trochaic bias in speech segmentation. They prefer to group together syllables that have a strong-weak stress pattern (trochaic) rather than syllables with a weak-strong pattern (iambic). How does this preference emerge? Swingley suggested that infants accumulate potential word-forms by grouping together clusters of syllables that occur together often and have a high conditional probability. He showed, with a computational model, that the majority of the potential word-forms have trochaic stress. Infants can use the early word forms to extract a dominant stress pattern that can then be used in the segmentation of novel words. Here too, children extract information from early, unanalyzed forms.

The view that children learn from unanalyzed wholes is also found in morphological acquisition. Children may learn inflections in a rote-based fashion. They start with inflected wholes and only later learn the morphological components. In support of this
idea, Wilson (2003) showed how the frequency of copula omission errors (as in he sleeping) varies according to the specific subject-type. Omission was more likely with lexical NPs than with pronominal forms, and more likely with lower frequency pronouns. Children’s acquisition of subject-verb agreement in Portuguese (Rubino & Pine, 1998) shows a similar pattern. Omission rates vary with the number/person of the subject nominal, going from zero errors for the more frequent first-person subjects to 43% errors with the less frequent third-person-plural subjects. These findings illustrate a whole-word path to morphological acquisition.

Studies within the usage-based tradition also imply a whole-chunk path to construction learning. They document the apparently item-based nature of children’s early utterances, compared to their later more abstract and productive knowledge of the relevant structures (Tomasello, 2000). For example, in transitive constructions, children seem to start with verb-specific uses and representations before they develop more abstract, verb-general ones (e.g., Abbot-Smith, Lieven & Tomasello, 2007; Akhtar, 1999; Tomasello, Brooks, & Stern, 1998). Similar arguments have been made for the acquisition of relative clauses (e.g., Diessel & Tomasello, 2000) and passives (Tomasello et al., 1998). Item-based productions appear not only with specific verbs, but also with specific noun-phrases, indicating the reliance on unanalyzed chunks: many early transitive utterances have the form I’m verb-ing it (Lieven, Pine & Rowland, 1998). In Gestalt terms, children’s construction learning seems to involve a big-to-small process: A move from unanalyzed chunks to more abstract constructions.

These findings suggest that Gestalt processes show up in learning on many levels of linguistic analysis. We can identify Big-to-Small processes in learning about phonemes, morphemes, words, and constructions. Quite a lot of language learning can be described as a move from larger, unanalyzed units to smaller more structured ones.

2.3 Multi-word Phrases in Children’s Language

Of particular interest here are processes that involve multi-word phrases. The role of multi-word phrases in learning was first highlighted by Peters in the late 70’s (Peters,
She reported on what she called Gestalt utterances in one child’s early speech. These utterances had the intonational contour of an adult phrase but did not contain identifiable ‘word’ units. At age 1;2, for example, the child produced /obe-da-do/ while banging on the bathroom door (to mean open the door). At this point, though, this child had only 10 recognizable words in his productive vocabulary, not including either door or open. Instead of producing the utterance by combining known words, the child seemed to be producing an unanalyzed, or under-analyzed, multi-word chunk.

Building on the case-study of this child together with earlier observational and diary studies, Peters (1983, 1985) presented a theory of language acquisition that emphasized the role of larger, unanalyzed units in the learning process. She suggested that children’s early units do not necessarily correspond to words or morphemes, and frequently consist of more than a word. Over time, the child segments these frozen or formulaic units using a set of strategies like SG:BEGIN (segment out the first syllable of a unit) or SG:STRESS (segment out a stressed syllable). In segmenting these early units from their surroundings, the child acquires structural information in the form of syntactic frames and slots. This insightful account of children’s early units diverged from the ‘dogmas’ of the time in three respects: (1) in suggesting that child units may diverge from adult ones, (2) in highlighting the role of units that cross word boundaries, and (3) in emphasizing the formulaic pre-fabricated nature of children’s input (adult language) and output (their own productions).

To support her proposals, Peters drew on many examples of early speech from other researchers. For example, she used the following dialogue, taken from Ruth Clark (1977) to show how children may extract the wrong-sized unit (also the wrong meaning in this case…). The participants in the exchange are Clark’s son, Adam (aged 2;3) and his caretaker.

(1) Adult: That’s an elephant, isn’t it? What is it?
   Adam: Intit
Adam treats ‘isn’t it’ as one unit and erroneously uses it to refer to an elephant for several weeks. Another example shows how children may extract syntactic frames from larger units. At 1;8, when his utterances were at most two words long, Satoshi learned to say what a nice bicycle you have. Soon after, he started producing utterances like what a nice elbow you have and what a nice Daddy you have presumably based on the frame what a nice X you have (Peters, 1983: p. 51).

These examples, and the others in the book, offer important insights into some aspects of children’s early language use, but it is hard to assess from them how commonly children produce multi-word phrases. While Peters talked about the processes that underlie children’s extraction and segmentation of units, (though without systematic empirical support), she did not address the role that such units might play in learning. That is, she suggested that multi-word phrases participate in learning process but did not argue that learning benefited from their presence.

Few studies have followed up on Peters’ insights, and few have addressed the theoretical and practical implications for the process of learning. This was probably driven, at least in part, by the fact that such units were not compatible with the dominant theories of grammar, or grammar learning at the time (as mentioned in the previous paragraph). As the theoretical climate changed, so did the space allocated for multi-word phrases in theories of learning.

Multi-word phrases do play a role in recent usage-based and exemplar models of language development (Tomasello, 2003; Abbot-Smith & Tomasello, 2006; Bod, 1998; 2009). In such models, both lexical and grammatical knowledge is learned by segmenting and analyzing larger units. Children start off with signs – non-abstract, lexically-realized linguistic patterns. These can be single words (e.g., mommy) or multi-word utterances (e.g., what is that?). These units are abstracted over to create schemas – partially realized frames with more abstract slots (e.g., what is X?), which eventually give rise to more abstract knowledge (Clark, 2003; Lieven, Pine & Baldwin, 1997; Lieven & Tomasello, 2008; Lieven et al. 2003). Multi-word phrases play a crucial role in such models - they are the units from which children extract grammatical knowledge. Such models echo Peters’ observation that children’s early units do not correspond to adult words. But they
add the prediction that abstract knowledge is learned from them. The validity of these models rests on the assumptions that (a) children can and do store multi-word phrases, and (b) that they use such units in learning.

2.3.1 Evidence from perception

On the perceptual side, there is evidence that infants can remember multi-word phrases early on. Infants as young as two-months-old remembered speech better when it was packaged in clausal units, as opposed to comparable, but non-clausal word sequences (like a word list, or a sequence that spanned clause boundaries, Mandel, Jusczyk & Kemler Nelson, 1994; Mandel, Kemler Nelson & Jusczyk, 1996). Two-month-olds could detect the subtle phonetic difference between the trained sequence *cats like park benches* and the test sequence *rats like park benches* when the former was spoken as one clause. But they could not detect the same difference when the same sequence was spoken with a list intonation (*cats, like, park, benches*) or when they corresponded to two phrasal units (*Cats like. Park benches*) in training.

This effect was not limited to sequences presented in isolation. Infants show better recognition of previously heard sequences embedded in fluent speech when they corresponded to a well-formed prosodic clause. Nazzi et al. (2000) familiarized 6-month-old infants with two versions of the same word-sequence: one that corresponded to a well-formed prosodic clause (*Rabbits eat leafy vegetables*) and one that crossed clause boundaries (*Rabbits eat. Leafy vegetables*). Despite equal exposure, infants listened longer to passages containing the well-formed clause. At the same age, infants’ showed better memory for well-formed prosodic clauses when they were familiarized with a multi-clause passage and then tested on an isolated sequence (Soderstrom, Kemler Nelson & Jusczyk, 2005). That is, infants not only detect clausal units, they also rely on such units in remembering and organizing speech. These findings show that infants remember chunks of language larger than words, and rely on such chunks in speech perception.
2.3.2 Evidence from production

Several findings for production support the claim that at least some units children start out with have the form of unanalyzed chunks that cross word boundaries. Peters reported examples of early multi-word chunks. For instance, the child studied in Horgan (1980), age 2;0, produced utterances like how about open more presents and how about what do you want to eat where how about functioned as a semantically bleached carrier phrase for requesting. Adam, at age 2;3, used I carry you when he wants to be carried by a grown-up (Clark, 1974). Tomasello (1992) noted that children use ‘frozen’ multi-word phrases like how-are-you or wat-iz-dat at a stage where most of their other productions consist of single words. Clark (1974) drew attention to “copied utterances that remained intact for several weeks without other lexical items being substituted” (p. 4). She noted her son’s use of wait for it to cool which he produced for a few weeks whenever a hot meal was brought to the table.

In a more recent study, Brandt et al. (2009) report evidence for chunked knowledge of certain complement-taking phrases like I think or I believe by showing that 4-year-olds have difficulty changing a first-person prompt (e.g., I think) into a third-person one (e.g., he thinks) for highly frequent collocations. Rowland (2007) showed how children’s inversion errors on questions can be partially traced back to high-frequency lexical strings in the input. Children were less likely to make inversion errors for strings that appeared inverted frequently in the input like What can X. Children’s correct productions may reflect knowledge of ‘chunked’ item-based constructions (Dabrowska, 2000; Rowland & Pine, 2000). Findings like these highlight the chunked aspect of early language – only some multi-word productions are fully analyzed and hence productive.

So far, there is mostly indirect evidence (the perceptual findings) or sporadic evidence (examples of chunked language) for children’s use of multi-word phrases. Only recently has there been more systematic investigation of children’s knowledge of multi-word phrases. Bannard & Matthews (2008) compared two- and three-year-olds’ repetitions of four-word phrases like a cup of tea and a cup of milk. These phrases differ in chunk frequency but are matched for part frequency (the frequency of the words,
bigrams, and trigrams within the phrases). Children’s productions of the identical first three words was faster and more accurate when these words appeared as part of a more frequent chunk than when they were part of a less frequent chunk. This suggests that children represent the entire phrase as a unit.

2.4 Evaluating the evidence

The findings I have reviewed here document the existence of big-to-small processes in language learning. While such processes have been traditionally under-studied, they are not rare. They appear in various linguistic domains including the learning of segmental phonology, lexical stress, and constructions. This part of the Starting Big Hypothesis seems to have a plausible basis. The findings also suggest that children can store multi-word phrases, a crucial prediction of the Starting Big Hypothesis. What remains unclear is how pervasive such chunks are and what their role in learning is.

In the next chapter, I explore the status of multi-word phrases in children’s language by asking if such units can enhance the production of irregular plurals, forms that are hard for children to master. Specifically, I ask if children produce irregular plurals better following familiar multi-word frames. This investigation has several goals. The first is to support previous findings (Bannard & Matthews, 2008) that children represent such units. The second is to establish that multi-word phrases play a facilitative role in children’s language, that is, such units can enhance the production of single words, even after segmentation has taken place. Such a finding could tell us something about the units children learn – both words and the larger patterns they appear in, and about the units they draw on in production. It would indicate that children use what they know about larger patterns in the production of smaller ones. This suggests that knowledge about words may be tied to the larger patterns those words appear in. Finally, if word production is affected by the linguistic context words are produced in, this could influence how we assess and characterize children’s lexical and morphological acquisition.

The prediction that multi-word phrases play an important role in learning is not easy to test. First, it is hard to identify children’s early units. Studies with pre-linguistic infants
show that they are capable of attending to multi-word units, but they cannot tell us about
the units they actually store in memory. It is even harder to document the process of
learning. How can we tell that a child’s correct use of a word originated from an
unanalyzed chunk? By the time they start producing words (between 12 and 22 months),
children are clearly capable of segmenting phrases into words (in fact they show this
ability in a rudimentary form as early as 7.5 months, Jusczyk & Aslin, 1995). But this
doesn’t mean that they don’t rely on larger chunks as well. By asking whether word
production is affected by the linguistic context that specific words appear in, we can learn
about the process involved (learning about smaller units by analyzing larger ones)
through a later symptom (relying on larger patterns in the production of smaller ones).
Finding that multi-word phrase can enhance word production is a first step in exploring
the role that larger units, namely multi-word phrases, play in the acquisition of smaller
ones (namely words).
3.1 Learning to Produce Words

In the previous chapter I reviewed various big-to-small processes (Gestalt processes) in language learning. I showed that while such processes have been traditionally understudied, they are not at all rare. There is evidence for them in various linguistic domains including the learning of segmental phonology, lexical stress, and constructions. Of particular interest to the Starting Big Hypothesis are Gestalt processes that involve units that cross word boundaries (multi-word units). In the current chapter, I ask whether multi-word phrases play a facilitative role in children’s language by looking at the production of irregular plurals in English.

3.1.1 What is in a word?

One of my colleagues recently told me the following story about her 5-year-old daughter. The family had recently arrived in an English-speaking country where the child was first exposed to English. After a few weeks in pre-school, she came home and asked her mother: Mom, what does the word ‘are-you-kay’ mean? the teacher uses it every time a kid falls down (the question was asked in Hebrew and the child used the term word to describe the utterance). This question highlights several interesting aspects of word learning in particular, and language learning in general. First, that unlike written text, or linguistic analysis, natural running speech does not provide a clear guide to different kinds of linguistic units (words, phonemes, morphemes). This child identified a phrase here as a word. Second, certain words tend to co-occur together, making the phrase more prone to identification as a single unit.

Despite the clear relation that words have to other words and phrases, much of the research on children’s lexical and morphological learning has largely ignored the role of
the linguistic context in children’s learning and production of words. Children’s production abilities are often assessed by eliciting single words in isolation. This is true for studies that assess lexical knowledge (e.g., naming objects or pictures) as well as morphological knowledge (e.g., producing a plural for a nonce word; Berko, 1958). Corpus studies also tend to focus on the properties of single-words (for example, their frequencies in child-directed speech).

Such studies have been invaluable in mapping out various factors (e.g., frequency, phonological form) that influence children’s learning. But they leave unexplored questions about the potential role of linguistic context on the learning and use of words. Underlying this single-word focus is an implicit assumption that lexical and morphological knowledge is independent of any larger contexts that words appear in. That is, that whatever children know about the correct label for an object, or the correct plural form of a noun, say, it is (a) constant across linguistic contexts, and (b) independent of their knowledge about larger patterns. The word is construed as the largest relevant unit of investigation for asking questions about lexical and morphological acquisition.

3.1.2 Re-thinking the single-word focus in production

Yet there are several reasons to think that larger linguistic units influence the way words are learned and used. First, children rarely hear words in isolation. Only around 10% of parental utterances to young children involve single-words alone (Aslin, Woodward, La Mendola & Bever, 1996; Brent & Siskind, 2001). While those utterances may play some role in assisting segmentation and learning new vocabulary items (Brent & Siskind, 2001, but see Aslin et al. 1996), children’s exposure in general consists of words embedded in linguistic contexts.

Moreover, the linguistic contexts for specific words are often repeated. In an extensive corpus study, Cameron-Faulkner, Lieven, and Tomasello (2003) showed that the linguistic frames that children are exposed to are highly repetitive. Half of all the child-directed speech (CDS) they analyzed starts with one of 52 simple frames, such as *Look at the ___?* or *Where’s the ___?* Of the remaining frames, many were repeated in
daily routines (feeding, bathing, dressing, etc.), making them frequent before specific nouns (e.g. Wash your____ before hands, Have a____ before bath). Frames like these are also used very consistently when adults introduce children to words that are unfamiliar, e.g., That’s a —, Those are —, This is called a — (Aslin et al. 1996; Clark & Wong, 2002; Estigarribia & Clark, 2007). Based on an analysis of child-directed speech, Mintz and his colleagues (2002; Mintz, 2003) further suggest that recurring frames for a single word may assist learners in extracting information about syntactic categories. Words that appear in similar frames tend to belong to the same syntactic category.

Studies of child-directed speech show that children mostly hear words in frames, and that those frames are frequent. Recent research shows that children attend to these frames, and that their own production is influenced by them (e.g., Bannard & Matthews, 2008). Two- and three-year-olds are more accurate and less disfluent when they are asked to repeat higher frequency four-word chunks (e.g., a drink of milk) compared to lower frequency ones (e.g., a drink of tea). These findings suggest that the repetitive nature of children’s linguistic experience has implications for their learning of language. Children produce more frequent chunks with greater precision and ease. Children attend not only to the general frequency of chunks but also to the frequency of specific frames (e.g., a drink of——) combined with particular words (e.g., milk vs. tea).

The findings reviewed so far show that children’s experience with words is to a large degree embedded in linguistic contexts, and that children are sensitive to such contexts. A recent study by Fernald & Hurtado (2006) provides a stronger indication that sentence frames play an important role in the use of words. They used a looking-while listening paradigm (where children listen to speech while looking at pictures of objects) to show that children’s object recognition is better when the word appears after a familiar frame frequent in child-directed speech. Eighteen-month-old infants were faster to orient to the correct picture when the word followed a frequent frame like Look at the --- compared to the same word presented either in isolation (Doggie!) or following a semantically equivalent, but infrequent, frame like Look! The ---. They attribute this effect to the greater ease of processing a familiar frame, which frees up resources for the processing and recognition of the final word.
The prevalence of recurring frames in children’s input, and the growing evidence that children rely on such units in production and recognition suggests that lexical and morphological knowledge may not be as context-independent as seems to have been assumed. Instead, children’s accuracy in producing single-words may be affected by the linguistic contexts they are learned and used in, reflecting (a) their knowledge of larger patterns, and (b) their reliance on such patterns in online production. If children learn about the larger patterns that words appear in, and rely on such knowledge in production, then word production may well vary in accuracy depending on the linguistic context of elicitation. To test this prediction, I looked at the effect of linguistic context on children’s production of irregular plurals in English.

3.1.3 Learning to produce irregular forms

Learning to produce morphologically inflected words such as plurals or past-tense verbs is not an easy task. It requires learning many components, among them the correct syntactic, semantic, and pragmatic distributions of the relevant morphemes, the functions of specific forms, and the regularities and irregularities within each paradigm in the morphological system. Such words, with the idiosyncrasies and regularities they display, provide a rich testing ground for theories of language learning. Studying the factors that influence children’s production of such words gives us a window onto the way their linguistic knowledge is learned, organized and accessed.

Ever since Berko’s (1958) study, where children were asked to pluralize nonce words (e.g. wug), children’s plural formation (with both real and nonce words) has been used to address a variety of theoretical issues: the degree of productivity of early morphological knowledge (Berko, 1958; Maratsos & Chalkley, 1980), the role of word frequency (Maslen, Theakston, Lieven & Tomasello, 2004), whether forms are rule-based or rote-learned (Berko, 1958; Marcus et al., 1992; Marcus, 1995; Pinker, 1991), and whether the acquisition of irregular morphology is better captured using single-system (Rumelhart & McClelland, 1986; Plunkett & Marchman, 1993) or dual-system models (Clahsen, 1999; Pinker & Prince, 1992; Pinker & Ullman, 2002; Ullman, 2001)
Irregular plurals have long been known to cause difficulty for children acquiring English (Berko, 1958; Cazden, 1968). Although children make sporadic correct use of them quite early on, they take several years to establish, and children’s uses tend to waver between over-regularized forms and the correct irregular forms (e.g., Cazden, 1968; Marchman, Plunkett, & Goodman, 1997; see also Kuczaj, 1977; Slobin, 1978). Children often produce over-regularized forms like ‘mouses’ or ‘foots’, with the regular plural inflection ‘-s’ used to create a plural for an irregular form. These errors are found in both spontaneous speech (Marcus et al. 1992; Maslen et al. 2004) and experimental elicitations (Arnon & Clark, 2008; Matthews & Theakston, 2006; Ramscar & Yarlett, 2007), and they have fueled heated debate about their source and significance in understanding language development. Why do children make such errors when these forms are not present in their input? There are two common answers. Under a rule-based explanation, errors reflect the application of the default regular rule (Marcus et al., 1992; Prasada & Pinker, 1993; Ullman, 2001). Activation-based accounts attribute errors to competition from the more frequent overall regular pattern (Rumelhart & McClelland, 1986; Plunkett & Marchman, 1993). I return to this debate in the general discussion.

This extensive body of research has been primarily concerned with how features of the words and their smaller components (e.g., simple words and morphemes) affect production. Like the general pattern described above, the study of irregular plurals has also had a single-word focus evident in the methods used to study children’s production. In experimental studies, children are asked to produce words in isolation, or in a linguistic environment assumed to be ‘neutral’ and held constant across all test items. For example, children are asked to produce a plural form by completing a phrase like “Now there are two ——” (Berko, 1958). Other studies have generally followed similar methods, using pictures together with neutral elicitation questions (e.g., Arnon & Clark, 2008; Graves & Koziol, 1971; Matthews & Theakston, 2006; Miller & Ervin, 1964; Ramscar & Yarlett, 2007). The same prompt is used with all the items, as if the linguistic context per se does not affect children’s accuracy in production.

Corpus studies have generally had a similar single-word emphasis. It is easier to control for multiple variables simultaneously in corpus studies compared to experiments.
But here too, researchers have focused on properties of the words or morphemes, not on the larger linguistic contexts these forms appear in (Maslen et al., 2004; Marcus et al., 1992). For example, Maslen et al. (2004) show that children’s production of irregular nouns and verbs is influenced by multiple factors – token frequency, type frequency, type/token ratio. But none of them relate to the linguistic contexts these words are produced in.

3.1.4 The current study

Irregular plurals make a good test case for looking at the role of linguistic context on the production of single words. Children’s correct performance is often poor and leaves much room for improvement: 4;6 year-olds produce irregular plurals correctly only about 30% of the time in experimental settings (Arnon & Clark, 2008; Ramscar & Yarlett, 2007). Previous research has shown that error-rate is affected by properties of the word. Children make fewer errors on more frequent nouns (Maslen et al. 2004). For example, they make fewer errors on the more frequent *teeth* compared to the less frequent *mice*. Errors are also affected by phonological properties – children make more errors when the singular noun ending resembles the plural inflection (nouns ending with sibilants like /iz/), when it has fewer phonological neighbors (Bybee & Slobin, 1982), and when it is similar in form to regularly inflected nouns (Matthews & Theaskton, 2006) or plural schemas (Köpcke, 1998). Inflection is also affected by semantic factors – nonce verbs are more likely to be inflected regularly if they resemble a regularly inflected verb in meaning (Ramscar, 2002). But no study to date has looked at the effect of linguistic context on error rate. The main goal of this chapter is to investigate the effect of linguistic context on word production. Focusing on irregular plurals has the added advantage of contributing to the theoretical debate on the underlying source of over-regularization. I will elaborate more on this in the general discussion.

I compared children’s production of irregular plurals in three conditions: (1) as the answer to a question (the labeling-question condition), (2) as the completion of the familiar frame *So many*, and (3) as the completion of a lexically-specific frame. By this I mean a frame that frequently precedes specific nouns in actual speech (e.g., *Brush your*
before teeth, or Three blind before mice). If, as in word recognition (Fernald & Huratado, 2006), word production is facilitated in familiar frames frequent in child-directed speech, then children’s production of irregular plurals should be better after a familiar frame than in isolation. If children are sensitive to both frame frequency and co-occurrence patterns between specific chunks and words, as suggested by Bannard and Matthews’ (2008) findings, then their production should be even better after a lexically-specific frame. I complement the elicitation data by looking at the linguistic contexts of naturally-produced errors to assess any parallels between experimental results and spontaneous speech.

This investigation represents a first step in exploring the role of linguistic context in children’s word production. For that reason, I was primarily interested in asking whether children’s word production was affected by linguistic context. If their accuracy varies with the linguistic context, this would provide evidence that their single-word production can be affected by the larger units that contain those words.

3.2 Experiment 1: Producing Irregular Plurals in Sentence-Frames

I used a picture-naming task to look at children’s production of English irregular nouns in three different conditions: (1) as an answer to the labeling-question ‘What are all these called?’ (2) as a completion of the frame ‘So many’, and (3) as a completion of lexically-specific frames found to appear frequently with each of the nouns tested. If the production of words is enhanced by the linguistic context, children should do better after a familiar sentence frame. If it is also affected by specific co-occurrence patterns between frames and words, they should do even better after lexically-conditioned frames.

3.2.1 Participants

Thirty-six children participated in the study (half boys, and half girls), with 12 in each condition. All the children were learning English as their first language, and all were developing normally. Their mean age was 4;6 (range 4;3-4;10) across the experiment and within each condition.
3.2.2 Materials

I looked at children’s production of seven irregular plural nouns: children, feet, fish, geese, mice, sheep, and teeth. I limited myself to irregular nouns that children are familiar with at the age tested (4;6 year-olds). I used a picture-naming task to elicit responses. The children saw pictures of multiple objects and had to name them. I compared their performance across the three conditions, using a between subjects design, and using the same pictures, in the same order, in all three. The children were asked to name the multiple objects shown in each picture, and, in all three conditions, produced one word for each picture. What differed between conditions was the linguistic context for the target irregular nouns. In the no-context condition, children produced each irregular noun after hearing the question ‘What are all these called?’ In the familiar-frame condition, they produced the irregular noun as a completion of the frame ‘So many—’. This frame was produced with rising ‘unfinished’ intonation by a puppet and the child was asked to help the puppet describe the pictures. In the lexically-specific condition, the children also completed two-word sentence frames produced by a puppet. But here each noun was matched with a frame that frequently co-occurs with it in spontaneous speech (e.g., brush your — for Brush your teeth; on your — for On your feet).

Selecting the frames

I used a six-million-word subset of the American English portion of the CHILDES database to extract the frames (MacWhinney, 1995). This included all speech spoken to and produced by non-impaired children between the ages of one and six. I extracted all the two-word frames that preceded our test nouns, limiting the search to cases where the resulting three-word sequence could be an independent utterance. I did this to ensure that the utterances used in the experiment sounded natural (for example, I excluded sequences like and the children). For each of the seven test nouns I selected the frame that appeared most often, and that appeared at least three times in the corpus. For the relatively more frequent nouns (mice, teeth, children, and feet) I could select two frames, while for the less frequent nouns only one frame met the criterion. Table 3.1 shows the frames selected.
for each noun. I also looked separately at the frequency of the frames in child vs. child-directed speech and found high agreement. The frames I chose were very frequent in both child and child-directed speech, but the frequencies differed for frames containing first- and second-person pronouns. For instance, *On your feet* was more frequent than *On my feet* in child-directed speech, while the reverse pattern was found in child speech.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td><em>All the children, The little children</em></td>
</tr>
<tr>
<td>Feet</td>
<td><em>On your feet, On my feet</em></td>
</tr>
<tr>
<td>Fish</td>
<td><em>Lots of fish</em></td>
</tr>
<tr>
<td>Geese</td>
<td><em>Buy some geese</em></td>
</tr>
<tr>
<td>Mice</td>
<td><em>Three blind mice, Afraid of mice</em></td>
</tr>
<tr>
<td>Sheep</td>
<td><em>Count some sheep</em></td>
</tr>
<tr>
<td>Teeth</td>
<td><em>Brush your teeth, Brush my teeth</em></td>
</tr>
</tbody>
</table>

Table 2.1: Lexically-Specific Frames

I used the same corpus to make sure that the frame used in the familiar-frame condition (*So many*) did in fact appear in child-directed speech (the frame occurred over 200 times), but did not often precede any of the test nouns. Unlike the question prompt or the *So many* prompt, fewer than half the lexically-specific frames contained any explicit cues to plurality (e.g., terms like *some, these*, or *all*, or numerals like *three*). By using lexically-specific frames that had fewer cues to plurality, therefore, we introduced a bias against our prediction that performance on irregular plurals would be best in this condition.

Filler items
In all three conditions the children completed a color-naming task in addition to the plural production task. They saw pictures of colored balloons and had to name each color in response to: ‘What color is this?’ In the two frame-completion conditions, children also
completed ten filler-frames in which the word to be supplied was a regular singular noun. For example, the child saw a picture of a red book and had to complete the frame *The red — ?*

### 3.2.3 Procedure

Children were tested in a quiet room at the nursery school. Children were first asked to name singular objects to make sure that they recognized the objects depicted and would use the intended labels for them. If they used a different label at this stage (like *lamb* for *sheep* or *kid* for *child*), the experimenter offered the intended label (*Right! But can we also call it a sheep?*). This was done to reduce the amount of non-intended responses in the plural-elicitation stage. The single-noun elicitation was followed by plural-noun elicitation (prompted by a question, a familiar-frame, or a lexically-specific frame, depending on the condition). The plural-elicitation started with two practice pictures. The first depicted a single house, the second five houses. In order to emphasize the plurality of the objects, the child was asked first to name the singular object (*What is this called?*) and then the plural objects (*and what are all these called?*). In the two frame-completion conditions the puppet was introduced after the practice pictures. The child was asked to help the puppet describe the pictures. The puppet would start a description and the child would finish it. Two practice trials with non-plural nouns followed to make sure the child understood the task (*It’s a — hat, It’s a — ball*).

**Coding**

Responses were coded for accuracy by a research assistant blind to the predictions of the study. They were coded as Correct if the child produced the correct irregular plural, but as Incorrect if the child produced an over-regularized form or a singular form. Such errors were further classified into three types: *s*-addition (adding an ‘s’ to a singular form), singular-form alone, and *s*-plural-addition (adding an ‘s’ to the plural form, e.g., *mices*). Finally, if children produced any alternative, non-intended, labels (e.g. *lambs, kids*), they were coded as Other.
3.2.4 Results and Discussion

In Experiment 1, I asked whether children’s production of irregular plurals would be better following a familiar frame than in isolation, and even better following a frame that often precedes those specific nouns in actual speech. Children gave responses in all of the elicitation trials. Their responses, when asked for an irregular plural form fell into five main categories: Correct, Added –s on singular, Added –s on plural, Singular alone, or Other, as shown in Table 3.2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Correct</th>
<th>Singular added –s</th>
<th>Plural added –s</th>
<th>Singular</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeling-question</td>
<td>28</td>
<td>32</td>
<td>2</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>So-Many</td>
<td>46</td>
<td>26</td>
<td>1</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Lexically-specific frame</td>
<td>63</td>
<td>2</td>
<td>0</td>
<td>26</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3.2: Percentage of each response-type elicited for irregular nouns (computed over all trials)

Figure 3.1 presents correct productions of the irregular plural forms for each condition out of total responses. I did not exclude Other responses since doing so would ignore the different proportion of Other responses elicited in the three conditions, 14% in the labeling-question, 12% in the so-many condition, and 9% in the lexically-specific one.
Figure 3.1 shows a clear effect of linguistic context on children’s accuracy. I used a mixed-effect logistic regression model to analyze the results. These models allow us to control simultaneously for subject and item effects (see H. Clark 1973; Baayen, Davidson & Bates, 2008). They also allow us to look at the effect of linguistic context while controlling for any effect of the frequency of the target nouns themselves. I ran mixed-effect models with accuracy (correct vs. incorrect) as the predicted variable, linguistic context and log (frequency) as fixed effects, and subject and item as random effects to compare each of the two conditions. I calculated the frequency of the nouns using the same six-million-word corpus we had used to select the frames.

As predicted, children did best in producing irregular plural forms correctly when a lexically-specific frame ‘introduced’ the target noun (63% correct). They continued to do
quite well after a *So many* (46% correct), and did least well when producing the irregular nouns in isolation (28% correct). Children did significantly better in the lexically-specific condition compared to the so-many condition, $B = 1.00$ ($SE = .42$), $p < .05$, and better in the so-many condition compared to the labeling-question condition, $B = 1.16$ ($SE = 39$), $p < .01$. In all three conditions, children were more accurate with more frequent nouns ($p$’s $< .001$). Performance in the lexically-specific condition was twice as good as in the labeling-question one. It was also twice as good as the performance levels previously reported for this age-level (e.g., Ramscar & Yarlett, 2007; Arnon & Clark, 2008).

Differences in accuracy cannot be attributed to knowledge of the relevant singular noun forms. Children successfully labeled singular objects before all three plural elicitation conditions: They were correct on 65% of singulars in the labeling-question condition, on 57% in the so-many condition, and on 66% in the lexically-specific condition. These differences were not significant (using the labeling-question as the baseline, $B$(so-many) = -.56 ($SE = .34$), $p < .1$, $B$(lexically-specific) = .04 ($SE = .34$), $p > .8$). In the remaining trials, children sometimes produced a non-target noun (e.g., *kid* for *child* 30% of the time) or, more rarely, used a plural form (5% of the time). Overall, though, children in both conditions were equally familiar with the objects depicted in the pictures, and with the singular forms of the relevant nouns. The children were also equally accurate, and at ceiling, (97% correct) on the color-naming task.

The plot in Figure 3.2 shows the mean accuracy for each condition by item. While the baseline accuracy differs for each noun (higher frequency nouns like *teeth* are produced accurately more often than lower frequency ones like *geese*), the general pattern is the same across items. Children were least accurate for irregular plurals after a labeling-question, better after a familiar frame, and better still after a lexically-specific frame.
These results represent a first demonstration of an effect of linguistic context on children’s accuracy in producing morphologically inflected words. They show that children’s accuracy in production varies with the linguistic context the target forms are elicited in. Before I turn to the broader implications of these results, and discuss possible mechanisms that might underlie them, I look for similar effects in spontaneous speech. If the experimental results reflect something fundamental about how children’s linguistic knowledge is organized and accessed, children should display similar patterns in everyday conversation. They should be equally sensitive there to the linguistic contexts for irregular nouns. They may produce fewer over-regularization errors for irregular plurals after lexically-specific phrases. I investigate this prediction in the next section.
3.3 A Corpus Study of Errors in Spontaneous Speech

In this part I explore the role of linguistic context on spontaneously produced errors. There are many corpus studies of irregular plurals but none have looked at the effect of context on error rate. Looking at naturalistic speech serves two goals – to increase the ecological validity of the experimental findings, and to expand our understanding of the factors that influence the production of irregular forms in natural speech. I wanted to see if children were less likely to make errors after the lexically-specific frames used in the previous study than after other two-word sentence frames. To do this, I looked at the linguistic material preceding correct and incorrect irregular plurals in child-produced speech.

3.3.1 Materials

I drew on all the child-produced speech within the six-million-word corpus I had used to select frames. This reduced the corpus to 1.7 million words, produced by over 50 children. 20 of them were followed longitudinally and the rest (the Hall corpus) were recorded over a period of two consecutive days. Using CLAN (MacWhinney, 1995), I extracted all the over-regularized forms produced by children, the utterances they were produced in, and the two utterances surrounding them (one utterance before and one utterance after). I expected to find two kinds of errors: addition of the regular ‘–s’ plural marker (a) to the singular form (e.g., *mouses*), and (b) to the irregular plural form (e.g., *mices*). Both error-types were attested in the experimental data. I therefore searched the corpus for all the following forms: *childs, childrens, fishes, gooses, geeses, mouses, mices, tooths, teethes, sheeps, foots, feets*. As in previous corpus studies (e.g., Marcus et al., 1992), we ignored errors of omission, where children used a singular form in lieu of a plural.

3.3.2 Results and Discussion

Children produced the nouns in question correctly 2118 times, and made 102 over-regularization errors (4.5%). Of these, 75% (77) involved adding ‘–s’ to the singular form
and 25% involved adding ‘–s’ to the plural form. Table 3.3 shows both the errors and the correct productions for each irregular noun.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Correct</th>
<th>Errors</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>205</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Feet</td>
<td>478</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>Fish</td>
<td>840</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>Geese</td>
<td>25</td>
<td>2</td>
<td>7.5</td>
</tr>
<tr>
<td>Mice</td>
<td>73</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Sheep</td>
<td>123</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Teeth</td>
<td>374</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2118</strong></td>
<td><strong>102</strong></td>
<td><strong>4.5</strong></td>
</tr>
</tbody>
</table>

Table 3.3: Correct and erroneous plurals for irregular nouns in spontaneous speech

Averaged across nouns, the error rate was 4.5%. However, the rate varied greatly from one noun to the next (from 1% for teeth to 21% for mice). These findings parallel the findings that different words are over-regularized at different rates at different points in development (e.g., Maslen et al., 2004; Marchman et al., 1997; see also Maratsos, 2000), and go against the uniform low-rate of over-regularization argued for by Marcus (1995; Marcus et al., 1992).

Since I wanted to compare experimental results with corpus data, I further analyzed only those errors produced following a two-word sentence-frame. This left 58 errors (of the now excluded errors, 15 appeared as one-word utterances, or were the first word of an utterance, and 29 were preceded by just one word).

As predicted, children seem to make fewer errors after lexically-specific frames. Of the 102 errors, 58 were produced after a two-word frame. (In addition, there were 15 over-regularizations produced as one-word utterances, or as the first word of an utterance, and 29 preceded by just one word.) None of the frames followed by an error, though,
were the lexically-specific frames used in our elicitation study: Children didn’t produce utterances like *Brush your teeths* or *on my feets*. They did produce the lexically-specific frames followed by an irregular plural noun (79 times), and in every case, the irregular form was produced correctly. They did produce errors like *Has stinky feets* (3;1) or *Found dinosaur tooths* (4;9) where the erroneous form follows a frame that doesn’t often preceded the specific nouns.

I also checked to see whether the frames that preceded errors ever preceded correct productions in child or child-directed speech (I used the full six-million corpus for this analysis). Two-thirds of the errors occurred following sentence-frames that children hadn’t heard or said before with that specific noun. For example, for *my baby feets*, we find that the correct form *feet* had not occurred after *my baby.* (Interestingly, eight of the errors (13%) appeared in frames that contained an additional mistake like *on him feets, they is mouses, from him foots*, almost as if errors were more likely in the context of other errors.)

These results are, of course, limited in that they are based on a relatively small set of errors, and in that I did not include cases where children incorrectly produce a singular form instead of a plural. But they are consistent with the prediction that children’s production of irregular forms is improved after frequent frames also in spontaneous speech, and that errors occur less often in phrases that children have not heard or produced before.

Discussion

3.4.1 Summary of findings

I set out to investigate whether children’s production of irregular plurals is affected by the larger patterns (sentence-frames) that these words are learned and used in. Experiment 1 showed that children’s accuracy was greatly influenced by the linguistic context the target words were produced in. Children showed the customary poor performance when producing irregular plurals in isolation (28% correct). They showed enhanced performance when producing such nouns after a familiar frame (46% correct), and
relatively high performance (63% correct) when producing the same nouns after a frame that often precedes them in natural speech. Note that this level of performance (70% correct) is twice as high as found in previous experiments with the same age-group and language population (Arnon & Clark, 2008; Ramscar & Yarlett, 2008). This pattern held across items and was clearer when noun frequency was controlled for. This effect is unlikely to reflect differential cues to plurality in the three conditions. Children’s production accuracy was highest following lexically-specific frames despite the fact that fewer than half the frames contained explicit cues to plurality (the use of plural terms like some or all, or numerals like three) while all the prompts had such cues in the other two conditions (What are all these called? So many —).

The corpus findings revealed a similar pattern. Children didn’t produce errors following any of the lexically-specific frames used as prompts in Experiment 1. They made over-regularization errors (4.5% averaged across all irregular nouns). But they did not make errors like brush my teeth or on your feet. In fact, most of the errors appeared following sentence frames that the child had not produced before with that noun. While the corpus results are limited in that they are based on a small number of errors, they provide additional evidence that children’s spontaneously produced errors are also sensitive to linguistic context.

This is a first demonstration that linguistic context affects children’s accuracy in producing morphologically inflected words. At a minimum, these findings show that children’s word production is not independent of the linguistic context that words are embedded in: They rely on distributional patterns in the process of word learning. Word production was affected by properties of the linguistic context (the presence of a sentence-frame and its frequency) as well as by properties of the words themselves (the token frequency of the nouns). The findings provide a cautionary note for the single-word focus that characterizes much research on lexical and morphological acquisition. Since linguistic context does affect single word production, asking children to produce words in isolation, or following a neutral prompt that rarely precedes the target words in natural speech, may not be the best way to assess children’s word knowledge. In fact, stripping away sentential context may tell us more about the meta-linguistic ability to produce and
comprehend language out of context (an ability that adults clearly possess), than about the developmental path of ‘natural’, contextually-bound, word learning. For a similar but more far-reaching argument against the study of language in non-communicative settings, see H. Clark, 1997. To get a more complete picture, we need to look at children’s productions of words both in isolation, and in varying and naturally-occurring linguistic contexts.

3.4.2 The units children learn

By looking at the production of single words in isolation and in different contexts, we can learn several things about what children know about those words. The first thing to notice is that children know about the relation between specific words and the (multiple) words that precede them. For children to be differentially affected by different frame-types, they must learn about larger patterns of language, specifically, the phrases that words are embedded in. The current findings join previous ones in showing that children can and do attend to multi-word phrases (Bannard & Matthews, 2008; Fernald & Huratado, 2006).

The current findings also go beyond previous ones in showing the facilitative role of phrases in producing single-words. Children were more accurate in producing irregular plurals when they were preceded by familiar and lexically-specific frames. This facilitation follows naturally if children, like adults, are sensitive to distributional information at multiple levels of analysis. Adult production is sensitive to co-occurrence patterns at many grain-sizes: from sound combinations (Colman & Pierrehumbert, 1997) through single words (Jescheniak & Levelt, 1994) to multi-word phrases (Bybee & Schiebman, 1999), and syntactic constructions (Bock, 1986; Gahl & Garney, 2004). Like adults, children rely on such information in production and show a gradient effect of predictability: words are easier to produce when they are more expected given the preceding context.

These findings demonstrate a link between larger and smaller units of language. The production of smaller units (words in this case) was affected by the larger patterns they appeared in (sentence-frames). This is exactly the pattern we would expect if children are
using larger units in learning about smaller ones, as predicted by the Starting Big Hypothesis. On some level, word knowledge is tied to the larger patterns words appear in. This is not to say that children don’t learn to differentiate words from the linguistic material that surrounds them. Word learning crucially depends on the ability to segment words from speech. But, while less studied, word learning also seems to depend on extracting, and preserving, the systematic relations that words have to the linguistic material that surrounds them.

While the current findings document a relation between words and phrases, they do not tell us how this relation developed. They do not prove that children’s lexical and morphological knowledge was learned by segmenting unanalyzed phrases like ‘three blind mice’\(^1\). At the time of testing, children clearly had knowledge of words that is separate from the phrases: the task required them to produce single words, not phrases. At the age tested (4;6) children have a vast amount of productive, combinatorial knowledge of language. Starting as one unit is one way of forming a relation between words and phrases. Alternatively, words can become more associated with phrases they often co-occur with. These two different processes will lead to a similar end-result: a reliance on larger patterns in the production of smaller ones.

The current findings show that children’s knowledge of irregular plurals was not static. They produced plurals with varying degrees of accuracy in different contexts. The goal of the current chapter was to show that this variance can, in part, be attributed to the effect of linguistic context. Performance is better understood when the relation between phrases and words is taken into account. But this variable performance raises broader questions about our ability to treat linguistic knowledge, or linguistic units, as static entities. Even adults can be induced to make errors on irregular forms under certain conditions (Stemberger, 1992; Stemberger & Middleton, 2003). Yet we would not want to say that they lack knowledge of the correct form. The realization of sounds differs in

---

\(^1\) It is interesting to explore whether any of the frames used in Study 1 were ones that words were initially segmented from. While this may be true for some of the lexically-specific frames I used (like \textit{brush your teeth}) it is probably not true for others (\textit{on your feet}). It may be possible to approach this question through a dense longitudinal corpus study that tracks the frames that words first appear in, in child-directed and child speech.
different words but we would not want to argue that language users do not learn about phonemes as well as words. The variability inherent in performance, the learning of units of varying sizes, and their inter-dependence can be cohesively integrated by adopting a more dynamic view of linguistic knowledge, and of the units that language users draw on (Elman, 1995; Thelen & Smith, 1994; Grossberg, 1980; Van Geert, 1991).

3.4.3 The units children learn – Take II

In a recent article, Goldinger & Azuma (2003) draw on Adaptive Resonance Theory (ART, Grossberg, 1980) to shed light on a long-standing debate about the basic units of speech perception. Are phonemes or syllables the fundamental unit of perceptual analysis? Thirty years of research have given rise to conflicting findings. Different units assume priority in different languages and tasks, sometimes phonemes are more prominent and sometimes syllables are (see Goldinger & Azuma, 2003 for a review). ART provides a way to make sense of these discrepancies. Cognitive behavior is seen as a complex dynamic system which achieves self-organization through adaptive resonance (Grossberg, 1980). The resulting system is fully inter-dependent, and perception occurs when bottom-up and top-down information bind into a stable state. For speech perception (Grossberg, 2003; Grossberg & Myers, 2000), the basic assumption is that listeners attend to speech units of all possible grain-sizes, and that the ones they depend on vary according to the nature of the task and stimuli at hand. The fact that both phonemes and syllables can achieve perceptual prominence is not surprising. It shows that “[…] unitization is a flexible, opportunistic process: The most predictive units will dominate behavior.” (Goldinger & Azuma, 2003: p. 308).

We can think about the units that children learn and use in a similar way. Children learn about both words and phrases. Using ART terms, the perception of words will involve the combination of bottom-up (word) and top-down (phrase) information. The extent to which children rely on, and attend to, either unit-type will be influenced by the nature of the task and the ‘usefulness’ of each unit. For example, multi-word phrases may be especially useful for conveying early communicative acts or participating in social routines (bye-bye, wat-is-dat). In her 1977 article Peters provides an example of this sort.
One of the earlier gestalt utterances of the child she studied, produced at 1;7, was /sil-i-ni?/, used to mean silly, isn’t it? In this case, the whole indeed seems more useful than its parts.

ART makes an additional prediction that is very relevant for the Starting Big approach. All being equal (which is rarely the case), processing will typically favor larger units. In the case of speech perception, this means that syllables will tend to be more prominent than phonemes, and words more prominent that syllables. But there may be conditions, like the one demonstrated in the example above, when phrases may be more prominent than words. There are no existing ART models to my knowledge that attempt to model language behavior beyond the word level. But it would be interesting to try and extend the predictions of ART to the relation between words and phrases. Using ART to think about the units of language learning can capture the inter-dependence between linguistic units and the reliance on larger patterns in producing smaller ones. Task and stimuli dependent behavior is expected and can be understood. Additionally, it may help us generate predictions about the situations where we might expect children to rely more on ‘chunked’ language. I discuss the usefulness of applying a dynamic systems approach to language learning and use in the final chapter of this dissertation.

3.4.4 Why are some frames better than others?

Children were affected not only by the presence of a frame, but also by its lexical content. They were more accurate after frames that commonly co-occurred with the target nouns, as expected if those frames increase the predictability of the correct form. But why was performance also enhanced (though less so) after a familiar but less highly collocated frame, namely ‘So many —’? As in word recognition (Fernald & Hurtado, 2006), word retrieval and production may benefit when the preceding context is more familiar, and hence easier to process. But there may be another explanation that is tied to the particular frame I used here. This frame rarely preceded any of the test nouns in the corpus I drew on, so there is little evidence for any sequential link between this frame and any irregular plural nouns. But there is a link between the frame and the notion of plurality.
The frame *So many* appeared over 200 times, and was followed by a plural noun 77% of the time (n = 155). (In the remaining cases, this frame appeared in utterance final position with no following word, e.g., *I got so many, there are so many.*) Even though the frame *So many*—was followed mainly by regular plural nouns (90% of the time), the fact that it is generally followed by plurals may have helped children retrieve irregular plural nouns there as well. Some recent experimental evidence suggests that activating the notion of plurality through exposure to regular plurals can facilitate the production of irregular plural forms as well (Ramscar & Yarlett, 2007). So, even though the frame *So many*--- was followed mainly by regularly inflected nouns (90% of the time), the fact that it was followed by plural nouns may have been enough to enhance the retrieval of irregular plural forms too. In short, children’s improvement in production after the frame *So many*--- may reflect the fact that this frame often precedes plurals in children’s input. This explanation too can be re-cast in terms of predictability: A plural form is more predictable after the frame *So many*--- than after the (neutral) labeling-question prompt.

3.4.5 Prosodic contour and retrieval

One additional factor deserves consideration here in accounting for children’s improved performance after frames. The two frame-completion conditions differed from the labeling question in prosody (two unfinished utterances vs. a question) as well as in the nature of responses they elicited (a word in utterance-final position vs. single-word answer). The prosodic characteristics of the linguistic context alone can’t account for the difference we found between frame-types: Children’s production was better after a lexically-specific frame than after a frequent frame (*So many*), even though both provided the same prosodic context (an ‘unfinished’ intonation contour). But prosody may nonetheless have contributed to the general improvement children showed in the frame-conditions. Children’s lexical retrieval may be affected by the prosodic contour a word is produced in.

These ideas are especially intriguing when we consider episodic theories of lexical access (Goldinger, 1996, 1998; Roediger & McDermott, 1993; Tenpenny, 1995), and their implications for speech perception and production. In such models “[…] groups of
detailed traces collectively represent individual words” (Goldinger, 1996: p. 1166). Those traces include, among other details, fine-grained acoustic information pertaining to the voice of the speaker (Goldinger, 1996, 1998; Hintzman et al. 1972; Shacter & Church, 1992), and the intonation contour a word was produced with (Church & Shacter, 1994). Word perception and production are facilitated when there are more shared features between the current stimuli and past experiences. For example, hearing a word in a voice that is perceptually more similar to a previous presentation enhances perception. More relevant to my study is the finding that word identification was better when a word was produced in the same intonation contour (question vs. statement, Church & Shacter, 1994). Word perception seems affected by the prosodic contour of the preceding linguistic material.

So far, studies investigating episodic theories of the lexicon have focused on properties of words in isolation. But given the continuous nature of speech, there is no reason to think that memory traces are limited to words in isolation. Instead, they appear more likely to “contain a rich linguistic history, reflecting words in various contexts, nuances, fonts, and voices” (Goldinger, 1998: p. 268; Hintzman, 1986). If memory traces for words include information about the prosodic context of production (e.g., sentence-final intonation), then perception and production may be enhanced in more familiar prosodic environments. Because words are more likely to appear as parts of larger utterances than as single-word utterances, this might explain why children did better after hearing a frame than in answering a question - the prosodic context the frame supplied may have facilitated retrieval.

Under this interpretation, the reduced performance in my study following the question prompt could reflect the prosodic fact that words are more often produced as part of utterances than in isolation. We can re-interpret the finding that children’s word recognition was enhanced following a familiar frame (Fernald & Hurtado, 2006) in this light. Children may have been faster to recognize words when they appeared following

---

2 This raises further questions about the relevant dimensions of prosodic context. Is it just in context vs. in isolation or are speakers actually sensitive to the locations of words within utterances (e.g. tend to be utterance initial, utterance final, etc.).
the familiar frame *Look at the bunny* compared to the semantically equivalent frame *Look! The bunny*, not only because of the lexical difference between them, but also because of the different prosodic contours they were presented in. I want to investigate this in future work by looking at children’s word recognition and production following sentence-frames that have undergone low-pass filtering that preserves the prosodic contour while removing any lexical information. This would allow analysis of the effect of the prosody of a frame and its lexical content independently. I want to know whether children show facilitation following such frames and to what degree – more than or equal to the facilitation provided by unfiltered familiar frames.

3.4.6 Rules and connections

The focus on irregular plurals has the added benefit of contributing to another theoretical issue. The source of children’s over-regularization errors has spurred heated debate. Do they reflect the application of a regular rule (e.g., Marcus et al., 1992)? Or do they reflect the higher activation of the more frequent regular form overall (e.g., Plunkett & Marchman, 1993)? This debate has been pursued in experiments (Arnon & Clark, 2008; Matthews & Theakston, 2006; Ramscar & Yarlett, 2007), corpus studies (Marcus et al. 1992; Maslen et al. 2004), and simulations (Plunkett & Marchman, 1993). But researchers have not looked for linguistic context effects, nor considered how such effects could be accommodated.

The current results have clear implications for the debate over the acquisition of irregular forms. Like Maslen et al. (2004), my corpus findings also show that that error-rate is not constant across nouns, but varies with frequency (from 1% to over 20% error rate). This is not easily accounted for in a rule-based account of errors (e.g., Marcus et al, 1992), but appears more explicable if errors reflect lower activation levels that are affected by the overall frequency of the inflection (type frequency), and by the frequency of the specific plural noun (token frequency).

My results go beyond previous findings in showing that accuracy can be boosted by manipulating the immediate linguistic context of the target words. This finding presents a challenge for both kinds of models. For activation-based accounts, the challenge is of a
more practical nature. Existing implementations do not incorporate linguistic contexts (Plunkett & Marchman, 1993; Plunkett & Joula, 1999; Rumelhart & McClelland, 1986). They model children’s behavior using units that roughly correspond to words. They model children’s productions with units that roughly correspond to words. To account for the current findings, such models would have to incorporate information about the larger patterns words appear in. This may be possible to do in a simple recurrent network (Elman, 1991). Such networks have been used successfully to model local-coherence effects (Tabor, Juliano, & Tanenhaus, 1997) and thematic fit effects (Tabor & Tanenhaus, 1999), both of which require keeping track of the distributional contexts. Effects of linguistic context on word production could likewise be modeled by tracking larger sequential chains. However, there are no existing connectionist simulations of irregular plural use that set out to capture such relations.

The findings pose a more serious challenge for rule-based accounts of inflection. Recent renditions of rule-based models accommodate the effect of irregular token frequency by suggesting that the application of the default regular rule is more easily blocked when the irregular in question is more frequent (Pinker & Ullman, 2002; Walenski & Ullman, 2005). But to account for the effect of linguistic context, such models would have to allow rule-application to be modulated by the larger linguistic context words appear in. That is, to be affected by elements external to the unit the rule operates over, namely the word. Allowing this would mean that the ‘rule’ no longer operates over clearly defined grammatical units (like Verb and Noun). The characteristics of the regular rule are made even hazier if inflection accuracy is affected by phonological and semantic features (Köpcke, 1998; Matthews & Theakston, 2006; Ramscar, 2002). It is not clear how rule application can be stochastically affected by non-grammatical features of the form in question.

3.4.7 Pairs of plurals in spontaneous speech

My findings on the role of the immediate linguistic context in children’s production of irregular plurals may also account for another puzzle – children’s simultaneous uses of correct and incorrect inflected forms. Some earlier studies have argued for a U-shaped
developmental curve. At first, children produce the correct irregular plural form—let say, *men*. Next they learn the general rule for producing plurals—and now substitute *mans* for their earlier *men*. And then, finally, they learn that the plural of *man* is *men*. However, even in the earliest report of this supposed U-shaped development in the acquisition of irregular nouns and verbs, Cazden (1968) reported that one of the three children she studied used a correct irregular past tense, *went*, alongside an over-regularized form, *goed*. Other children have also been reported as fluctuating between correct and over-regularized forms of the same verb, even within the same conversation (Kuczaj, 1977; Slobin, 1978). Parental reports corroborate side-by-side uses of correct and over-regularized plurals: Many children used two plural forms (correct and incorrect) for at least one irregular noun (Thal et al., 1997; Marchman et al., 1997). Arnon & Clark (2008) provide further evidence for the prevalence of side-by-side uses. In this study, children (aged 4;6) were asked to produce irregular plural forms in various tasks so we elicited multiple productions of the same form for each child. We found that 19 of the 22 children (86%) produced both correct irregular forms and over-regularized forms of the same word within the same experimental session, e.g., producing *mouses* in one task and *mice* in another. In short, children do not ‘drop’ the correct irregular plural forms but use them side-by-side with the over-regularized plural.

The linguistic context the erroneous forms are produced in may account for these side-by-side uses. Children may be producing over-regularization errors in less familiar frames, alongside correct uses in familiar frames. Further research is needed to see if frame-type can account for the variability in accuracy within children. If so, it would offer a novel way to make sense of children’s interchangeable uses, and a useful framework for studying the processes that such side-by-side uses reflect. More generally, the pervasiveness of side-by-side uses, and their potential relation to linguistic context, cast doubt on the existence of a U-shaped learning curve for irregular forms – while incorrect uses are initially rare, they do not fully replace correct uses in later development.

46
3.4.8 Some limitations

The results are limited in several ways. First, I only looked at irregular plurals. One would need to extend these findings to other domains (e.g., verb inflections) to see how widespread the effect of frames is on word production. Second, I only looked at the effect of context on accuracy: forms were either correct or incorrect. My findings might receive further support from more subtle measures of production such as voice-onset time and duration: the preceding linguistic material may affect not only accuracy (whether the correct form is retrieved) but also the speed and fluency of correct forms as well. The analysis could benefit from (1) obtaining online measures of production like voice-onset time and duration, and (2) from looking for parallel effects on word recognition. If the facilitation effect reflects the way linguistic information is organized in memory, we would expect word recognition to also benefit from lexically-specific frames. Third, I used raw frequency of occurrence (how often a frame appeared with a target noun) to select the lexically-specific frames. There are other measures that are relevant for measuring the degree of association between sentence-frames and words. One factor that is emerging as important is entropy – how much information is added by the random variable (in my case, how much information does the word add to the sentence-frame. The more often the two co-occur, the less information it will carry). Entropy has recently been used to predict how productive sentence-frames will be (Mathews & Bannard, 2008). Finally, I used sentence-frames produced by adults (or a puppet) to prompt children’s production. While the corpus study suggests that the same pattern holds when children are producing the entire utterance on their own, it would be useful to elicit whole utterances in an experimental setting as well.

3.5 Summary

The goal of the current chapter was to demonstrate (a) that children learn about the larger patterns that words appear in, (b) that they draw on such knowledge in the production of words, even when segmentation has already taken place, and (c) that this reflects the inter-dependence between linguistic patterns of varying sizes: Larger patterns
are utilized in the production of smaller ones. The chapter shows that multi-word phrases facilitate children’s production of irregular forms. If the role of multi-word phrases in children’s production reflects a basic property of the organization of linguistic knowledge, namely, the ability to attend to, and draw on linguistic patterns of varying sizes and levels of abstraction, then we should find a similar sensitivity to multi-word phrases in adult language. In the next chapter I explore this prediction by asking whether adults are sensitive to the frequency of compositional multi-word phrases like *don’t have to worry*. 
4.1 The Building Blocks of Language

In the previous chapter I tested the predictions that children attend to multi-word phrases and that those larger units can be used to facilitate the production of smaller units (i.e. single words). Both predictions were supported. Children were in fact better at producing irregular plurals correctly when they were preceded by a familiar sentence-frame. They were even better when the preceding frame was one that often appeared with the nouns in actual speech. To give a concrete example, children produced the irregular form ‘mice’ more accurately in *Three blind mice* than in *So many mice*, and more accurately in *So many mice* than in isolation.

These findings highlight the importance of multi-word phrases in first language acquisition. They show that young speakers are sensitive to the frequency with which different phrases appear; to know that *feet* is often preceded by *on your---* the child has to be able to attend to the entire phrase. More importantly, the findings show that young children utilize such co-occurrence information in online production. They add a dimension to what we think children know about their language: not only words but also the larger patterns that they appear in. But is this knowledge restricted to young speakers? Or do adult speakers also possess knowledge about multi-word phrases? In the current chapter I explore the idea that adults continue to represent multi-word phrases, and discuss implications of this for models of the lexicon and grammar.

4.1.1 Dual-system and single-system approaches

There has been long-standing tension in the study of language between approaches that assume a clear distinction between the mental lexicon and grammar (dual-system theories, Chomsky, 1965, 1995; Fodor, 1983; Pinker, 1991; Pinker, 1999; Pinker &

---

3 The work reported in this chapter was done in collaboration with Neal Snider.
Prince, 1988; Ullman, 2001; Ullman, 2004) and ones that do not (single-system theories, Bates & MacWhinney, 1989; Elman, 1999; MacDonald, Pearlmutter & Seidenberg, 1994; Rummelhart & McClelland, 1986; Seidenberg, 1994). Dual-system models distinguish between the mental lexicon – an inventory of memorized forms, and the mental grammar – the rules or constraints used to combine the memorized elements. This distinction echoes the one made in many generative models of language (Chomsky, 1981; Jackendoff, 2002; Kaplan and Bresnan, 1982; Pollard and Sag, 1994). The mental lexicon contains the linguistic units that cannot be derived: simple words (e.g. cat), morphemes, irregular nouns and verbs. The bulk of the lexicon consists of atomic words. Multi-word phrases are only stored when they cannot be derived (as in the case of certain idioms like kick the bucket, Pinker, 1999). The most clearly articulated model of this kind is presented by Pinker and his colleagues (Pinker, 1991; Pinker & Prince, 1988; Pinker & Ullman, 2002). In this model, the two components of language (lexicon and grammar) are thought to be learned differently, to involve different cognitive abilities and to be governed by different neural substrates (Ullman et al. 2005; Ullman, 2001). Forms created by grammar are distinguished from those that originate in the lexicon.

No such distinction is posited by single-system theories. Instead, all aspects of language depend on one computational system. The same cognitive mechanism is used to process all linguistic experience, whether it is a non-compositional lexical item like ‘cat’ or a compositional phrase like ‘I don’t know’. Word-object mappings and grammatical rules are learned in a qualitatively similar fashion – by abstracting and generalizing from linguistic experience. There are now a growing number of models that implicitly or explicitly take a single-system stand. In connectionist models, the unity of lexicon and grammar is made explicit by using one single network to capture all linguistic experience (e.g., Rumelhart & McClelland, 1986, Seidenberg, 1994). Exemplar-models of language also dispense with the distinction though they do it in a different way (Bod, 1998, 2006; Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001): by having linguistic units and categories correspond to clusters of memory traces.

Exemplar models differ from connectionist models in several important respects: the use of symbolic or non-symbolic representations; the implementation of higher level categories like nouns or verbs; the way linguistic similarity is captured (see Bybee & McClelland, 2005 for a discussion). But for the purpose of the current chapter, neither model makes a clear distinction.
These models are closely related to what are often labeled usage-based approaches to language (Bybee, 1998, 2006; Goldberg, 2006; Barlow & Kemmer, 2000; Langacker, 1986; 1987; Tomasello, 2003) where grammatical knowledge emerges from linguistic experience. The lexicon is not ‘reserved’ for atomic elements, but instead contains patterns of varying sizes and levels of abstraction. There is no a priori limit on the size of the units that are stored; as long as they can be attended to and remembered, units can be of varying length (word, two-word, multi-word phrase) and levels of abstraction: ranging from single words, through partially realized constructions to fully abstracted ones (give, give me a break, give NP a break, give NP NP, V NP NP).

4.1.2 The representational status of compositional phrases

The contrast between single and dual-system models has been mostly studied in the domain of morphological processing and representation (see Pinker & Ullman, 2002 vs. McClelland & Patterson, 2002 for a more recent round). But given their very different views on what language is and how it should be represented, these models should make different predictions in many other domains. Here, I investigate one that has rarely been studied: the way the two approaches handle compositional multi-word phrases (like don’t have to worry).

I use the term ‘compositional multi-word phrase’ to refer to multi-word strings of words whose meaning and syntactic form is compositional. Such phrases differ from multi-word expressions as studied in the syntactic and computational literature (e.g. Moon, 1998; Sag et al., 2001; Sinclair, 1991). In this literature, the term is used to refer to “idiosyncratic interpretations that cross word boundaries or spaces” (Jackendoff, 1997, adopted by Sag et al., 2001). Since the phrases I will study do not have idiosyncratic semantic or syntactic features, I use the term ‘multi-word phrases’ instead.

The different ways the two approaches handle words are clearly articulated in existing models (e.g., Rumelhart & McClelland, 1986; Pinker, 1991). Their predictions about larger units are not clearly stated in any existing model but can be extrapolated from their general assumptions about language. Under a single-system model, compositional

between the linguistic units that are memorized and the ones that are the product of ‘rules’. All linguistic experience is expected to be learned, processed and used in a similar fashion.
phrases should be like any other linguistic pattern. Every encounter with a compositional phrase is predicted to add to its representation and influence future processing. Compositional phrases should be impacted by the same factors that impact the processing of simple words and regularly inflected words (e.g., frequency). A different prediction is made by dual-system models. Compositional phrases can be predictably derived. Unlike non-compositional phrases (like certain idioms, Pinker, 1999), they do not have to be stored in the mental lexicon. In fact, given the goal of these models to minimize storage, compositional multi-word phrases seem like very unlikely candidates for storage in the lexicon. The extent to which compositional multi-word phrases are part of the mental lexicon, and the extent to which their status can be distinguished from that of non-compositional phrases, has an important role in evaluating these views on language.

I address this question by investigating whether comprehenders are sensitive to the frequency of the phrases when part frequency is controlled for. In doing so, I draw on a similar method used in the morphological literature (e.g. Taft, 1979) and on a recent application to the study of larger units (Bannard & Matthews, 2008).

4.1.3 Frequency effects and mental representation

Frequency plays a very different role in single and dual-system models. In single-system models, frequency effects reflect the way linguistic knowledge is learned and organized. Frequency is an approximation of experience. As such, it plays an important role in the emergence and entrenchment of linguistic units. The more a pattern is experienced, the easier it will be to access and use (Bybee, 2006; Bod et al. 2003, Bybee & Hopper, 2001). Single-system models differ in the specific mechanism they use to explain the processing advantage of more frequent forms (by impacting the weights in a connectionist network, by lowering the threshold of activation in spreading activation networks, or by enhancing the activation of a memory trace in exemplar models). But they share a common belief that frequency effects tell us about the units that speakers attend to and a common prediction that frequency effects should be found for all linguistic units: simple or complex.

Frequency effects are viewed differently in dual-system models. For starters, the role of frequency in language representation and use is rarely discussed explicitly in these
models (e.g. Pinker, 1999). This lack of discussion seems to echo the traditional Chomskian belief that frequency effects are irrelevant to the study of language because they reflect real-life probabilities or performance issues that are separate from, and immaterial to linguistic knowledge (Chomsky, 1957, recently re-argued for in Newmeyer, 2003). In specific models (Ullman & Wellensky, 2005) frequency effects are relegated to the mental lexicon. This allows such models to account for the wide-spread frequency effects found for word production and comprehension (see Monsell, 1991 for a review) while maintaining a distinction between ‘stored’ and ‘computed’ elements. ‘Stored’ elements should exhibit frequency effects but ‘computed’ elements should not.

4.1.4 Lessons from morphology

To distinguish between the two approaches presented above we need a way to determine whether compositional multi-word phrases are represented in the mental lexicon. This is not an easy thing to do. How can we assess whether a linguistic form is part of the mental lexicon or not? A similar question has been debated in the morphological literature. The battle between single-system and dual-system models has been played out in most detail in the domain of morphologically inflected words. Specifically, in the study of the representational status of regular and irregular inflected forms (e.g. walked vs. felt). Dual-system models (Marcus et al., 1992; Pinker, 1991, 1999; Pinker & Prince, 1991; Pinker & Ullman, 2002; Ullman et al. 1997) predict that irregular forms will be stored in the mental lexicon while regular forms will be generated by the grammar. Single-system models predict that all forms will be represented by the same associative memory mechanism (Rumelhart & McClelland, 1986; Plunkett & Marchman, 1991, 1993; Marchman, 1993).

Frequency effects have been used to contrast the different representational status of regularly inflected forms in both accounts. As noted by Alegre & Gordon (1999) “If regularly inflected forms are not represented as whole words, then the base form must be accessed every time an inflected word is processed. Thus access speed should depend on the cluster frequency of the base form, defined as the aggregated frequencies of the base and all its inflectional variants (e.g., for walk, the summed frequencies of walk, walks, walked, and walking). On the other hand, if whole-word representations are available for
regularly inflected forms, then the frequency of the inflected form itself should determine access time.” Put differently, finding that the frequency of the inflected form is predictive of processing would indicate whole-form representation, as argued by single-system but not dual-system models.

The representational status of regularly inflected forms is tested by using a whole-form frequency manipulation. Starting with Taft (1979) and continuing in later studies (Alegre & Gordon, 1999; Baayen et al. 1997, 2002; Sereno & Jongman, 1997; Stemberger & McWhinney, 1988; Taft, 1979), researchers have asked whether processing is affected by the frequency of the inflected form itself (*walked*) when the frequency of the base form (*walk*) and the inflectional morphemes (*-ed*) is controlled for. To be affected by its frequency, speakers must have a whole-word representation of the inflected form.

4.1.5 Whole-form frequency effects for phrases

In a recent study, Bannard and Matthews (2008) extended this argument to multi-word phrases. They wanted to show that children store multi-word phrases. This is a crucial assumption for certain models of language development where grammatical knowledge is learned by abstracting over stored utterances (Tomasello, 2003; Abbot-Smith & Tomasello, 2006). To do this, they used a whole-form frequency manipulation with phrases. They compared imitation of phrases that differed in phrase frequency but were matched on all other frequency measures. For example, *a drink of milk* is more frequent than *a drink of tea* in British child-directed speech. But the two phrases are matched on substring frequency: *tea* is as frequent as *milk* (probably because of the British habit of drinking tea...), *of milk* is as frequent as *of tea*, and so on. They are also equally plausible. Any difference in performance has to reflect the properties of the full phrase itself. Two and three-year-olds were faster and better at repeating higher frequency phrases compared to lower frequency ones. Just as in the morphological literature, the authors took these effects to indicate lexical storage of the phrase.
4.2 The Current Study

In the current study I use a similar manipulation to assess the representational status of compositional multi-word phrases in adult language users. We can roughly distinguish between three more detailed models. In a words-and-rules model like that presented by Pinker and his colleagues (Pinker, 1999; Pinker & Ullman, 2002), such phrases are not part of the mental lexicon. Only non-compositional expressions, like idioms, are stored, treated as special vocabulary items that are learned and accessed like simple words (Jackendoff, 1995). This explains why idioms are processed faster than their non-idiom counterparts (Swinney & Culter, 1979). I call this a ‘strong’ dual-system model. In a ‘weak’ dual-route model, phrases that are of sufficient frequency can also attain independent representation as a way of making processing more efficient (Bybee & Hopper, 2001; Goldberg, 2006). Researchers differ on whether frequency is the only criterion for storage (as in the lexical bundles literature, Biber et al. 1999), or whether other factors also play a role in determining whether a phrase is stored (e.g., compositionality in the case of Goldberg’s Construction Grammar, 2006, or internal structure and context of use in Wray’s study of formulaic language, 2002). The ‘weak’ dual-route model maintains a distinction between phrases that are stored and ones that are not while allowing for more phrases to be stored in the lexicon.

No such distinction exists in a single-system model. In this framework, every instance of usage affects representation and processing. Compositional phrases are represented in the same way that simple words and non-compositional phrases are. The frequency of a phrase will influence its entrenchment and future processing (Bybee, 1998; Bybee, 2006; Bybee & Hopper, 2001; Bod et al. 2003). The difference between higher and lower frequency phrases is one of degree (the level of activation), and not of kind (stored vs. computed). This approach also predicts frequency effects for lower frequency phrases. It also predicts that that there will be a relation between the actual frequency of a phrase (the number of times it appears) and processing latencies.

This model draws heavily on usage-based theories of grammar (Bybee, 1998; Langacker, 1987; Tomasello, 2003) and single-system models of morphology (Elman, 1999; MacDonald, Pearlmutter & Seidenberg, 1994; Rumelhart & McClelland, 1986; Seidenberg, 1997). Since multi-word phrases are not discussed explicitly I extrapolate from how other compositional units are treated.
4.2.1 Previous findings

There is surprisingly little research that can allow us to distinguish among these three views. Many studies have shown that two-word (bigram) frequency affects processing. Object relative clauses are processed faster when the embedded clause consists of a frequent subject-verb combination (Reali & Christiansen, 2007). Pronunciation of words is phonetically reduced when the word appears as part of a frequent bigram (Bell et al., 2003, 2009; Gregory et al., 2004; Jurafsky et al., 2001). These studies show that people keep track of co-occurrence patterns for single words. But capturing such relations doesn’t require any representation beyond the single word.

Few studies have looked beyond the bigram level. In a seminal study, Bybee and Scheibman (1999) found that *don’t* was phonetically reduced in frequently recurring phrases (e.g., *I don’t know*). They argued that this provides evidence that very frequent phrases are represented in the lexicon. But although they extracted three-word sequences, they examined the effects of the preceding and following word separately, hence limiting their results to bigrams which could be modeled without representing larger units. Levy and Jaeger (2007) found that speakers were less likely to produce the optional relativizer in English relative clauses like *How big is the family (that) you cook for?* when the subject of the relative clause (*you*) was more predictable given the previous two words (*the family*). They show that a model that includes the last one, two and three words of the pre-relative clause utterance predicts speakers’ use of the optional relativizer, but because they do not report the independent effect of each string size (this was not the goal of their paper), we cannot know whether their results show an effect of three-word frequency when bigram and unigram frequency are controlled for. Bell et al. (2003) found that words were phonetically reduced when they were more predictable given both the previous and the following word (e.g., in the trigram *middle of the*, the predictability of *of* following *middle* and preceding *the*), again suggesting that speakers represent the phrase. But they did not find any effect when they looked separately at the predictability of a word given the two preceding or two following words. Moreover, this investigation was limited to the ten most frequent words in English.

Underwood et al. (2004) used eye-tracking to look at participants’ eye-movements while reading formulaic sequences of up to six words. They compared fixation times for
the same word in a formulaic sequence and in a non-formulaic one (e.g., \textit{fact} in: \textit{as a matter of fact} and \textit{it's a well-known fact}). They found fewer fixations when words appeared in formulaic sequences. They interpreted this finding as evidence that people represent the sequences as a whole. But there is an important limitation to this study. The authors did not control for the frequency of the substrings or for the frequency of the bigram that the words appeared in. Since those differed between the formulaic and non-formulaic sequences (e.g., \textit{of fact} and \textit{well-known fact}), the effect could have been driven by bigram frequency rather than sequence frequency.

The only study to control for part frequency is the one conducted by Bannard & Matthews (2008). Their findings present a challenge to dual-system models: Children showed frequency effects for compositional multi-word phrases. But the results are limited in several ways. First, the findings are limited to children. We do not know if the same effects will be found with adults. More importantly, they took high frequency phrases from the top third of the frequency scale and low frequency ones from the bottom third and only tested 12 phrases. As they stand, the results do not distinguish between a ‘weak’ dual-route model and a single-route one – they could still be accommodated if only very frequent phrases were represented.

The effects reported so far provide limited evidence that adults represent compositional multi-word phrases; we need more evidence from adults, with part frequency controlled for. Previous studies provide no evidence to distinguish between a single-system model and a ‘weak’ dual-system: To do that we need to look at the cases where the two accounts make different predictions. Only the single-system model predicts (1) that lower frequency phrases will be processed like higher frequency ones, and (2) that there will be a clear relation between the exact frequency value and processing latencies.

4.2.2 The information speakers are sensitive to

There is another incentive for asking whether processing is affected by phrase-frequency. There is mounting evidence that language users are sensitive to distributional information at many grain-sizes (sound patterns - Colman & Pierrehumbert, 1997, Hay et al. 2004; single-words - Morton, 1969; see Monsell, 1991 for a review; verb sub-categorization
frames – MacDonald et al., 1994, just to name a few). Such findings have influenced models of the lexicon. Word frequency effects for example, led to a revised model of lexical access (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986), where such effects could be captured. Finding that speakers’ maintained information about the voice a word was uttered in led to the creation of episodic models of the lexicon (Goldinger, 1996). Speaker-dependent phonetic effects fueled the development of phonetic exemplar models where such variability can be accommodated (Pierrehumbert, 2001). Phrase-frequency effects would have a similar effect.

We know little about the way the frequency of multi-word phrases (lexically realized phrases longer than two words) affects processing. Such effects would expand our understanding of the units people attend to – multi-word phrases in addition to words and constructions - and inform and limit the kind of the language models we use to accommodate frequency effects. Word and bigram frequency effects can be easily accommodated via links between words (or a non-symbolic representation of them) but frequency effects beyond the bigram (e.g. four-word frequency effects) call for the representation of larger chains of relations (sequential information), not only between single words but also between word strings of varying sizes.

More broadly, widespread frequency effects have led to the development of frequency-sensitive comprehension and production models that can account for the way different sources of information are integrated in real-time processing. Such models include (but are not limited to) constraint-satisfaction and expectation-based models of comprehension (Hale, 2001; Jurafsky, 1996; 2003; MacDonald, 1994; MacDonald et al., 1994; Tanenhaus et al., 1995; McRae et al., 1998). Similar frequency-sensitive models have been developed for production (e.g. Dell, 1986; Dell, Chang, & Griffin, 1999; Levy & Jaeger, 2007; Jaeger 2006). Uncovering the full range of distributional information that speakers are sensitive to becomes important for (a) developing adequate processing models, and (b) tackling the grain-size issue (Mitchell et al. 1995) that results from being able to estimate frequencies at multiple levels of linguistic analysis (e.g. syntactic construction, syntactic construction given a verb, syntactic construction given a verb and object, etc.). What is the relevant grain-size for a given calculation and how should
different frequency measures be integrated? To address this question, we need to know the full range of grain-sizes that language users attend to.

4.2.3 Study goals

In the current study I use the same whole-form frequency manipulation used by Bannard and Matthews (2008) to investigate the effect of phrase frequency on adult comprehension. The current study has several goals. The first is to see if adults are sensitive to the frequency of compositional four-word phrases when the frequency of the substrings is controlled for. Such effects are expected under a single-system model where compositional phrases are no different from any other experienced pattern. They are also expected if processing latencies reflect expectations derived from linguistic units of varying grain-sizes, including multi-word phrases. The second goal is to distinguish between a ‘weak’ dual-system model and a single-system one by looking for frequency effects along the continuum (also for lower frequency phrases) and by testing whether actual frequency predicts reaction times. A ‘weak’ dual-system model would predict frequency effects for very frequent phrases but not for lower frequency ones. Finding effects across the continuum would argue against the idea that very frequent phrases are represented differently from lower frequency ones. The third goal is a methodological one. Though the effect of frequency on processing is often assumed to be continuous (e.g., Bybee, 2006), in practice, items are often binned into two categories, high frequency vs. low frequency. By comparing how well a binary measure of frequency (high vs. low) predicts processing latencies compared to a continuous measure, I can test whether the assumption that effects of frequency are continuous is actually supported by empirical reaction time data, and how much better a continuous measure captures latency differences, compared to a binary one.

I investigate these questions by conducting two reaction time experiments where I compare processing latencies for pairs of compositional four-word phrases that differ in phrase frequency (the frequency of the four-word phrase) but are matched for part frequency (e.g. *don’t have to worry* vs. *don’t have to wait*). I then conduct a meta-analysis of the reaction times taken from the two experiments to ask whether a continuous measure of frequency predicts processing latencies and whether it does so
better than a binary measure. Processing latencies were measured using a phrasal decision task. People saw four-word phrases and had to judge whether they were possible in English. I use this task for two reasons. First, lexical decision tasks are often used in the study of morphologically complex words (Baayen et al. 1997). Since I am using a similar frequency manipulation (varying the frequency of the whole form vs. the parts), I wanted to also use a similar task. Second, the task allows for the presentation of the phrase as a whole and encourages participants to attend to each phrase as one unit. I controlled for the frequency of the sub-strings by comparing phrases that differed only on the final word, and by controlling for the final word, the bigram, and the trigram, both in the item selection and in the statistical analysis of the results.

4.3 Experiment 3: High and Low Frequency Phrases

In the first experiment in this chapter I compared the processing of multi-word phrases that differ in phrase frequency but are matched for substring frequency. I looked at two frequency bins. In each bin the cutoff point between high and low was set to a different value. In the high bin it was set at 10 per million – this is often considered a threshold for representation in the lexical bundle literature (Biber et al. 1999). Finding that very frequent phrases are processed faster than lower frequency ones would not distinguish between a ‘weak’ dual-system model (where very frequent phrases have special representational status) and a single-system one. To distinguish between those two approaches I looked at a lower frequency bin. In the low bin the cutoff point was set at one per million (high frequency phrases appeared more than once per million while low frequency ones appeared below once per million). If language users have whole-form representations for phrases, they should respond faster to phrases of higher frequency. If this is true for phrases across the continuum (and not just for very frequent ones) we should see similar effects in the two frequency bins.

4.3.1 Participants

Twenty-six students (mean age 20 years) from Stanford University participated in the study. All were native English speakers and were paid $10 in return for their participation.
4.3.2 Materials

I constructed 28 items (16 in the high cutoff bin and 12 in the low cutoff bin). Following Bannard and Matthews (2008), each item consisted of two four-word phrases (I counted orthographic words) that differed only in the final word ("don’t have to worry vs. don’t have to wait"). In each pair the phrases differed in phrase frequency (high vs. low) but were matched for substring frequency (word, bigram, and trigram). The phrases did not differ in the frequency of the final word, bigram or trigram. Any effect of phrase frequency could not be attributed to a difference in substring frequency. All phrases were constituents of the same kind (two VPs, two NPs, etc) that could form an intonational phrase.

The items were constructed using the Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) corpora that were combined to yield a 20 million word corpus. Both corpora consist of American English collected in telephone conversations. I used these corpora because I wanted to create items that were natural (appeared in spontaneous speech), that could form an intonational phrase, and whose frequency was not driven by a specific genre (e.g. Wall Street Journal). I selected all the 4-grams that fulfilled the following criteria: (1) the first trigram (e.g. "don’t have to" had to have a high frequency (over 30 per million); (2) the last word in the 4-gram (the one that differed between the high and low frequency variants) had a frequency of at least 50 per million. Both criteria were used to increase the reliability of the frequency estimates for low frequency sequences, and (3) the last word in the 4-gram could not be a determiner (which would create an incomplete intonational phrase) or a demonstrative. In final position a demonstrative like "that" could also be interpreted as a modifier (e.g. part of "that boy"). Since I do not know what additional processing that may entail, such items were excluded.

I selected our actual test items by choosing item pairs that had the same first trigram, and were matched for the frequency of the final word, bigram, and trigram (see Appendix A.1 for complete item list). I used the same corpus to calculate the frequency of the last word, the last bigram and the last trigram of each selected 4-gram. I selected 16 pairs of four-word sequences for the high frequency bin, and 12 pairs for the lower frequency bin. Table 4.1 shows example items for each bin.
Table 4.1: Mean frequency (per million words) and example items in the High and Low frequency bins

<table>
<thead>
<tr>
<th>High bin (High: 19.48, Low: 3.61)</th>
<th>Low bin (High: 3.5, Low: 0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t have to worry</td>
<td>Don’t have any money</td>
</tr>
<tr>
<td>15.3</td>
<td>2.35</td>
</tr>
<tr>
<td>Don’t have to wait</td>
<td>Don’t have any place</td>
</tr>
<tr>
<td>1.5</td>
<td>0.2</td>
</tr>
<tr>
<td>I don’t know why</td>
<td>I want to sit</td>
</tr>
<tr>
<td>35.5</td>
<td>3.6</td>
</tr>
<tr>
<td>I don’t know who</td>
<td>I want to say</td>
</tr>
<tr>
<td>7.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

High cutoff bin

In this bin the cutoff point between high and low was set at 10 per million words. In each pair, the high frequency variant appeared over 10 times per million and the low frequency one appeared under 10 times per million. The mean frequency of high frequency sequences (19.48 per million) was higher than that of low frequency sequences (3.61 per million), t(30) = 5.86, \( p < .001 \). But there was no difference in the frequency of the final word (high: 922 per million, low: 2235 per million, t(30) = -1.12, \( p = .27 \)), the final bigram (high: 295, low: 174, t(30) = .73, \( p = .47 \)), or the final trigram, (high: 40, low: 23, t(30) = 1.4, \( p = .17 \)) between the high and low frequency sequences. There was also no difference in the number of letters between the high sequences (mean 12.63 letters) and the low ones (mean 12.44 letters), t(30) = .25, \( p = .8 \).

Low cutoff bin

The cutoff point for the lower-range bin was once per million. I selected 12 pairs of four-word sequences for this bin. In each pair the higher frequency variant appeared between once and five times per million and the lower frequency variant appeared under once per million. I added an additional restriction that the lower frequency variants had to appear at least twice in the entire 20-million word corpus (.1 per million words). The mean frequency of high frequency sequences (3.5) was higher than the mean frequency of the low frequency ones (0.2), t(22) = 4.66, \( p < .001 \). There was no difference in the frequency of the final word (high: 433 per million, low: 278 per million, t(22) = 1.08, \( p = .29 \)), or the final bigram (high: 76, low: 32, t(22) = 1.44, \( p = .14 \)) between the high and
low frequency sequences. There was also no difference in the number of letters (high: 12.75 low: 12.33), \( t(22) = .44, p = .66 \).

Plausibility
I also wanted to control for the real-world plausibility of the events depicted by the low and high frequency phrases. To do so, I used an online survey to gather plausibility ratings for all the selected items. 25 different participants rated the selected items for plausibility on a scale from 1 to 7 (1 - highly implausible, 7 – highly plausible). Plausibility was defined as “describing an entity or situation that is likely to occur in the real world”. Selected items had a high plausibility rating in both bins (high cutoff bin: 6.66, low cutoff bin: 6.51). Importantly, high and low frequency phrases were rated as equally plausible in the high cutoff bin (high: 6.7, low: 6.7, \( W = 113.5, p > .5 \)), and in the low cutoff one (high: 6.6, low: 6.4, \( W = 43.5, p > 0.1 \), I used Wilcoxon tests because the ratings were skewed towards the plausible end of the scale).

Fillers
In addition to the 56 test items (16 pairs in the high cutoff bin, 12 in the low one, and two variants per pair), I created 80 four-word fillers (I wanted a higher number of fillers than test-items). Twelve of them were possible sequences in English and 68 were impossible ones. This design resulted in an equal number of possible and impossible sequences. Fillers were impossible either because they had scrambled word order (e.g. I saw man the) or because they had inappropriate prepositions (e.g. jump during the pool or put from the shelf). Most of the impossible fillers (75%) were scrambled and only 25% of them had an inappropriate preposition. Fillers were chosen so as not to overlap lexically with the experimental items.

4.3.3 Procedure
The experiment was run using Linger\(^6\). Participants sat in a quiet room in front of a computer and completed a phrasal-decision task. In this task participants saw four-word phrases centered on the screen and had to decide as quickly as possible whether they were possible sequences in English. Phrases appeared in their entirety (all four words

---

\(^6\) Developed by Douglas Rhode, http://tedlab.mit.edu/~dr/Linger
appeared at once on the screen) and response times were measured from the moment the phrase appeared. Participants used a keyboard to indicate if the phrase was a possible sequence or an impossible one. They used the ‘j’ key to make a ‘yes’ response and the ‘f’ key to make a ‘no’ response (the keys are equally positioned). *I saw the man* was given as an example of a possible sequence. *I saw man the* and *jump during the pool* were given as impossible examples. It was stressed that the sequences did not have to be full sentences to be judged as possible. Participants saw six practice items at the beginning of the experiment.

Each participant saw all 56 experimental items (the 28 item pairs). The task was divided into two blocks with one variant of each item appearing in each block. This was done to reduce possible priming effects from seeing two very similar sequences (e.g. *don’t have to worry* and *don’t have to wait*). The order of presentation of the blocks was counter-balanced between participants (half saw Block 1 first and half saw Block 2 first). Each block took about 5 minutes to complete. The blocks were separated by an unrelated lexical decision task which took five minutes to complete. None of the words from the main experiment were used in the distracter task. None of the words were semantically or phonologically related to the final words in the test sequences.

4.3.4 Results

Reponses under 200 ms and over 2 standard deviations from the mean per condition (high vs. low) were excluded. This resulted in the loss of 5% of the data. Accuracy for target items was at ceiling in both frequency bins for high (99%) and low (99%) items. The results are presented separately for the two frequency bins. I analyze the results using mixed-effect linear regression models with subject and item as random effects. I used log (response times) as the predicted variable to reduce the skewness in the distribution of response times. I added the log frequency of the final word and the final bigram as controls (I used log frequencies to correct for non-normal distributions). These were the only substrings that differed between the high and low variant of each pair, and while their frequency was matched, I wanted make sure that any effect of frequency was caused by phrase-frequency and not substring-frequency. These substring-frequencies were calculated from the same corpus used to select the target items.
In all analyses, I checked for collinearity between the fixed effects (for example between phrase-frequency and the frequency of the final bigram), and reduced it by regressing one of the collinear factors (the factor of interest, if one was involved) against the collinear covariates, and using the residuals of these regressions instead of the original variables in the final models I report (this reduced all correlations between factors to less than .19).

High cutoff bin
As predicted, participants were sensitive to phrase-frequency. They were faster to respond to high frequency phrases (mean 1056 ms) than to lower frequency ones (mean 1107 ms). Figure 4.1 shows the mean reaction times for the two frequency bins (the error bars overlap because this calculation doesn’t control for the different unigram, bigram, and trigram frequencies of the different item pairs, for instance the difference between don’t have to worry and a lot of work, but it illustrates the overall difference between the two frequency conditions).
Figure 4.1: Response times averaged over items for High and Low frequency variants in the High frequency bin (not controlling for part frequency), with 95% confidence intervals.

The final model had phrase-frequency (high vs. low), log(final-bigram), log(final-unigram), number-of-letters, and block-order (whether the item was seen in the first or the second block of the experiment) as fixed effects, and subject and item as random effects. Because low frequency was coded as the baseline, I expect the coefficient to be negative, indicating that high frequency trials were faster (took less time). The model showed a significant effect of phrase-frequency. Participants were faster on high frequency phrases when controlling for substring frequency, number of letters, and order-of-presentation, $\beta = -0.053$ (SE = .02), $p < .05$. Model comparisons using the likelihood ratio test showed phrase-frequency to be a significant predictor; the full model with phrase-frequency fit the data better than a model without it, $\chi^2 (1) = 8.93$, $p = .002$.

In addition to phrase-frequency, response times were affected by block order, and number of letters. Decision times were faster in the second block, $\beta = -0.11$ (SE = .02), $p$
Importantly, block-order did not interact with phrase-frequency \((p = .65)\); phrase-frequency affected both blocks similarly (I also conducted analyses on each block separately, and all reported results still held). Unsurprisingly, decision times were slower for phrases that had more letters, \(\beta = .03\) (SE = .01), \(p < .001\), with slower decision times for longer phrases. The frequencies of the final word and the final bigram were highly correlated, \(r = .42\), \(p = .02\). As a result, the coefficient estimates are not necessarily reliable. Instead I report the results of model comparisons using the likelihood ratio test (comparing the full model to one without the final word and one without the final bigram) which show that neither the final word, \(\chi^2(1) = 2.11\), \(p = .14\), nor the final bigram, \(\chi^2(1) = .004\), \(p = .94\), were significant predictors of processing latencies.

Low cutoff bin

As predicted by the single-system approach, but not by the `weak’ dual-system one, participants were also sensitive to phrase-frequency in the lower frequency range. They responded faster to items of higher (but still low) frequency (mean 1080 ms) than lower frequency (mean 1131 ms. see Figure 4.2, the error bars overlap for the same reasons mentioned for Figure 4.1).
Figure 4.2: Response times averaged over items for High and Low frequency variants in the Low frequency bin (not controlling for part frequency), with 95% confidence intervals.

As in the previous bin, I wanted to control for the frequency of the final word and bigram (those differ between the high and low variant in each pair). I ran a mixed-effect model with phrase-frequency (high vs. low), log(final-bigram), log(final-unigram), number-of-letters, and block-order (whether the item was seen in the first or the second block of the experiment) as fixed effects, and subject and item as random effects. I took the same measures as in the previous analysis to reduce any collinearity between the fixed effects.

In this bin also, participants were faster to respond to phrases of higher frequency, $\beta = -.06$ (SE = .02), $p < .02$. Model comparisons using the likelihood ratio test showed phrase frequency to be a significant predictor, $\chi^2(1) = 8.09$, $p = .004$. Again, a negative coefficient is expected because that shows that higher frequency phrases have shorter reaction times.
As in the previous bin, response times were also affected by block order, and number of letters. Decision times were faster in the second block, $\beta = -0.07$ (SE = .02), $p < .001$, and block-order did not interact with phrase-frequency ($p = .51$; I also conducted analyses on each block separately, and all reported results still held). Decision times were slower for longer phrases, $\beta = .03$ (SE = .01), $p < .001$). Because they were correlated, $r = .67$, $p < .001$, I estimated the effect of final word and final bigram using model comparisons using the likelihood ratio test (by comparing a full model to one without the final word and to one without the final bigram). Both final word, $\chi^2(1) = .16$, $p = .68$, and final bigram, $\chi^2(1) = .73$, $p = .39$ were not significant predictors.

4.3.5 Discussion

Experiment 3 set out to test the predictions that people are sensitive to phrase-frequency, and that this is true not only for ‘special’ very-frequent phrases, but for phrases across the frequency continuum. The results showed an effect of phrase-frequency on recognition times for phrases of varying frequency. Since sub-string frequency was controlled for, the effect could not have been driven by the frequency of the sub-strings the phrase is made up of. Since the high and low frequency phrases were also matched for real-world plausibility, it is unlikely that responses reflected knowledge about the frequency of the events depicted by the phrases. These are the first findings to show that four-word phrase-frequency affects adult processing latencies. They mirror effects found for children (Bannard & Matthews, 2008), suggesting that sensitivity to phrase-frequency is not limited to the developing lexicon. They add fully-realized multi-word phrases to the units that influence processing latencies.

These results imply that language users have some representation of the phrase itself to attach frequency counts to. That is, they store information about the compositional multi-word phrases themselves. As in the morphological literature (Baayen et al. 1997; Tuft, 1979), finding phrase-frequency effects suggests that on some level, people have an independent representation of such complex units. Such results provide evidence against a ‘strong’ dual-system model like words-and-rules (Pinker, 1999) – people seem to represent linguistic units that can be easily generated from their parts. Finding that more frequent phrases on the lower frequency range were also responded to faster is not
compatible with a ‘weak’ dual-system model of representation where only linguistic units of sufficient frequency can attain independent representation (Biber et al., 1999; Goldberg, 2006; Wray, 2002). Instead, the results are more compatible with a single-system view of representation, where multi-word phrases are one of the many linguistic patterns represented in the mental lexicon.

The results of Experiment 3 showed that language users are sensitive to phrase-frequency on the high and low end of the frequency scale. In experiment 4 I look at phrases from the middle of the frequency range that fall between the frequency ranges tested in experiment one. If phrase-frequency effects are found along the continuum, as predicted by the continuous approach, then mid-frequency phrases should be recognized faster than lower frequency ones. Experiment 2 serves an additional goal. I want to test the prediction that actual phrase-frequency will predict representation strength. To do this, I need observations for phrases across the frequency continuum. Looking at mid-frequency sequences will complement the low and high frequency sequences tested in Experiment 3. By gathering observations for items across the frequency continuum, I will also be able to conduct a methodological investigation of the relative merit of using a continuous measure of frequency (as opposed to a binary one) in predicting processing latencies.

4.4 Experiment 4: Mid Frequency Phrases

In this experiment I looked at the effect of phrase-frequency for a third frequency bin in between the high and low ranges of the first experiment. I set the cutoff between high and low frequency items at five per million (in experiment one the cutoff points were one per million for the lower bin, and ten per million for the higher one). High frequency phrases appeared between five and ten times per million, and low frequency ones appeared between once and five times per million.

4.4.1 Participants

Twenty-three students from Stanford University participated in the study. All were native English speakers and were paid $10 in return for their participation.
4.4.2 Materials

I constructed 17 test items (see Appendix A.2 for a complete item list). As in experiment one, each item consisted of two four-word phrases that differed only on the final word. In each pair, the phrases differed in phrase-frequency but did not differ in frequency of the final word, bigram, or trigram. The items were constructed using the same corpus and the same selection criteria used in experiment one. Only the cutoff point distinguishing high and low frequency items was changed. High frequency phrases appeared between five and ten times per million. Their low frequency counterparts appeared between one and five times per million.

The mean frequency of high frequency phrases (7.6) was higher than the mean frequency of the low frequency phrases (2.0), \( t(32) = 12.24, p < .001 \). There was no difference in the frequency of the final word (high: 2445 per million, low: 2267 per million, \( t(32) = .25, p = .8 \)), the final bigram (high: 658, low: 424, \( t(32) = 1.05, p = .3 \)), or the final trigram, (high: 44, low: 26, \( t(32) = .96, p = .34 \)) between the high and low variants. There was also no difference in the number of letters (high: 12.94, low: 13.06), \( t(32) = -.16, p = .86 \). As in experiment one, the selected items were also matched for real-world plausibility. 25 different participants rated the selected items using the same online survey used in experiment one. Selected items had a high plausibility rating (mean 6.05). Importantly, high and low frequency phrases were rated as equally plausible (high: 6.5, low: 6.3, \( W = 101.5, p > .1 \)). The same fillers were used as in experiment one.

4.3.3 Procedure

The procedure was identical to Experiment 3. Participants completed a phrasal-decision task in two blocks. One variant of each item appeared in each block. Unlike the previous study, the blocks were separated by a Stroop task that took 5 minutes to complete. In the Stroop task, participants had to give the font color of color words that appear on the screen. In Experiment 3 there was a strong effect of block-order: participants were much faster in the second block. The effect of block-order did not interact with the effect of frequency but we wanted to see if it would be reduced if we changed the distracter task from a lexical-decision task to a task that did not involve explicit linguistic judgment (namely the Stroop task).
4.3.4 Results

Responses under 200 ms and over 2 standard deviations from the mean per condition (high vs. low) were excluded. This resulted in the loss of 5% of the data. Accuracy for test items was at ceiling for high (99%) and low (98%) items. As in Experiment 3 I analyzed the results using mixed-effect linear regression models to predict logged reaction-times.

The results showed a similar pattern to that of Experiment 3. Participants were faster to respond to higher frequency phrases (1198 ms) compared to lower frequency ones (1276 ms., see Figure 4.3, the error-bars overlap for the same reasons cited in the previous two figures).

![Figure 4.3: Reaction Times by Condition (mid frequency bin)](image)

Figure 4.3: Response times averaged over items for High and Low frequency variants in the Mid frequency bin (not controlling for part frequency), with 95% confidence intervals.

As in the previous experiment, I wanted to control for the frequency of the final word and bigram (those differ between the high and low variant in each pair). I ran a mixed-effect
model with phrase-frequency (high vs. low), log(final-bigram), log(final-unigram), number-of-letters, and block-order (whether the item was seen in the first or the second block of the experiment) as fixed effects, and subject and item as random effects.

In this experiment also, participants were faster to respond to phrases of higher frequency, $B = .053$ (SE = .02), $p < .01$. Model comparisons showed phrase frequency to be a significant predictor, $X^2(1) = 9.25$, $p = .002$. The positive coefficient shows that indeed, low frequency phrases had slower reaction times.

As in the previous experiment, response times were affected by block order. Decision times still were faster in the second block despite changing the intervening task, $B = -0.08$ (SE = .01), $p < .001$, and block-order did not interact with phrase-frequency ($p = .81$). Number-of-letters was not a significant predictor, $B = .005$ (SE = .007), $p = .5$. As in experiment one, the frequency of the final word and the final bigram were highly correlated, $r = .42$, $p = .01$. I estimated the effect of the final word and final bigram using model comparisons (by comparing a full model to one without the final word and to one without the final bigram). The effect of the final word was significant, $X^2(1) = 5.9$, $p = .01$, but the effect of the final bigram was not, $X^2(1) = 1.17$, $p = .27$.

I also ran a mixed-effect model on each block separately, with frequency, log (final-bigram), log (final-word), and number-of-letters as fixed effects and subject and item as random effects. Because half the participants saw block1 first and half saw block2 first, I had observations for each test item in each block. The effect of phrase-frequency remained significant in both blocks (block1: $B = .05$, SE = .02, $p = .02$, block2: $B = .05$, SE = .02, $p = .02$). As in the full analysis, there was not a significant effect of number-of-letters (block1: $B = .003$ (SE = .006), $p = .57$, block2: $B = .01$ (SE = .02) $p = .29$). Also as in the full analysis, the effect of the final word was significant in both blocks (block1: $X^2(1) = 12.66$, $p < .001$, block2: $X^2(1) = 4.65$, $p = .03$) and the effect of the final bigram was not (block1: $X^2(1) = 2.75$, $p = .1$, block2: $X^2(1) = .45$, $p = .5$).

4.3.5 Discussion

As in Experiment 3 phrase-frequency had a significant effect on reaction times. Participants were faster to respond to mid frequency phrases than to lower frequency ones. Whether the cutoff point between high and low was set at ten, five (in experiment
three), or one per million (in Experiment 4), participants were faster on higher frequency phrases. Experiment 4 provides additional evidence that people store information about compositional multi-word phrases across the frequency range, and that their frequency influences processing.

I now have responses for phrases along the frequency continuum; from ones appearing less than once per million to ones appearing over ten times per million. Table 4.2 shows the ranges and the means of the three frequency bins I tested in Experiments 3 and 4.

<table>
<thead>
<tr>
<th>Frequency bin</th>
<th>Condition</th>
<th>Median</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lo</td>
<td>Lo</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1-0.4</td>
</tr>
<tr>
<td></td>
<td>Hi</td>
<td>2.5</td>
<td>3.5</td>
<td>1.3-9.7</td>
</tr>
<tr>
<td>Mid</td>
<td>Lo</td>
<td>1.9</td>
<td>2.0</td>
<td>0.8-4.1</td>
</tr>
<tr>
<td></td>
<td>Hi</td>
<td>7.1</td>
<td>7.6</td>
<td>5.4-9.8</td>
</tr>
<tr>
<td>Hi</td>
<td>Lo</td>
<td>2.8</td>
<td>3.6</td>
<td>0.6-8.9</td>
</tr>
<tr>
<td></td>
<td>Hi</td>
<td>15.4</td>
<td>19.5</td>
<td>9.1-44.8</td>
</tr>
</tbody>
</table>

Table 4.2: Item properties in the three frequency bins (in words per million)

I can use these data to test a prediction put forth by specific usage-based models that actual frequency predicts representation strength (Bybee, 2006). If true, then the higher the frequency, the faster recognition times should be. I did not test this prediction in Experiments 3 and 4. While I looked at phrases along the frequency continuum, I used a binary measure to model the results. In each bin I compared high frequency to low frequency. To test the prediction that actual frequency of occurrence predicts reaction times, I conducted a meta-analysis the items from Experiments 3 and 4, using log (frequency) as a predictor. This investigation serves another goal. The effect of frequency on language processing is often assumed to be continuous. But it is often modeled using binary measures (e.g., Grainger, 1990; Schiling, Rayner & Chumbley, 1998). I can now test whether the assumption that effects of frequency are continuous is actually supported.
by empirical reaction time data, and how much better it captures differences in processing latencies compared to binary groupings.

4.5 Meta-Analysis\(^7\)

I performed a meta-analysis of the items used in experiment three and four. The items from both experiments provide observations for phrases across the frequency range. These offer a flatter distribution of frequencies (with items spread out equally along the frequency continuum) instead of a bimodal one (with items divided by an arbitrary cutoff point). This distribution allows me to test whether actual frequency predicted reaction times, and whether it is a better predictor than a binary frequency measure.

4.5.1 Data and materials

I used all the items from Experiments 3 and 4 in the meta-analysis. This yielded 45 pairs of 4-word sequences (each pair had the same first 3 words). I took the reaction times of all the participants from experiments 3 and 4 (49 native English speakers). I only used the trials included in the analyses of the previous experiments (hence excluding trials that were answered incorrectly and trials with reaction times more than 2 standard deviations from the mean). This yielded a total of 2105 trials.

4.5.2 Results

I analyzed the data using mixed-effect linear regression models, as in Experiments 3 and 4. As a first step I wanted to see whether log (phrase-frequency) was a significant predictor of reaction times. As a second step, I wanted to see if it was a better predictor than a binary measure. To do this, I conducted a breakpoint regression analysis to find the breakpoint that best fits the data (the one where two models fit on either side of it have the maximum summed likelihood, Baayen, 2008). I then compared how well that measure (set at 4.94 per million) fared in comparison with a continuous measure (log phrase-frequency). By choosing the binary measure based on the most likely breakpoint in the data, I am biasing against our prediction that a continuous measure will be a better

---

\(^7\) This part in particular could not have been completed without the insight and guidance of Neal Snider.
predictor of processing latencies. Before conducting these analyses I had to address potential confounds arising from the use of items that are no longer as well controlled (I am no longer comparing pairs that differ only in phrase-frequency).

Reducing collinearity and over-fitting

In the regression models used in Experiments 3 and 4, I controlled for the frequencies of all the substrings that differed between the low and high variants (they only differed on the final word and bigram). Now that I am treating frequency as a continuous variable, each item pair is effectively treated as two items and the differences between items taken from different pairs are much greater. Items now differ also on the first tri-gram (e.g. don’t have to worry vs. go back to work). To look at the role of phrase-frequency I need to control for the frequencies of all the smaller elements (the two tri-grams, the three bigrams, the four unigrams). But this would risk over-fitting the results (I now have nine frequency measures in addition to the continuous frequency measure, the binary frequency measure, and the number-of-letter and block-order variables).

To address this, I ran a model with all the variables except the two phrase-frequency ones (the nine frequency controls, block-order, and number-of-letters) as fixed factors and with log (reaction-time) as the dependent variable. Following Baayen (2008), I then removed all the variables whose standard error was greater than the value of their coefficient in the model. This left us with six variables: four frequency control variables (log(unigram3), log(unigram4), log(bigram1), and log(trigram1)), block-order, and number-of-letters. These six variables will go into the full model that includes the two phrase-frequency variables (continuous and binary), along with the random effects of subject and item. In addition to block-order and number-of-letters, some of the frequency control factors still significantly (or marginally) predicted reaction times, even when 4-gram frequency was included: log(unigram4), \( p < .05 \) and log (trigram1), \( p < .09 \).

Analyzing the results

As predicted by usage-based approaches, log frequency of occurrence was a significant predictor of reaction times. I established this by comparing a full model with log (phrase-frequency) and the control variables (not including the binary measure for now), to a model without the continuous frequency measure. The difference between the models
was significant, $X^2 (1) = 14.86, p < .001$. Figure 4.4 plots the model fit for the reaction times to all sequences.

![Experiments 1 & 2 and model fit](image)

**Figure 4.4:** Model fit to reaction times for all phrases. Log reaction time by phrase frequency. Circles represent means for each bin, with 95% confidence intervals. The fit line is derived from a regression model with a continuous measure of frequency and all control covariates (it also includes 95% confidence intervals).

I now wanted to see whether the continuous variable accounted for more of the variance than the binary one. To do this I compared a full model with both frequency measures (and the control variables) once to a model without the continuous measure, and once to a model without the binary measure. The comparisons showed that the continuous frequency was a significant predictor, $X^2 (1) = 10.5, p < .002$, but the binary measure was not, $X^2 (1) = .69, p > .4$. In other words, when continuous frequency was taken into account, the binary measure was no longer significant; it didn’t explain any additional variance.
4.5.3 Discussion

These results are consistent with a single-system approach where every additional occurrence of a sequence strengthens its activation. The results also show that a continuous measure of frequency is a better predictor of reaction times than a categorical one, even when analyzed with a conservative model that controls for the frequencies of all the sub-strings and words making up the four-word sequence (although some of these still contribute), as well as participant and item effects. They provide qualitative support for the commonly accepted observation that the effect of frequency on language processing is continuous. This is not a controversial idea but there have been no previous empirical investigations that pit a binary variable against a continuous one. Methodologically, the results highlight the advantage of using (1) continuous measures of frequency as predictors of reaction times, and (2) statistical tools of analysis (e.g., regression models) where such continuous predictors can be used.

4.6 Discussion

4.6.1 Summary of findings

I set out to investigate whether phrase-frequency affects processing, and whether these effects are found not only for ‘special’, very frequent phrases, but whenever a higher frequency phrase is compared to a lower frequency one. Experiments 1 and 2 and the meta-analysis provided an affirmative answer to both questions: higher frequency phrases are responded to faster. The effect was found across the frequency range and was a gradient one. It was better captured when frequency was not binned but instead treated as a continuous variable. The meta-analysis showed a direct relation between frequency of occurrence and processing latencies: The more a phrase was experienced, the faster it was processed.

These effects cannot be attributed to part frequency – the pairs of phrases I compared were matched for the frequency of all sub-strings. It is also unlikely that they reflect a difference in the real-world likelihood of the events denoted by the phrases since they were rated as equally likely/plausible. Because the phrases differed on the final word, it would not be enough to store co-occurrence information for words or even two-word
sequences. To recognize that one phrase is more frequent than the other one would need to know how often the entire phrase appeared.

These results advance our understanding of frequency effects in several ways. First, they show frequency effects for four-word phrases. Though emergentist models (e.g. usage-based, connectionist, exemplar) have been around for over 20 years, there has been little empirical research testing their predictions for larger chunks of language. My findings fill an empirical gap, since phrase-frequency effects are predicted under such models, but were not previously reported. Second, my results highlight the loss of power when frequency is treated as a binary variable. While frequency is often viewed as a continuous variable, in practice, items are generally binned into two categories, high frequency vs. low frequency. By pitting a binary measure against a continuous one I demonstrate the advantage of using a continuous measure as a predictor of processing latencies.

4.6.2 Phrases in the lexicon

I started out by contrasting three views on the status of compositional multi-word phrases. The current findings are hard to accommodate within a words-and-rules model (a ‘strong’ dual-system model) where frequency effects are taken as a marker of lexical storage. Compositional units (regular words or compositional phrases) are not expected to display whole-form frequency effects because they are not stored. One way to explain these effects is to allow for compositional forms to be stored. This is the solution adopted by Ullman & Wallenski (2005) to account for the frequency effects found for regularly inflected words. The current findings would require this model to extend the lexicon dramatically to include many (if not all) compositional phrases. It is no longer clear what, if anything, remains outside the lexicon, thus undermining the distinction between the mental lexicon and grammar that these models depend on. The distinction can also be maintained by suggesting that compositional forms can be both ‘stored’ and ‘generated’. Sometimes phrases are stored (resulting in frequency effects) and other times they are
generated. This solution runs into an equally difficult problem. It is not clear when speakers use each type of phrase or how this can be empirically tested\(^8\).

The results are also not easy to accommodate within a ‘weak’ dual-system model that posits a unique status for very frequent forms. There was no indication of a clear difference between very frequent phrases and low frequency ones. Frequency effects were found for all the tested phrases. Using a frequency threshold as a determiner of storage is problematic because speakers cannot know a priori which phrases will become frequent enough to merit storage. Whatever information is maintained for very frequent phrases must have once been registered for all phrases. This information could be discarded at later stages of learning, but this seems unlikely. A similar argument can be made against using idiosyncrasy of meaning as a criterion for inclusion in the lexicon (Goldberg, 2006; Wray, 2002) From the perspective of the child learner who has yet to home in on the regularities of his/her language, all linguistic input starts out being idiosyncratic and ‘irregular’ to some degree.

The results are most compatible with single-system models of language where frequency is expected to affect all linguistic forms in a similar way. Compositional phrases showed whole-form frequency effects like those displayed by simple and inflected words. Furthermore, I found no evidence for a dichotomous distinction between very frequent phrases and all other phrases. More broadly, these findings argue against a clear distinction between the linguistic forms that are ‘stored’ and the ones that are ‘computed’. Instead, they enhance an emergentist, single-system view where all linguistic material is represented and processed in a similar fashion.

\(^8\) Ullman (Ullman 2001; Ullman et al. 2005 and many more papers) suggests that ‘stored’ and ‘computed’ forms can be distinguished on the basis of the neural substrates used to process them. This is part of a larger argument about the parallels between procedural memory and grammar, and declarative memory and the lexicon. He presents convincing evidence for double dissociations in patients. Patients who have impairments in procedural memory (e.g. who suffer from Parkinson’s) have a harder time with regular past-tense forms compared to irregular ones. Patients who have impaired declarative memory (e.g. Alzheimer’s) show the reverse pattern: they have more trouble with regular forms compared to irregular ones. These claims have come under attack (Penke & Westermann, 2007; Plunkett & Bandelow, 2006). Also, it is not clear how one would distinguish the neural substrates in non-impaired populations. Nevertheless, these findings provide important support for dual-system models.
The distinction between ‘stored’ and ‘computed’ forms

The distinction between ‘stored’ and ‘computed’ material is further blurred by recent findings on the processing of idioms. Idioms are often seen as prototypical candidates for ‘storage’ (Pinker, 1999; Jackendoff, 1995 but see Nunberg, Sag & Wasow, 1995 for an argument that only few idiomatic phrases are entirely non-compositional). However, several recent experimental results reveal parallels between the processing of idiomatic and non-idiomatic phrases. Sprenger et al. (2006) show that idioms can prime and be primed by words that appear in them (e.g. hit the road prime road), suggesting that like compositional phrases, they have internal structure. Konopka & Bock (2009) show that idiomatic and non-idiomatic phrasal verbs (e.g. pull off a robbery) can prime particle placement (whether the particle appears before or after the direct object) in non-idiomatic phrases that have no lexical overlap (e.g. knocked over the vase). Using acceptability judgment of familiar and invented idioms Tabossi, Wolf, & Koterle (2009) suggest that the syntax of idioms is governed by syntactic and pragmatic principles that are qualitatively like the ones that govern non-idiomatic language. These findings highlight the difficulty in distinguishing between ‘stored’ and ‘computed’ forms.

The difficulty in finding a clear criterion for inclusion in the lexicon leads Elman (2009) to the radical solution of “lexical knowledge without a lexicon”. Elman reviews numerous studies detailing the rich information language users have about verbs (from the agents it appears with to the discourse situations it evokes), and the way this information is rapidly used in online processing. To explain the rapid availability of such detailed, situation-specific lexical information in online processing, Elman suggests that “either the lexicon must be expanded to include factors that do not plausibly seem to belong there; or else virtually all information about word meaning is removed, leaving the lexicon impoverished”. He argues for a third alternative, an emergentist model in which linguistic knowledge is viewed as a constantly changing dynamic system and where the lexicon doesn’t contain fixed units but dynamic patterns. I propose that phrasal frequency effects similarly require a model that transcends traditional notions of the lexicon.
4.6.3 The information speakers are sensitive to

At a minimum, the current findings add multi-word phrases to the units that influence processing in adults. They show that language users are sensitive to co-occurrence patterns beyond the bigram level. This raises questions about how to integrate different frequency measures in a processing model, and how to capture and predict phrase-frequency effects when modeling linguistic knowledge.

As they stand, the results cannot be taken as evidence that the phrases were accessed as unanalyzed wholes – we do not know whether, and to what extent, the parts of the phrases were activated. My experiment was not designed to test this - the experimental items were designed to keep substring frequency maximally similar within an item pair (e.g. don’t have to worry vs. don’t have to wait). Looking back at the results of the meta-analysis, there is evidence that substring frequency (the frequency of the fourth word and the first trigram) still affected reaction times when controlling for phrase-frequency, as would be expected given the wide-spread word frequency effects in reading where words are read as part of a larger linguistic context. Nor can we know whether the phrases were processed incrementally faster due to their increased predictability - I only obtained reaction times for the entire phrase. More work using different experimental paradigms such as self-paced reading is needed to study the way part and whole frequency interact and the way phrase-frequency effects arise over time.

Word frequency effects led to revised models of lexical access (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986). Finding that speakers maintained information about the voice a word was uttered in led to the creation of episodic models of the lexicon (Goldinger, 1996). Finding speaker-dependent phonetic effects fueled the development of phonetic exemplar models where such variability can be accommodated (Pierrehumbert, 2001). Finding phrase-frequency effects can have a similar effect of extending existing models. Many of the models currently available focus on modeling frequency effects at the word level or below, or at the level of syntactic constructions. To capture phrase-frequency effects, such models would have to incorporate larger frequency relations.
4.6.4 A possible lexical model

One possibility, in line with exemplar models of language (Bod, 1998; Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001, 2006), is to implement the representations produced by the exemplar-based syntactic models of Bod (1998, 2006) in a spreading-activation network, as proposed in Snider (2008). In the model that Bod presents, syntactic productivity is achieved by starting with arbitrarily large linguistic units and deducing syntactic structure by means of statistical inference. The resulting lexicon has structurally analyzed chunks of different grain-sizes, along with a mechanism for constructing larger structures out of them. The processing of units is influenced by the probability of the smaller units used to form them (Bod 2006).

Implementing these representation in a spreading-activation network (Snider, 2008) will result in patterns of varying levels of abstraction (from fully realized strings of words, to fully abstract constructions) that are linked to each other, and whose activation is related to frequency of occurrence. Multi-word phrases are naturally represented in this model, and are linked to the words and smaller strings they consist of. For example, the phrase *don’t have to worry* would be linked to *don’t*, *have*, *to*, and *worry* as well as *don’t have*, *to worry*, and so on. Put differently, *don’t* would be linked to all the expressions it appears in (*don’t have to worry, don’t have to wait*, and so on). Multi-word phrases are also linked to the more abstract units they are instances of: verb-phrases, constructions, etc. The same would apply to all phrases, regardless of their frequency, and would lead to complementary representations at different grain sizes.

In this model, frequency effects on processing reflect a complex interaction between the various frequency measures of the different grain-sizes. This fits in nicely with previous findings that processing latencies are better captured when frequencies at different grain-sizes are taken into account simultaneously. For example, many studies have found that comprehension is affected by the sub-categorization biases of the verb (Trueswell & Tanenhaus, 1994; MacDonald, 1994; Garnsey et al., 1997; Gahl, 2002). But not all studies find such effects (Ferreira & Henderson, 1990; Kennison, 2001; Pickering et al 2000). Hare et al. (2003, 2004) suggest that the empirical discrepancies arise because verb-sense (an additional grain-size) was not taken into account. In similar spirit, Crocker and Brants (2000) reinterpret the early preference for an NP continuation
regardless of verb-bias reported in Pickering et al. (2000) as reflecting the overall lower probability of the S-complement analysis (Hale, 2003 and Jurafsky, 1996 also include the frequency of syntactic rules in their parsing models). The model described above naturally captured frequency estimates on multiple levels of linguistic analysis.

Phrase-frequency effects could also be modeled in other ways. It may be possible to capture four-word frequency effects in a simple recurrent network (Elman, 1991). Such networks have been used successfully to model syntactic processing (e.g., object vs. subject relative clauses, MacDonald & Christiansen, 2002). Combined with a dynamic system approach, they have been used to model local-coherence effects (Tabor et al, 1997) and thematic fit effects (Tabor & Tanenhaus, 1999) both of which require keeping track of the distributional contexts that words appear in. Phrase-frequency effects may be similarly modeled by keeping track of larger ‘frequency’ chains. However, there are no existing connectionist networks that set out to capture such relations.

6.4.5 Wholes and parts

Incorporating linguistic units at varying grain sizes raises interesting questions about the relation between multi-word phrases and two kinds of linguistic units (1) the substrings they are made of (e.g. ‘to worry’) and (2) the more abstract units they are instances of (e.g. an infinitive clause). Under the model sketched above, multi-word phrases do not lose links to their internal structure. The strength of the links may vary but they remain present over time. Because highly frequent phrases are represented in the same way, the same should hold for them. This is inconsistent with predictions made by specific usage-based models (e.g., Bybee, 2002) where the repetition of highly frequent phrases leads to them being stored and accessed holistically, as unanalyzed whole. However it is compatible with the recent idioms work mentioned earlier. This work demonstrates that idioms still maintain links to their internal structure (e.g., they can prime constructions). It would be interesting to extend this line of work to frequent, compositional, phrases to see whether such phrases also prime the constructions they are an instance of. For example, will a phrase like ‘give it to me’ prime the prepositional object variant of the dative alternation? If so, this would further support the notion that larger units maintain links to more abstract ones.
Kapatsinski & Radicke (2008) argue for competition between larger units and their parts when the whole-form is of sufficient frequency. Participants had to respond whenever they detected the particle ‘up’ in a verb-particle combination (e.g., give up). Reaction times were faster the more frequent the collocation. But for collocations in the highest frequency bin, there was a slowdown in reaction times. A similar result is reported in a separate study on the function word of (Sosa & MacFarlane, 2002) where detection of of was slower in highly frequent collocations like kind of (ones that appeared over 800 times per million). These findings are interpreted as evidence for competition between the part and the whole when the whole is frequent enough. Reaction times speed up when the particle is more predictable, given the sequence, but slow down when there is competition between the particle and the “chunked” collocation. This study suggests an interesting way to reconcile claims about the ‘special’ status of very frequent units (e.g. Bybee, 2002; Goldberg, 2006) with the current findings. I did not find evidence for a distinction between very frequent phrases and lower frequency ones. Phrase-frequency effects were found across the continuum. However, very frequent phrases may differ in the degree to which the parts activate the whole and vice versa.

4.6.6 Processing phrases

In the current study, I assumed that phrase frequency (the raw frequency of the string) is a relevant measure: higher frequency phrases were processed faster. But other measures are known to influence phrase processing, although they have mostly been looked at only for two-word sequences. Predictability is one: how predictable a word is, given the previous words. Mutual information is another: a measure of the mutual dependency of the two words. Roughly speaking, we can classify these into external factors that pertain to the phrase as a whole (e.g., phrase frequency), and internal ones that have to do with the internal relations within a phrase (e.g., mutual information). Both types of factors probably influence phrase processing. One way to reconcile the current study with the research on idioms is by incorporating internal measures as predictors. Idioms are processed faster than the same words in a non-idiomatic phrase (Cacciari & Tabossi, 1988; Swinney & Cutler, 1979), yet they don’t necessarily have a raw frequency count that is very high (though it may be higher than the same phrase in other configurations).
However, they rate highly on internal measures of predictability and mutual information, and this may contribute to faster processing (Church et al. 1991, but see Manning & Schütze, 1999 for limitations of this specific measure).

4.6.7 Limitations

The results of the current study are limited in that all the phrases that I used were constituents: verb phrases, noun phrases, prepositional phrases. They possessed some structural consistency. This doesn’t in any way undermine the effect of frequency, but I cannot rule out the possibility that people are only sensitive to the frequency of multi-word sequences that are also constituents. This would pose an interesting challenge for emergentist models of language. The phrases were always presented out of context. It is likely that like other linguistic units, the processing of multi-word phrases will also be influenced by expectations formed on the basis of prior linguistic context. In fact, finding that manipulating the linguistic context can affect phrase-processing would provide additional support for treating phrases as units of processing.

The results are also limited in that phrases always differed on the final word and that word was always a content word (e.g., *worry, wait*). I do not know if the same effects would hold when the phrases differ in function words, or when the words that they differ on are not in final position. For example, in a corpus study of word duration, Bell et al. (2009) found that different predictability measures affected the duration of function words and content words. Both content and function words were shorter when they were predictable given the following word, but only very frequent function words were sensitive to predictability given the preceding word. Further research is needed to see if these effects will hold for non-constituents, and when the different word is in non-final position. The results do not tell us why certain phrases are more frequent than others. They do not address the multiple linguistic and real-world factors that make certain linguistic configurations more frequent, but they show that whatever the underlying causes, frequency differences influence language use.
4.6.8 Phrases in parsing and production

Words have served an important role not only in models of the lexicon but also in parsing and production models. Word frequency influences interpretation: parses reflect the more frequent uses of a word (e.g. the garden-path caused by a sentence like *The old man the bridge*, in which *man* is used as a verb). But what if phrase frequency affects parsing in a similar way? For example, ambiguity resolution may be driven not only by how often a verb appears as a past participle and how likely a noun is to be an Agent, but also by the exact frequencies of the noun-verb combination. Patterns such as this have been observed in the processing of object relative clauses where chunk frequency influenced processing speed (Reali and Christiansen, 2007). If the effect of chunk frequency on parsing is widespread, then (1) parsing models will have to take into account chunk frequency, and (2) chunk frequency will have to be controlled in experiments. Production models have also assumed that creating an utterance involves a stage of word selection that is separate from the syntax level (Levelt, 1999). What if multi-word phrases are also selected in production? Speakers’ choices could be driven by a tendency to use constructions with higher phrase-frequency. These ideas must for the moment be considered speculative, but the current findings highlight the need to look at the role of multi-word phrases as well as single words in parsing and production.

4.7 Summary

I started this chapter by contrasting single and dual approaches to language. The representational status of compositional phrases is just one difference between these views, but it is an important one. It is important because it embodies larger differences in the way language is viewed; in what speakers need to learn and how that knowledge is acquired and represented. Larger units are an essential element of the lexicon if structure emerges from usage; they are the building blocks for extracting grammatical relations. They are not essential if the process of language use involves composing phrases from non-derived atoms. The findings of the current chapter show evidence of whole-form representation for compositional phrases. These findings reinforce an emergentist view of language where units of varying sizes co-exist in the lexicon.
The core of the Starting Big Hypothesis is in the emphasis on units that cross conventional word boundaries: on their role in learning and their status as psychologically real mental units that speakers attend to, store, and use. The main prediction is about first language acquisition – about how grammatical and lexical knowledge emerges from segmenting larger phrases of language. For this to hold, children must be able to store multi-word phrases. But the model also makes predictions about adults’ representation of phrases. If the developing lexicon can accommodate phrases, why wouldn’t the adult lexicon do the same? It is this question that connects questions of language acquisition and representation. What we learn about the units that children have can inform our view of grammar, and of how it is represented in the mind. In this chapter I have argued that adults still store multi-word phrases, and that this is a pervasive characteristic of language that should influence our models of linguistic representation.

In the next chapter, I test the main learning prediction made by the Starting Big model. If adult learners are worse language learners because they start off with more segmented units, then exposing adult learners to larger chunks may facilitate their language learning abilities.
CHAPTER 5: MULTI-WORD PHRASES IN ADULT LANGUAGE LEARNING

In the previous chapters, I looked at the role of multi-word phrases in children’s language and adult processing. I have shown that such units can facilitate children’s morphological production and affect adult processing – adults were faster to process higher frequency phrases. These findings highlight the role of larger units in native speakers’ language use, and suggest the intriguing possibility that by acknowledging the role of such units in native language use, we may gain insight into the seemingly unrelated puzzle of children’s better language learning skills.

5.1 Learning Language in Adulthood

In the current chapter, I test the prediction that part of the difficulty experienced by adult learners is related to the more analyzed and segmented units that they learn from. In a nutshell, the argument is that adults’ prior experience and knowledge about language leads them to learn from linguistic units where words are more clearly differentiated. This in turn may impede their learning of the grammatical relations between words (e.g., the agreement patterns between articles and nouns). I test this prediction by looking at adult learning of grammatical gender in an artificial language learning paradigm where I can manipulate the kinds of units that learners are initially exposed to. If adult learning is affected by the units that they learn from, then manipulating these units should result in changes in learning. Specifically, initial exposure to larger chunks of language should result in more child-like learning outcome.

5.1.1 Why is it so hard?

Why is acquiring a language to native proficiency in adulthood so difficult? Numerous studies have revealed that the expertise levels of native and non-native speakers diverge across many aspects of language; pronunciation (Moyer, 1999), morphological processing (Johnson & Newport, 1989), the use of formulaic speech and idioms
(VanLancker–Sidits, 2003). In accordance with the larger theme of this dissertation (the focus on multi-word units), I focus on non-native learning of grammar (morphology and syntax), and not on the way sound inventories (phonology) are learned. Given the many differences between children and adults, in terms of cognitive and neural development, and in terms of the social contexts in which they learn languages, it is perhaps not surprising that children and adults differ in their ability to learn. What is surprising, given adults' proficiency when it comes to learning in other domains, is that children appear to learn languages more successfully than adults.

Various approaches have been taken in seeking to understand this pattern. Lennenberg (1967) argues that adults no longer have access to a biological window of opportunity for learning language. Newport (1990) and Elman (1993) emphasize differences in cognitive capacity, suggesting that adult's increased memory hinders correct generalization by preventing them from ignoring some of the variability and complexity in their input (I will expand more on this approach in the discussion). Other researchers (Kuhl, 2000; Neville & Bavelier, 2001) highlight the changes in neural plasticity and the way early neural commitment shapes consequent learning (e.g., learning the phonetic distinctions that are relevant to your language changes the sensitivity to non-phonemic distinctions, Werker & Tees, 1984). Adults and children differ not only in their internal make-up (cognitive, neural) but also in the setting they learn language in. First and second language learners differ in the amount and kind of input they receive, and in the communicative setting they learn language in.

All these factors probably contribute to the different path and end-state of language learning in children and adults. But the challenge is to translate the differences between children and adults into concrete behavioral predictions about the linguistic domain in question. To do that, we need to identify the aspects of language learning that adult L2 learners have particular difficulty with, and explain how this difficulty can be traced back to any (or many) of the above-mentioned factors. For example, why is it that cognitive limitations result in worse learning of idiomatic language? How is it that biological limitations lead to the kind of learning outcomes seen in adult learning of morphology and syntax?
Theories of L2 learning often offer a link between certain characteristics of the learner, or the learning situation, and the general pattern of increased difficulty. But they rarely make explicit how these characteristics result in the specific learning outcomes observed (e.g., what is the mechanism that turns increased cognitive capacity into decreased learning of morphology?), and why they would affect certain linguistic domains more than others.

Overall, L2 learners show reduced proficiency compared to native speakers. But this is not equally true for all aspects of the novel language. For example, adult L2 learners seem to have little trouble learning vocabulary items, but struggle more with syntax and morphology (Weber-Fox & Neville, 1996). They show better learning early on of syntax and morphology, but not of pronunciation, compared to children (Krashen, Long & Scardella, 1979; Snow & Hoefnagel-Hohle, 1978). Also, the large literature on near-native speakers (Birdsong, 1992; Sorace, 2003) suggests that it is easier to reach native-like proficiency in some domains of language compared to others (e.g., syntactic constraints vs. discourse-based ones). A closer look at the Johnson & Newport (1989) study of L2 English-learners (often cited as evidence for an age-related boundary for second language learning) similarly reveals that the decline of linguistic ability with age was apparent in some but not all domains of grammar learning (e.g., there was no age effect for basic word order and the progressive). What is needed is an explanation of L2 learning that can predict this differential pattern of difficulty, and provide a systematic account of the mechanisms underlying it.

5.1.2 What adults learn from

Here, I propose looking at the linguistic units that adults learn from, and how they might differ from the ones children use, as a way of explaining differences between child and adult learning. I suggest that the different background knowledge that children and adults bring to language learning shapes the linguistic units they employ in early language learning, and this in turn shapes subsequent learning. Adults come to the task of language learning with a lot of prior knowledge about language; they know about words and grammar, and know the words and grammatical elements of their first language. Children, on the other hand, have none of this knowledge, and as a result are far more
likely to be learning segmentation, meaning, and structure, interdependently, at the same
time.

These differences in background knowledge may influence the linguistic units
learners employ: Adults learn from more segmented representations – with word
boundaries more clearly marked – while children begin with larger, less segmented
representations (including ones that cross word boundaries). The evidence presented in
chapters one and two of this dissertation suggests that children indeed attend to and store
units that cross word boundaries (chapter two), and that they make use of such units in
online production (chapter three). Here, I make the further prediction that learning from
more segmented representations may hinder learning about the relations between units.
Adult’s more segmented representations may make it harder for them to learn about
grammatical relations like gender-marking, case-marking, complex morphology, or even
idioms, where mastery is dependent on the ties between linguistic units.

I do not want to argue that the units of learning are the only factor underlying adults’
difficulty in achieving native-like mastery. But such a focus allows us to make headway
in tying specific characteristics of adult learners to specific patterns of learning. We can
now make concrete and testable predictions about the linguistic domains that adults may
have particular difficulty with (e.g., gender-marking), and the conditions under which
that difficulty may be attenuated (e.g., by initial exposure to larger chunks of language).
Focusing on the units of learning had the additional advantage of being relatively easy to
manipulate. While we cannot change the way an adult mind or brain has developed, we
can manipulate the kind of input learners are exposed to in learning a novel language.

To examine this idea, I focus on learning the agreement patterns between articles and
nouns in languages with grammatical gender, an aspect of language that non-native
speakers have considerable difficulty with (see e.g., Harley, 1979; Scherag, Demuch,
Roesler, Neville & Roeder, 2004). If some of this difficulty is related to the units that
adult learners employ, manipulating these units should result in changes in learning.
Having adults learn from larger units of language should enhance learning. Specifically,
starting with sequences of language in which the article and the noun are less
differentiated should facilitate learning of the relation between them.
5.2 Learning Grammatical Gender – A Case Study

Grammatical gender is a system found in many languages. It assigns all nouns (including inanimate ones) to noun classes, and marks neighboring words for agreement (Corbett, 1991). In Hebrew, for example, verbs and adjectives are marked for gender. In Spanish and French, articles have to agree in gender with the nouns they precede. Knowing a noun’s gender in gender-marking languages is essential for correct sentence construction.

5.2.1 Learning patterns in native and non-native speakers

Grammatical gender provides a good test case for studying differences between L1 and L2 learning. Native and non-native speakers show different patterns of learning and using grammatical gender. Children master grammatical gender relatively early (see Slobin, 1985 for cross-linguistic reports), and make few mistakes in spontaneous speech. In contrast, L2 learners have persistent difficulty with grammatical gender even after extensive exposure (Dewaele & Veronique, 2001; Holmes & de la Batie, 1999; Rogers, 1987; Scherag et al., 2004). Native and non-native speakers also differ in their ability to use the gender information conveyed by the article in real time processing. Native speakers (adults and children) can use this information to guide lexical access; they anticipate a feminine noun following a feminine article (Lew-Williams & Fernald, 2007a) and slow down if there is a gender mismatch between the article and the noun (Grosjean et al., 1994; Dahan et al., 2000). Non-native speakers do not show these effects (Guillelemon & Grosjean, 2001; Scherag et al., 2004; Lew-Williams & Fernald, 2007b). These findings suggest that native speakers treat the article and the noun as a more cohesive unit than do non-native speakers; this allows them to select the correct article in production, and use it to facilitate noun recognition in comprehension. Non-native speakers, on the other hand, seem to display a weaker association between the article and the noun.

Several suggestions have been advanced to explain these different patterns of grammatical gender learning, including: that grammar cannot be fully mastered in adulthood (Clahsen & Muyksen, 1986), or at least not aspects of it that are not found in the learner’s native tongue (Hawkins & Chan, 1997); that L2 learners have difficult accessing and retrieving gender information (Jiang, 2004); and that L2 learners form
more shallow grammatical representations that are hard to access in real time (Clahsen & Felser, 2006). These accounts describe the difficulty L2 learners have with grammatical gender, but they do not fully explain why this difficulty arises.

5.2.2 The units of learning

Could the units that learners start out with play a role in creating the different learning patterns? Researchers from various theoretical backgrounds have suggested that children initially treat the article and the noun as a single unit, rather than two separable ones, as an adult might (J. Carroll, 1939; S. Carroll, 1989; Chevrot et al., 2008; MacWhinney, 1978; Pinker, 1984). This would be a natural consequence of the way children encounter nouns, which most often is in the company of articles. This pattern is especially prominent in gender-marking languages (Bassano, 2008; Mariscal, 2009). Additionally, articles and nouns often form a single prosodic unit (Selkirk, 1995), which may contribute to their treatment as a single unit.

Numerous findings support this observation. Children’s early knowledge of articles appears to be lexically-specific. Instead of productively using articles with many nouns children often initially use a given article with a given noun. (e.g., produce only the definite article with one noun and only the indefinite with another, Bassano, 2008; Mariscal, 2009; Pine & Lieven 1997). Patterns of liaison acquisition in French (variation in pronunciation of the final consonant of certain articles depending on the beginning of the following word) also support the idea that articles and nouns are initially stored as a single unit; children make mis-segmentation errors where the liaison consonant is incorrectly treated as part of the noun (Chevrot et al., 2008; Dugua et al., 2008). For example, children mis-segment the article-noun combination un ours (a bear), to produce nour (with an initial consonant) instead of our (the correct form).

Adults on the other hand, may be less likely to treat the article and the noun as a single unit. Adult L2 learners may know that nouns and articles are separate entities from their experience with their first language, and the way they encounter nouns and articles, especially in a classroom setting, may emphasize their independence (Doughty & Williams, 1998; Lew-Williams & Fernald, in preparation). Whereas none of children’s early language input is written, adults are likely to learn from written input in which the
distinction between the article and the noun is explicit and visually salient. Finally, there is evidence that while adults can use cognitive control to selectively attend to particular aspects of the input, children may largely lack this facility (Ramscar & Gitcho, 2007). In other words, adults not only know that articles and nouns are separate, but they can also ‘choose’ to focus their attention on one or the other.

The findings reviewed here point to two different ways of learning grammatical gender. In the more ‘child-like’ mode, learners start by associating an object with an article-noun sequence. Over time, they learn to dissociate the article and the noun. In the more ‘adult-like’ mode, learners start by associating an object with its semantic label – the noun, and then learn which articles the specific noun can appear with. But this description on its own does not explain the different learning outcomes. Why is it that starting from the noun-label, and learning about the relation between the article and the noun from co-occurrence patterns results in a weaker association between the article and the noun?

5.2.3 Insights from learning theory

One possible explanation lies in learning theory, and specifically, in the effect of blocking on learning (Kamin, 1969). The basic notion here is of error-driven learning. Learners learn from violation of expectation. If an event is fully predicted, no further learning occurs. Blocking occurs when a new cue is introduced into a situation where a set of previously learned cues fully predict a response. In the absence of any discrepancy between what was encountered and what was anticipated, the new cue will not be associated with the event (Rescorla & Wagner, 1972; Rescorla, 1988). Blocking predicts different learning patterns depending on order-of-acquisition: what gets learned depends on what is already known.

To take an example from animal learning, if a pigeon learns to associate bell ringing with food (every time a bell rings, it receives food), it will have a hard time associating a new, simultaneously presented cue (say light-flashing) with the same event. Since the event of receiving food is fully predicted by the bell-ringing, the new cue of light-flashing will not become associated with the event. These principles have been applied to human learning in various domains, most notably in the realm of categorization and

They can be extended to grammatical gender in a relatively straightforward manner. Learners starting with noun-labels will initially associate an object and a noun-label: Their knowledge about the object will center on the noun. This will make it harder to later learn about the relation between the article and the noun: because the noun will fully predict the object, the article will add no information. Because the largest gains in associativity come in the earliest stages of learning (Rescorla & Wagner, 1972), the more adults treat articles and nouns as separate in these stages, and the more they associate the noun alone with an object, the less they will come to associate the article with the noun. In effect, initially focusing on the noun may cause learners to ‘listen through’ the article, because it doesn’t add any information.

In contrast, if learning starts with larger article-noun sequences, the initial association will be between the object and the article + noun sequence. Generally, an article can appear with many nouns, but a noun will appear with a more limited set of objects. Because of this, over time, cue competition will cause objects to become more strongly associated with nouns. The presence of a noun with an object but not the article will strengthen the association between the noun and object, and the occurrence of an article but not a given object or noun will weaken its association with them. Speakers will thus come to largely dissociate articles and nouns, but, crucially, the article will still remain associated to the object and the noun as a result of initial learning (Rescola & Wagner, 1972; see also Ramscar, Yarlett, Dye, Denny & Thorpe, submitted).

We now have a principled explanation of why starting out with noun-labels may result in a weaker association between the article and the noun. This explanation allows us to make concrete predictions about the conditions under which L2 learning of grammatical gender may be enhanced. Specifically, learning should be enhanced by initial exposure to article-noun sequences. I use an artificial language learning paradigm to test this prediction – we cannot manipulate the units that adult learners already possess in a language, but we can manipulate the units they are exposed to in an artificial language. The paradigm of artificial language learning has been used before to study how
arbitrary and non-arbitrary noun classes are learned. I review those findings in the next section before turning to the current manipulation.

5.2.4 Previous studies

A number of previous studies have used the artificial language learning paradigm to study how noun classes may be learned (Braine 1987; Braine et al. 1990; Brooks et al., 1993; Frigo & McDonald, 1998; MacWhinney, 1978; Williams & Lovatt, 2003). These studies manipulate the nature and kind of noun classes that learners are exposed to, most often, whether there are any semantic or phonological cues to class membership. They distinguish between arbitrary noun classes – where co-occurrence with a specific marker is the only cue for class membership, and non-arbitrary ones – where elements in a class share certain semantic or phonological features.

By and large, these studies have documented learner’s inability to learn arbitrary noun classes (but see Williams, 2003). That is, ones that have no additional phonological or semantic cues to class membership. Learners show above chance performance on trained items (Braine et al., 1990; Brooks et al., 1993; Frigo & McDonald, 1998)). But they show chance performance on novel items across studies (except in Williams, 2003), suggesting that they didn’t learn the ‘abstract’ noun-class category (a natural language equivalent would be not using the correct indefinite article if you have only heard the noun with a definite article). Once phonological (Brooks et al., 1993; Frigo & McDoland, 1998) or semantic (Williams, 2003) cues are introduced, learners show improved learning of trained items, and an ability to generalize to novel items. These studies have led researchers to suggest that learners cannot learn fully arbitrary mappings (and as a result, to characterize computational models that can, like that of Maratsos & Chalkely, 1980, as an inadequate model of human behavior, e.g., Frigo & McDonald, 1993).

But natural language does present learners with such arbitrary mappings. Gender classes often have semantic or phonological cues to class membership (Corbett, 1991; Kopcke, 1982). In German, for example, there are semantic and phonological correlates to class membership (Zubin & Kopcke, 1981), while in Spanish and French phonological cues are more prominent. But all those languages also include items for which gender assignment is non-transparent (some nouns ending with X in Y are masculine while
others are feminine), or even misleading (in Spanish problema has a feminine ending but is masculine). Yet even those items are easily learned by young language learners. Why then, do adult learners seem incapable of learning such arbitrary associations?

One feature that all previous studies have in common is that their training regime consists of an initial stage of vocabulary learning where nouns and markers are learned separately, followed by a subsequent stage of exposure to larger chunks of language demonstrating the relation between the nouns and the class-marker. In many of the studies (Braine, 1987; Braine et al., 1990; MacWhinney, 1978; Williams & Lovatt, 2003) the vocabulary learning stage lasted until participants had fully mastered the noun-labels. If the ideas sketched in the introduction are correct, then this type of exposure may have actually exacerbated adults’ difficulty in learning the noun-classes. Having adults learn nouns and markers separately may have hindered their ability to learn the arbitrary mappings between them. I discuss the concept of arbitrary mappings in language more in the general discussion of this chapter.

5.3. Experiment 5: Training Study

If adults' difficulty with grammatical gender results at least in part from their beginning with more segmented linguistic units, then adult learning should improve if the linguistic environment emphasizes larger linguistic units initially. To examine this, I created an auditorily presented novel language with arbitrary noun-classes, and contrasted the effect on learning of initially exposing adult learners to article-noun sequences, where the boundaries between articles and nouns are less prominent, with that of initially presenting them with the noun-labels as identifiable units (I use the term noun-label to refer to a noun appearing without an article).

Learners were divided into two groups. In the sequence-first group, learners were first exposed to a block of article + noun sequences in whole sentences and then to a block of noun-labels. In the noun-label-first group, learners were first exposed to noun-labels and then to full sentences. By the end of the experiment, both groups had received exactly the same input, but in different orders. By manipulating the initial units that learners were exposed to while keeping frequency of exposure constant, I could examine the way that initial learning with different sized units affected subsequent learning. I then assessed
how well participants learned the article-noun pairings. I predicted that participants in the sequence-first group will be more likely to produce the correct article for a given noun, and more able to detect a mismatch between the article and the noun. Like native-speakers, they should show evidence of a stronger association between the article and the noun than participants in the label-first condition.

5.3.1 Participants

Thirty-two native English-speaking undergraduate students at Stanford University were paid $10 for participation. Participants included 18 females and 14 males with a mean age of 20 years and three months.

5.3.2 Materials

The artificial language had 14 novel labels for familiar concrete objects (e.g., piano-‘slindot’), two articles (‘sem’ and ‘bol’) and a carrier phrase (‘os ferpal en’). The nouns were divided into two “noun classes”; each noun could only appear with one article. There were no semantic or phonological cues to class membership. Articles always followed the carrier phrase and preceded nouns (see Appendix A.3 for a complete stimuli list). An example of a full sentence in the language is given in (1).

(1) Os-ferpal-en bol slindot
Carrier phrase article-A “piano”

All noun labels were two syllables long. The objects were matched for familiarity, and for the frequency and Age-of-Acquisition of the English word (Stadthagen-Gonzalez & Davis, 2007). Participants were exposed to auditory stimuli of two kinds: noun-labels, and full sentences consisting of the carrier-phrase and an article + noun sequence. A male speaker recorded the carrier phrase, the articles, and the nouns separately. These were concatenated using Praat to create the full sentences. One recorded token of each noun, each article and the carrier phrase was used throughout the experiment to ensure that the nouns had the same prosody in full sentences and in isolation and that the articles had the same acoustic features with all nouns. The duration of the two articles was kept identical.
ensuring that neither had any acoustic prominence (sem = 410 milliseconds, bol = 409 milliseconds).

Another block of phrases in the artificial language was constructed in addition to the experimental items. This “distracter block” comprised the same carrier phrase, seven different noun-labels matched to seven additional objects, and two additional articles (‘tid’ and ‘gob). In contrast to the experimental items, the mapping between the articles and the nouns was not consistent (nouns could appear with either article). The “distracter block” was created to provide a buffer zone between training and testing.

5.3.3 Procedure

The experiment was divided into two phases: learning trials, and test trials. Participants were told they will be tested on the novel language and were asked to repeat the sounds they hear to enhance learning. The experiment lasted 25 minutes (20 minutes of training and around 5 minutes of testing). Training and testing sessions took place in a quiet room. All sessions were video-taped. Forced-choice responses and reaction times were collected using a response box. The experiment was run using e-prime 2.0.

Learning trials

Pictures of objects were presented on screen with an accompanying “description” in the artificial language. Participants were exposed to two kinds of stimuli: noun-labels and full sentences (carrier-phrase + article + noun) that were presented in separate blocks of trials. In noun-label trials, a picture of the named object was presented on screen alone; in full-sentence trials a picture of the named object was presented on screen along with a picture of a male gesturing to the object. Stimuli presentation was timed; objects appearing with full-sentences stayed on the screen for 3500 ms and objects appearing with noun labels stayed on the screen for 2000 ms. Participants in both learning conditions were exposed to the same number of noun-labels (each noun-label was repeated five time, a total of 70 labels) and full sentences (each noun in a sentence five times, a total of 70 sentences).

Participants in the sequence-first condition heard a block of full sentences followed by a block of noun-labels while participants in the label-first condition heard a block of noun-labels followed by a block of sentences. The only difference between the two
conditions was the order of the blocks. Following the two learning blocks, participants in both learning conditions were exposed to a distracter block of 35 sentences (accompanied by pictures of the objects). The distracter block was introduced to eliminate recency effects during testing, and ensured that the last block before testing was identical in the two learning conditions.

Test trials
Test trials were identical in the two learning conditions. Participants completed a forced-choice task and then a production task. In the forced-choice task, participants saw a picture, heard two sentences and had to indicate which sentence was the correct one in the language. They were told that only one sentence was correct.

Half of the forced-choice trials tested knowledge of the article + noun pairing. On these trials, the incorrect sentence had the right noun label but the wrong article (e.g. participants saw a piano and heard: *Os-ferpal-en *sem* slindot versus Os-ferpal-en *bol slindot*). The other half of the trials tested knowledge of the noun-labels. On these trials, the incorrect sentence had the right article but the wrong noun-label (saw piano and heard: *Os-ferpal-en bol *viltord* versus Os-ferpal-en *bol slindot*). Because participants heard a full sentence, they could also use the mismatch between the article and noun as a cue. Each object was presented once in an article trial and once in a noun trial yielding 28 forced-choice trials (half testing article knowledge and half testing noun knowledge). Order of presentation was randomized for each participant.

In the production task, participants saw a picture and had to produce a full sentence to describe it. They were encouraged to produce full sentences even if they were unsure about all the parts. There were 14 production trials (one for each object). Order of presentation was randomized for each participant. Responses were coded for accuracy by a research assistant blind to the study goals (reliability with coding by the author was high, $\kappa = .95$). Nouns and articles were coded as correct if they didn’t differ from the target in more than one sound (*slipdot* for *slindot*, and *vol* for *bol* were coded as correct). An article+noun sequence was coded as correct only if both the article and the noun were correctly produced. The carrier-phrase was coded for accuracy on a scale from 1-3 (1-fully accurate, 2-partially accurate, 3-not accurate).
5.3.4 Results and discussion

Forced-choice results
I excluded responses with RTs above 4500 ms. and below 150 ms (many of the lower reaction times were caused by a technical error in e-prime 2.0 that caused slides to be skipped). This resulted in the loss of 16% of the data.

As predicted, participants in the sequence-first condition showed better learning of the article + noun pairing. They were significantly above chance (61%) in choosing the correct article $t(15) = 3.55, p = .003$, while participants in the label-first condition were at chance $t(15) = .81, p > .4$. Figure 5.1 shows the proportion of correct responses by trial type (noun-label vs. article) and learning condition.

![Forced-choice accuracy by trial types and learning Condition](image)

Figure 5.1: Proportion correct by trial type (article or noun) and learning condition (sequence-first vs. label-first), with error bars marking +/- standard error from the mean.

A mixed-effect regression model with trial type and learning condition as fixed effects, and subject and item as random effects, revealed a main effect of learning condition that was not qualified by a significant interaction: Participants in the sequence-first condition
were more accurate overall (80% vs. 71% correct, B = .44 (SE = .21), p < .05). They were better at selecting both the correct article (61% vs. 54%) and the correct noun-label (98% vs. 92%). Not surprisingly, given the difficulty of learning grammatical gender, participants selected the correct noun-label more often than they selected the correct article (95% vs. 57.5% correct, B = 2.72 (SE = .26) p < .001).

Participants in the sequence-first condition also made correct responses faster than participants in the label-first condition (990 ms. compared to 1188 ms.). This was true when selecting the correct determiner (1194 vs. 1436) and when selecting the correct noun (854 vs. 1032). The effect of learning condition on reaction times was significant across items, F2 (1,13) = 6.57, p = .01, but not across subjects, F1 (1,30) = 1.63, p = .2. Overall, participants were faster to select the correct nouns than the correct articles, F1 (1,30) = 5.89, p = .01, F2 (1,13) = 16.32, p < .001. There is some evidence that the advantage of the sequence-first condition was evident in reaction times as well as accuracy measures.

Production results
I excluded responses made under 150 ms (these were caused by a technical error in e-prime 2.0 that caused slides to be skipped) and over 12,000 ms. This resulted in the loss of 9% of the data. The production results showed a similar effect of learning condition. Participants in the sequence-first condition were more likely to produce a correct article + noun sequence (40% of the time) than were participants in the label-first condition (25% of the time), B = .15 (SE = .06), p < .05. Overall accuracy rates were not high, which is not surprising given the short exposure time (20 minutes) and the number of noun-labels taught (14). Importantly, there was no difference between the groups in the production of the carrier-phrase, t(30) = -1.08, p > .2. That is, participants in the sequence-first condition showed better learning precisely of the association between the articles and the nouns. Figure 5.2 shows the proportion of correct article-noun sequences in each learning condition.
In summary, both measures (forced-choice and production) produced the same pattern of performance. Participants in the sequence-first condition showed better learning on all measures: recognition of the correct article, recognition of the correct noun, and production of the article + noun sequence.

5.4 Discussion

5.4.1 Summary of findings

The artificial grammatical gender system was learned better when participants started with “less segmented” input, where the boundaries between individual segments (in this case articles and nouns) were less prominent. Participants in the sequence-first condition were more likely to choose the sentence with the correct article in a forced-choice task,
and more likely to produce the appropriate article for a given noun in a production task. This was despite the fact that by the end of training, all of the participants had seen exactly the same training items exactly the same number of times. These results demonstrate an effect of order-of-acquisition (what gets learned first) and initial unit size (whether your start from more or less differentiated chunks of language) on learning. The relation between the article and the noun was better learned when learners were exposed first to larger, less segmented chunks of language.

As I noted in the introduction, there is reason to believe that adults are more likely to focus on noun-labels in learning. Thus, the current findings offer a novel perspective on why adult learners struggle to learn grammatical gender, and why the representations they learn are shallow and hard to access in real time (Clahsen & Felser, 2006). Starting from noun-labels may hinder learning about the relation between articles and nouns, despite the consistent co-occurrence patterns between them. Put differently, adults’ ability to focus on the semantic label of an object may in turn cause them to form a weaker association between the article and the noun. This tendency may be exaggerated for learners whose first language does not have gender categories, leading to an effect of L1 on L2 learning of grammatical gender (Franceschina, 2001; Sabourin, Stowe, & de Haan, 2006). By focusing on the units that adults learn from, and drawing on principles from learning theory, we can offer an explanation for what has so far been a description: adults’ weaker association between articles and nouns.

The notion of blocking plays an important role in the explanation offered here. One concern that may be raised is that the kind of experimental blocking I used doesn’t have a correlate in natural language learning. Learning in the label-first condition was hindered by initial exposure to a block of noun-labels, but it is unlikely that adult learners of natural languages are exposed to blocks of noun-labels before hearing article-noun sequences. Even in classroom settings, where noun-labels are often presented separately from articles, learners probably encounter article-noun sequences as well. Instead, the blocking manipulation should be understood as a way of simulating a more child-like learning strategy. When noun-labels and articles co-occur in speech, child learners ‘chunk’ them, because (a) that is how they appear, and (b) they don’t know better. Adults, on the other hand, are able to focus on the noun-labels, and this takes away from
their ability to treat the article and noun as a single, cohesive unit. In the artificial language, adults are stripped from their knowledge about what the relevant units are. Blocking enables us to focus their attention on one unit (the noun-label) or the other (the whole sentence). The effectiveness of this method highlights the importance of early exposure to article-noun chunks in L2 teaching, a strategy that has been suggested before (e.g., Holmes & de la Batie, 1999).

5.4.2 Comparison to previous studies

In the introduction, I reviewed previous studies using artificial language learning to explore how noun classes are acquired (Braine, 1987; Braine et al., 1990; Brooks et al., 1993; Frigo & McDonald, 1998; Williams, 2003). The current findings expand on previous ones by showing that the ability to learn arbitrary noun classes (ones that have no phonological or semantic cues to class membership) is enhanced when learners start out with larger chunks of language. Given that all previous studies have had a label-first set-up, where participants learn isolated vocabulary items before being exposed to the relations between them, the conclusion that adults cannot learn arbitrary noun-classes (Braine et al., 1990; Brooks et al., 1993; Frigo & McDonald, 1998) may have been too hasty. At least some of the difficulty reported in these studies may have been caused by the label-first training regime, and not by an underlying inability to learn certain kinds of noun classes. Fully arbitrary classes – where none of the items in a class have any shared features, may still be impossible to learn. Such classes, if they exist, are very rare in natural language. Nor do the current findings detract from the argument that noun classes are easier to learn when they have more shared features. But they do raise the need to examine how exposure type interacts with the ability or inability to learn arbitrary noun classes.

On a more philosophical level, we can ask whether first language learners need to be able to learn arbitrary classes. Natural languages provide learners with cues to class membership. But it is interesting to see how modern teaching methods tend to emphasize rote memory of larger chunks of language (e.g., Rosetta Stone, Berlitz) over learning of grammatical rules. There is a big literature on pedagogy in L2 learning that may both inform, and be informed, by the findings presented in the current chapter.

---

It is beyond the scope of the current chapter to evaluate and compare various L2 teaching methods. But it is interesting to see how modern teaching methods tend to emphasize rote memory of larger chunks of language (e.g., Rosetta Stone, Berlitz) over learning of grammatical rules. There is a big literature on pedagogy in L2 learning that may both inform, and be informed, by the findings presented in the current chapter.
membership. Kopcke (1982) identified 44(!) such cues for gender assignment in German. These cues allowed a model to predict gender membership with 90% accuracy, rendering the assignment less arbitrary. Such a model, however, differs from a child learner in two important respects. First, the model, but not the child, ‘knows’ that noun classes exist. Second, the model, but not the child, is ‘told’ what the relevant cues are. A child learner, on the other hand, may start out with a collection of individual exemplars and learn over time that there are noun classes, and that they have phonological and/or semantic correlates. Until this is achieved, gender-marking will remain arbitrary to some extent. Under this analysis, being able to learn arbitrary mappings becomes a more central component of early learning, and understanding the conditions under which such learning can be achieved becomes of greater relevance.

5.4.3 The effect of early units on learning

The present study offers a novel perspective on adults’ difficulty with grammatical gender by highlighting two aspects of learning: (1) the role of larger, less analyzed units in the learning process, and (2) the effect of early units on subsequent learning. While these aspects were only explored for grammatical gender, their effect on language learning is probably not limited to this domain. Instead, what this chapter offers is a view of L2 difficulty that is rooted in the units that learners attend to. This view naturally addresses two challenges that face theories of L2 learning: the challenge of how to link between specific features of the learner and the kinds of learning outcomes observed, and the challenge of why L2 learners find certain linguistic domains harder than others.

The focus on how early units shape later learning allows us to make headway on both questions. Adults’ prior knowledge (about language and about the world) results in more segmented linguistic representations which in turn lead to difficulty in learning certain grammatical relations (the how question). This difficulty may be especially pronounced when the relations in question are more semantically opaque (the why question). That is, when they do not correspond to observable semantic features (compare for example plural marking to grammatical gender). In these cases treating the two elements initially as a more cohesive unit may facilitate learning the relation between them. In the current experiment, nothing about the object could tell you which article it appeared with.
Learning that the article and noun could appear separately (because the article occurred with a variety of nouns) was easier than learning that they belong together.

5.4.4 Other linguistic domains

Seen in this light, the current results may reflect a broader pattern where learning units individually (articles and noun-labels in this case) has the cost of blocking later learning about the relations between them. Equipped with the link between early units and later learning, we can start to delimit the conditions under which starting from more segmented representations may hinder learning. What kinds of grammatical relations are more susceptible to the influence of early unit size?

One candidate set was proposed in the previous paragraph: grammatical relations where the link between units is more semantically opaque. The idea that semantic opacity makes it harder to learn grammatical relations receives support from a recent study of L2 learners of Spanish (Lew-Williams & Fernald, 2009). Previous studies have shown that non-native speakers have trouble using the gender information conveyed by the article to guide lexical access (Guillelemon & Grosjean, 2001; Scherag et al., 2004)). Lew-Williams & Fernald (2009) show that this effect is limited to objects that don’t have natural gender (ball compared to girl). Non-native speakers showed native-like processing of gender information for items like ‘girl’ or ‘man’. In these cases, the gender-marking of the article corresponds to an observable semantic feature of the object, making the relation easier to learn even when starting from a more segmented representation of the article-noun combination.

I explore two other domains that also exhibit a certain degree of semantic opacity. One is the pairing between verbs and prepositions (that you say elaborate on and not elaborate about) and the correct use of phrasal verbs, and the other is idiomatic language. In both domains the semantic relation between units is not always transparent, and both domains seem to pose a challenge for L2 learners (DeKeyser, 2007). Verb-preposition pairings vary cross-linguistically, in English you hit someone, but in Hebrew you hit to-someone. The choice of preposition is semantically motivated but there isn’t a one-to-one
mapping between the event type and the preposition used\textsuperscript{10}. For example, in Hebrew you hit \textit{to-someone} but kick \textit{in-someone} even though both events depict a similar negative impact delivered by body parts. In a similar way, the meaning of phrasal verbs is often non-transparent (e.g., the meaning of \textit{out} in \textit{figure out}). Idioms, by definition, are also semantically non-transparent. Their meaning is not derived solely from the literal meaning of their parts.

L2 learners seem to have trouble in both domains. The lack of idiomatic language is often seen as a hallmark of non-native fluency (Fillmore, 1979; Pawley & Syder, 1980; Wray, 1999, 2002). Empirical investigations suggest that non-native speakers don’t process idioms like native-speakers. Many of these investigations focus on phrases like ‘\textit{She missed the boat}’ that are ambiguous between a literal and figurative interpretation. Native speakers are faster to process the figurative meaning compared to the literal one (Gibbs, 1980; Swinney & Cutler, 1979; Tabossi, Fanari, & Wolf, 2009). Non-native speakers showed no such effect (Cieslicka, 2006). In addition, non-native speakers were less able to use prosodic cues to distinguish between figurative and literal meanings (Vanlancker–Sidits, 2003), and seem to judge idioms as more decomposable than native speakers (Abel, 2003). Looking at production, several researchers have noted the lack, or misuse, of idioms in non-native speech. In his study of a Spanish-speaking immigrant in New-York, Yorio (1989) notes the use of blended idioms like “\textit{it always strikes the mind of the employer}”, or “\textit{Give up their freedom of mobility}”. More generally, non-native speakers seem to treat idioms as more flexible and less cohesive than they actually are (Wray, 2004).

There is much less data on the acquisition of verb-preposition pairings and phrasal verbs by L2 learners. Anecdotally, my own use of English suggests that non-native speakers do have difficulty in these domains. The idea that L2 learners struggle with prepositions is prevalent in the literature (for example: “prepositions… are known to be among the most difficult items to master in a second language” Gass & Selinker, 2008: 47). But there is relatively little research to back up this claim. There is, however, some

\textsuperscript{10} The choice of preposition is motivated by many semantic factors, from which generalization can be extracted. I do not wish to argue that there aren’t semantic cues to preposition choice, but just to note that they are not always transparent or easy to deduce.
evidence that non-native speakers avoid the use of verbal phrases, especially ones that have a less literal meaning (Dagut & Laufer, 1985; Liao & Fukuya, 2004), and that they use fewer figurative or idiomatic verb-particle combinations (Yorio, 1989).

The interaction between initial unit-size, semantic opacity, and subsequent learning offers an interesting way to connect L2 learning patterns in these two different domains. In both domains, learners need to learn semantically non-transparent grammatical relations between adjacent linguistic units (verbs and prepositions, parts of an idiom). In both domains, learning may be hindered by initially treating the units as more separate. We can extend the predictions about the effect of unit-size to these domains. Learning may be improved by a ‘starting big’ process where smaller units are initially part of a larger chunk. In the case of verbs and prepositions, this would mean first associating an event with a verb-preposition pairing and only then constructing the semantic meaning of the preposition. In the case of idioms, learning may benefit from associating a ‘chunk’ with its non-literal interpretation. I would like to test both predictions in the future, as well as investigate further the differences in learning outcomes between native and non-native speakers. If the ideas presented here are correct, then L2 difficulty should increase with semantic opacity: the more non-transparent an idiomatic meaning is, the harder it will be to learn.

5.4.5 Unit size and statistical learning

Another way of stating the difference between the two learning conditions (label-first vs. sequence-first) is that what was learned about the relationship between units may be affected by the information they convey. This offers a way of reconciling the current results with the extensive research demonstrating speakers’ ability to detect and use co-occurrence information in learning language (see Saffran, 2003 for a review). Both children and adults learn transitional probabilities for sound sequences in a robust and reliable fashion (Saffran, Aslin, & Newport, 1996). However, participants in my experiment did not learn the relations between articles and nouns equally well in both conditions, even though they had access to the same co-occurrence information. They learned them better when they were segmenting speech and learning semantics at the
same time. That is, when learning the relation between the article and noun was part of learning about the relation between language and the world.

Unlike many other studies of statistical learning (especially ones concerned with word segmentation), the current study had a semantic component. Learners were exposed to pairings of objects and speech. Using the artificial language to ‘talk about’ the world alters the distributional structure of the input – there are now co-occurrence patterns between linguistic units and objects as well as between the linguistic units themselves. It also alters the task learners are engaged in, by adding a ‘communicative’ component of understanding how the novel language corresponds to the world. These factors may in turn alter what participants learn. Participants will learn transitional probabilities when that is all that is present in the input (as in many word segmentation studies). But when there is other information in the input (e.g., semantic), participants will learn what is most relevant to the task at hand (for relating objects and speech, the co-occurrence patterns between linguistic elements and objects). This fits in nicely with a more ‘communicative-oriented’ view of statistical learning, where learners extract the statistical information that best helps them make sense of language and the world. Applying these ideas to word segmentation, Frank & his colleagues (Frank, Mansingha, Gibson & Tenenbaum, 2006; Frank, Goodman & Tenenbaum, in press) argue for a referential view of word segmentation where transitional probabilities are useful precisely because they help learners in the task they are already engaged in: finding out what things are called.

5.4.6 Order-of-acquisition effects in language learning

The current results also underline the effect that prior knowledge has on what gets learned: if you already know noun-labels, you learned about articles differently. There are many other examples of this: the initial sound patterns children learn influence the acquisition of later forms (Kuhl et al., 1992); the stress pattern of children’s first words in English affects the segmentation of later words (Swingley, 2005); they are more likely to pick up new words that conform to their existing production templates (Velleman & Vihman, 2002); their knowledge about the phonological contrasts in their language influences how speech-sound variation will be interpreted (infants will interpret ‘tan’ and ‘taan’ as two words in Dutch, where vowel length is a contrastive features, but as one
word in English, Dietrich, Swingley & Werker, 2007). The current findings join previous ones in highlighting the way prior knowledge shapes subsequent learning in non-obvious ways.

5.4.7 Starting big or starting small?

In this chapter, I’ve argued for a link between the units of learning, and what gets learned. This chapter cannot be concluded without considering two previous proposals that tie early unit size to differences in L1 and L2 learning: Newport’s Less is More (1990), and Elman’s Starting Small (1991, 1993). In many ways, the current study is inspired by this previous research, which first suggested that properties of the learner (reduced working memory) can affect the units used in learning. However, my proposal differs from earlier ones in emphasizing the advantage of larger units, rather than smaller ones. Below I discuss the two proposals in more detail, and the consequences for the role of larger units in learning.

Newport’s less-is-more hypothesis (1990) was developed against the backdrop of generative explanations that attributed children’s better language learning to the availability of a time-sensitive “language acquisition device” (Chomsky, 1965). Newport, instead, proposed a link between children’s reduced cognitive abilities in other domains (e.g., working memory), and their greater ability to learn language. She proposed that children’s more limited memory resources are actually an asset in language learning because they allow learners to ignore variability and focus on smaller linguistic units as a basis for learning grammar. This idea has considerable intuitive appeal; it links two disparate and well-studied observations (better language learning and smaller working memory), and elegantly turns a disadvantage into an advantage. But its validity is dependent on demonstrating a causal link between reduced processing capacity and enhanced language learning. As stated by Rohde & Plaut (2003): “[…] for such a theory to be plausible, the potential benefit of limited resources must be demonstrated both computationally and empirically”. In their paper, Rohde & Plaut present an eloquent and convincing critique of the findings cited in support of the less-is-more hypothesis. I present their arguments and elaborate on them below.
There are relatively few studies that set out to empirically test the less-is-more hypothesis. Among them are computational simulations (Elman, 1991, 1993; Goldowsky & Newport, 1993) and experimental studies (Cochran, McDonald & Parault, 1999; Kersten & Earles, 2001). Rohde & Plaut (2003) raise substantial doubts about their intended support for the less-is-more hypothesis; they question their empirical validity and the way ‘less’ is operationalized.

Elman (1991, 1993) found that a connectionist network could learn an English-like artificial grammar only when ‘starting small’: when either the network’s memory or its training corpus was initially limited (I return to how ‘small units’ are defined here in the next section). Rohde & Plaut (1999) fail to replicate these findings and instead find that starting with simplified input (only simple clauses) hinders learning, especially when the language encodes semantic relations between nouns and verbs (e.g., verbs appear more with certain nouns, like in real languages). The effect of limited memory was only found when starting with simplified input, and it did not lead to performance that was better than in any of the complex training regimes. Moreover, Elman’s computational model (1991, 1993) is often seen as a direct test of the less-is-more hypothesis (this is the stance also in the Rohde & Plaut paper). But a closer look reveals that the units Elman considers as small are different from the ones proposed by Newport and in fact consist of multi-word units. When the input to the network is manipulated, ‘starting small’ means starting with simple clauses instead of embedded ones; when the memory of the network is limited, ‘starting small’ means having a window of 3-4 words. In both cases, ‘small’ units correspond to multi-word phrases of a minimum of three words, this is very different from the truncated morphemes Newport imagines children’s building blocks to be. Rohde & Plaut (2003) also fail to replicate another computational result reported by Goldowsky & Newport (1993) whereby learning complex form-meaning mappings was improved when parts of the input were randomly removed by a filter (one could ask whether this the right way to simulate reduced memory capacity). Rohde & Plaut point out that in the original study, filtering was only beneficial for the lower performing model (that selected the correct mapping only 20% of the time). They then demonstrate that learning in a connectionist network is actually impeded when a filter is introduced.
The results of the experimental studies are also questioned. Cochran et al. (1999) report that non-deaf participants learned agreement patterns in American Sign Language better when they were placed under high memory load. The main problem with this study is that participants were asked to extract a generalization that was not supported by the input (to detect verb-types based on only one example, see Rohde & Plaut, 2003 for more details). What looks like better learning may actually reflect more random performance (participants in the high-load condition also tended to omit motions altogether). Kersten & Earles (2001) argue that participants exposed to ‘staged’ input (one word at a time) showed better learning of an artificial language than ones who were exposed first to full sentences. But this was true only for the vocabulary items that were learned first and did not extend to learning of grammatical relations (which are the kinds of relations argued to benefit from reduced memory in the less-is-more hypothesis). Both experimental studies can also be criticized for the way ‘less’ is implemented. It is not clear that engaging simultaneously in a tone-counting task (as in Cochran et al. 1999) or receiving one isolated word at a time (as in Kersten & Earles, 2001) accurately simulate children’s smaller memory capacity. Several additional studies (not reviewed by Rohde & Plaut) fail to find a facilitative effect for starting with smaller or reduced input (Conway, Ellefson & Christiansen, 2003; Ludden & Gupta, 2000). In summary, despite the intuitive appeal of the less-is-more hypothesis, there is little empirical evidence to back it up.

Rodhe & Plaut (1999, 2003) point to the weakness of the empirical and computational support for the less-is-more hypothesis, but there are also more conceptual issues at stake. When laying out the proposal, Newport (1990) suggests that children perceive and store only parts of the complex linguistic input they are exposed to. But it is not clear what these parts correspond to, or how they would lead to better learning. This vagueness is a major limitation. Do children simply truncate their input after a certain amount of time or syllables? If so, how is this parameter determined? A window of X seconds in English would result in a very different kind of unit than the same window in a morphologically agglutinating language like Turkish. Moreover, how will children learn morphology from morphologically impoverished representations where only parts of the words are maintained? The way ‘less’ is implemented in Goldowsky & Newport (1993) suggests that children randomly filter their input (though I know of no evidence that this is the
case). This may do more harm than good by creating a less consistent and messier distribution. Newport (1990) cites children’s omission of morphemes as evidence for their truncated representations, but we know that infants attend to elements that they are not yet producing, suggesting that at some level, these elements are represented (Hirsh-Pasek & Golinkoff, 1996). For example, preverbal infants can perceive function words (Shafer, Shucard, Shucard & Gerken, 1998; Shi, Werker & Cutler, 2006), and 18 month-old infants can use function words as a cue to comprehension even at a stage when they often omit them in production (Gerken, Landau & Remexz, 1990; Kedar, Casasola & Lust, 2006). In sum, there may be other, more plausible ways to define the smaller units children learn from, but to date, no such definition exists. This state of affairs casts doubt on the less-is-more enterprise. It is hard to argue for the benefit of learning from ‘parts’ until we know what these ‘parts’ look like.

The Starting Big approach is similar to the less-is-more hypothesis in the belief that the units of learning have an effect on the learning outcome. But it differs in suggesting that adults, and not children, are more likely to ‘break language down’ in ways that may be adversary to learning. This is not to say that children cannot or do not segment their input. It is clear that children can segment speech very early on. But it is also clear that they are capable of storing larger chunks of language early on as well (see literature review in chapter 1). Here, I argued that it is children’s tendency to under-segment (not over-segment) that contributes to their better language learning abilities. The Starting Big approach makes more concrete predictions, and provides a clearer definition of the units of learning (prosodic units) than the less-is-more hypothesis. Unit size is not seen as an across-the-board explanation for all the differences between first and second language learners, but as a useful way to understand differences in certain linguistic domains.

5.5 Summary

I have shown in this chapter that adults were better at learning grammatical gender in an artificial language when they were first exposed to full sentences and only then to noun-labels, demonstrating an effect of unit size and order-of-acquisition on adult learning. Adults showed improved learning of grammatical relations when exposed first to larger, and less differentiated chunks of language. These results enhance the Starting Big
prediction that starting with larger units, and slowly increasing segmentation with learning, may prove advantageous. More broadly, the chapter outlines a view of L2 difficulty that relates the units that adults learn from to their difficulty in learning certain aspects of a second language. This account makes concrete predictions that should be tested in future research. In the next chapter, I re-evaluate the Starting Big Hypothesis in light of the findings about the role of multi-word phrases in children’s production of irregular plurals (chapter 2), adult processing (chapter 3), and L2 learning (the current chapter).
CHAPTER 6: MULTI-WORD PHRASES IN LANGUAGE LEARNING AND USE

6.1 Summary of Findings

I started this dissertation by asking why children are better language learners than adults. I hypothesized that the answer lies, at least in part, in the linguistic units that children and adults learn from. Specifically, children learn from units that are larger and less analyzed than the ones adults learn from. This process, of using larger units to learn about smaller ones, can lead to better learning of certain grammatical relations (the Starting Big Hypothesis = SBH). The idea that early units can shape subsequent learning, and that larger units may play a crucial role in achieving native-like proficiency, prompted an investigation into the role of multi-word phrases (units that cross word boundaries) in language learning and use. Before we can argue that L2 learners make insufficient use of larger units in learning, we have to first show that such units are an integral part of the native speaker’s inventory.

In chapter 2, I brought together existing findings from production and perception to establish that young children can, and do, represent multi-word phrases, a crucial assumption of the SBH. Among the findings cited were infants’ ability to remember multi-word phrases (Mandel et al. 1994, 1996); children’s use of under-analyzed multi-word phrases in early production (Peters, 1983); and children’s sensitivity to whole-phrase frequency in later production (Bannard & Matthews, 2008). In the same chapter, I documented the prevalence of big-to-small processes (Gestalt processes) in first language learning. While combinatorial (small-to-big) processes are often emphasized in developmental research, children’s mastery of many linguistic domains can be understood in terms of a move from larger less well-analyzed units to smaller, more structured ones. Such processes can be found in the acquisition of segmental phonology (Ferguson & Farwell, 1975; Velleman & Vihman, 2002), morphology (Rubino & Pine, 1998), and construction learning (Brandt et al., 2009; Rowland & Pine, 2000). These
findings provide support for the advantage, postulated in the SBH, of learning smaller units from larger ones.

The next chapter presented novel evidence that multi-word phrases can facilitate children’s language. Children were better at producing irregular plurals in English when they were preceded by a familiar frame like *So many*. Performance was further enhanced when the nouns were preceded by a frequent sentence-frame like *Brush your* before *teeth* or *Three blind* before *mice*. That is, children not only represent multi-word phrases, they can also use their knowledge of them to facilitate the production of smaller units, namely words. The findings also show that lexical knowledge is not independent of the linguistic context that words appear in, and support the prediction that multi-word phrases are part of the inventory of units that children learn and use.

Chapter 4 showed that such units are also part of the adult linguistic inventory. Adult speakers were faster to process higher frequency compositional four-word phrases like *don’t have to worry* compared to lower frequency ones like *don’t have to wait*. Since the phrases were matched for part frequency (the frequency of all the unigrams, bigrams and trigrams) and plausibility, the effect must reflect speakers’ knowledge of the frequency of the whole-phrase. That is, on some level adult speakers have memory traces of phrases that could be easily generated. These results were brought to bear on the debate between single-system and dual-system approaches to language. In dual-system accounts (e.g., Pinker, 1999), but not in single-system ones (e.g., Rumelhart & McClelland, 1986), there is a distinction between ‘atomic’ and ‘derived’ forms. The former are stored in the mental lexicon while the latter are generated by grammar. Finding that speakers represent ‘generated’ forms, like compositional multi-word phrases, undermines this distinction and suggests instead that all linguistic experience is processed and represented by the same cognitive mechanism.

Chapter 5 provided support for the prediction that adult difficulty is related to the more segmented units they learn from. Adults showed enhanced (and more child-like) learning of grammatical gender, a notoriously hard domain for L2 learners, when first exposed to larger chunks of language. Participants who heard nouns as part of a sentence (with articles) before hearing them on their own (noun-labels) were better than participants who heard nouns on their own before hearing them in sentences. In accord
with the SBH, learning was facilitated when participants were learning segmentation, semantics (the labels for objects), and grammar (which article a noun appears with) at the same time. These findings constitute the first empirical demonstration of two components of the SBH: (1) early units can influence subsequent learning, and (2) there is an advantage to learning grammar by segmenting larger units.

Together, the findings presented in this dissertation highlight the importance of multi-word phrases in child language, adult processing, and second language learning. They are consistent with many of the components of the SBH as presented in the introduction, and raise additional questions. In the next sections I discuss some implications and open questions in three domains: child language learning, adult language learning, and the units of language more generally. I conclude by presenting a revised Starting Big Approach that emphasizes the importance of larger units in learning and processing.

6.2 First Language Learning

6.2.1 The nature of early units

In this dissertation, I emphasized how the units children learn from can influence what gets learned. To understand how children are learning, we need to identify what they are learning from. We need to determine what their early units look like. This task, which is not easy to accomplish, is crucial for any theory of learning that rests on generalizing from input – we need to discover the units that form the basis for generalization. As discussed in chapter 2, this is quite difficult to do. By the time they start talking, children have already learned a lot about their language, moreover, the units they produce may differ from the ones they represent or learn from (Clark & Hecht, 1983).

The prosodic factor

We can look for a possible answer in the literature on infant speech perception. Theories of grammatical development rarely integrate findings about children’s developing ability to perceive and produce speech. The two fields of study, learning grammar and learning sounds, remain largely separate. But what children can attend to in speech must shape the linguistic units that they extract. We can learn about their early units by looking at the
cues they first use to segment speech: Early units may correspond to what infants are first capable of detecting in fluent speech.

These units may be equivalent to clausal-level boundaries, which are marked by various acoustic properties (see Cooper & Paccia-Cooper, 1980; Beckman & Edwards, 1990). Infants are sensitive to the acoustic correlates of major prosodic units (namely clauses) very early on, but take longer to detect smaller prosodic units like phrases or words. By two months, they remember speech better when it is packaged in a well-formed prosodic clause (Mandel et al. 1994, 1996). By 4.5 months, infants can detect prosodic correlates of clauses in fluent speech (Jusczyk, 1997). By 6 months they are capable of recognizing a clause previously heard in infant-directed speech (Nazzi et al. 2000), and of extracting a clausal unit from a multi-clause passage (Soderstrom et al. 2005). The ability to use prosodic cues to detect smaller prosodic units, like phrases, develops later. By 9 months, but not before, infants are sensitive to phrase boundaries (Jusczyk et al. 1992), but even then, they do not perceive all phrase boundaries (for example, they don’t notice a pause if it is inserted between a pronominal subject and a main verb, Gerken et al. 1994). A month later (at 10 months), infants can perceive phonological phrase boundaries (the next unit on the prosodic hierarchy, Nespor & Vogel, 1982) in fluent speech. After being familiarized with a target word (e.g., paper), infants turned their head more often when they heard the target word itself than when they heard its constituent syllables separated by a phonological phrase boundary (paper vs. pay#per, Gout, Christophe & Morgan, 2004).

Infants gradually develop the ability to detect smaller prosodic units in fluent speech. This pattern matches the perceptual salience and cross-linguistic consistency of prosodic cues to clause boundary. The cues to clause boundaries (like pauses, syllable lengthening, and a decline in F0 for English, Cooper & Paccia-Cooper, 1980) are

---

11 There are studies showing that new-born infants are sensitive to the acoustic correlates of phonological phrases (Christophe et al. 1994, 1996, 2001): they distinguish between the same sequence extracted from within a Spanish word (geLATIna), and across word boundaries (goriLA TiMida). This may seem to contradict my suggestion that infants gradually develop sensitivity to smaller prosodic units. But these studies did not look at fluent speech: they do not demonstrate that infants use such cues to segment fluent speech but just that they can detect acoustic differences between two isolated tokens.
acoustically more salient than the cues to phrase (or word) boundaries (Cho & Keating, 1999; Cooper & Paccia-Cooper, 1980; Shattuck-Hufnagel & Turk, 1996), especially in infant-directed speech (Fisher & Takura, 1996a, 1996b). They are also more consistent across languages (Venditti, Jun & Beckman, 1996; Juscyzk et al. 1992). Phrase-level prosody is more variable, and more dependent on language-specific features (like head-directionality) than clause-level prosody. Since infants do not know which language they are born into, it makes sense for them to detect cues that are more robust cross-linguistically before others that are more specific to particular languages. “This suggests a picture of the development of prosodic competence that proceeds from larger, more language-general units (the clause), through intermediate units (phrases) to smaller, more language specific units (the word).” (Soderstrom et al., 2003: 263).

This developmental trajectory suggests that children’s early units correspond to major prosodic boundaries. These are the boundaries that infants detect first. They are similar across languages, and they are acoustically salient. In fact, they are more salient in infant-directed speech than in adult-to-adult-speech. One of the characteristics of infant-directed speech is the use of exaggerated prosody (Ferguson, 1964; Fernald, 1983; Fernald & Simon, 1984). Utterance-level boundaries in particular are more prominent in infant-directed speech compared to adult speech across languages (Bernstein-Ratner, 1986, 1996; Fernald & Simon, 1984; Fernald et al., 1989; Morgan, 1986; Stern, Spieker, Barnett, & MacKain, 1983), making them even easier to detect. Six-month-old infants were better at detecting clause boundaries in paragraphs of child-directed speech compared to the same material spoken in adult speech (Kemler Nelson et al., 1989).

This is not to say that all of children’s early units are multi-word ones. Given the nature of child-directed speech, extracting units that correspond to major prosodic breaks will result in a mix of single words, short phrases, and multi-word ones. Around 10% of child-directed speech is made up of single-word utterances (Aslin et al. 1996; Brent & Siskind, 2001; Fernald & Simon, 1984), which can be used to aid segmentation, recognition, and vocabulary learning (Brent & Siskind, 2001; Fernald & Hurtado, 2006). Children’s input also includes many sentence-fragments. Up to 18% of the speech spoken to English children consists of multi-word fragments like over here or show me produced with utterance-like prosody (Fisher & Takura, 1996a; Cameron-Faulkner et al., 2003).
These fragments may assist learning by providing a more accessible demonstration of certain grammatical relations (Fisher & Takura, 1996b; Snow, 1972). Many of the multi-word units that children extract will be (a) relatively short and (b) match syntactic boundaries. Of the utterances directed to English-speaking children, 56% had three words or fewer (Bernstein-Ratner, 1996, see also Snow, 1972), and, unlike in adult speech, more than half the syntactic clauses in speech to children are followed by major pauses (Fisher & Hakuta, 1996a).

The perceptual findings offer a way of systematically predicting what early units look like. We can refine the Starting Big notion of ‘learning from larger units’ to mean that children learn from units that correspond to major prosodic boundaries (often utterance boundaries). Given children’s input, these early units, which form the building blocks for later learning, are not restricted to whole clauses, but include words, sentence fragments and multi-word phrases.

This prosodic view of early learning allows us to make concrete predictions about children’s initial inventory while taking into account what children’s input looks like in terms of (a) its prosodic characteristics and (b) its more fragmented nature. It also provides a more principled explanation for why children may extract multi-word units. There is nothing a-priori special about multi-word phrases, instead, these units happen to correspond to the prosodic boundaries that infants can first attend to, and that adults make prominent in speech to children. If children extract units based on major prosodic boundaries, they will end up with many multi-word ones.

6.2.2 Multi-word chunks

Another factor that may increase the likelihood of a multi-word phrase being included in children’s early inventories, and treated as one chunk (where the whole phrase is treated as one, not fully analyzed, unit), is if it is highly collocated - if its components frequently co-occur, more frequently than expected given the frequency of their parts. The treatment of such sequences as one unit is attested in several computational models of speech segmentation (Brent & Cartwright, 1996; Goldwater, Griffiths & Johnson, 2009; Swingley, 2005). These models were not built with the goal of identifying multi-word chunks. But, using different cues to segment speech into words (stress, transitional
probabilities), they all end up with some ‘word’ units that consist of more than one word. Such units are often discarded as errors that are insignificant to the learning process, but they may actually provide insights into the nature of early units.

Swingley (2005) extracted early word-forms by clustering together syllables that score high on frequency and mutual information (how informative syllables are about one another). In the English data (the Dutch data patterned similarly), 40 of the 44 bi-syllabic ‘false alarms’ (two-syllable sequences erroneously classified as one word) were frequently co-occurring word pairs like come on or give us. Of the 22 tri-syllabic false alarms, 15 were well-formed multi-word phrases like want some more and mummy’s here (sequences longer that three syllables were not investigated). Using mutual information between syllables as a cue to word boundaries resulted in a substantial number of multi-word ‘words’.

Using transitional probabilities as a cue to word boundaries also led to a number of multi-word ‘words’ (Goldwater et al. 2009). When words were assumed to be statistically independent (an implicit assumption in artificial language learning experiments on speech segmentation, Saffran et al. 1996), the resulting lexicon had many two and three-word ‘words’ - 30% of these proposed ‘words’ were multi-word strings like look at this or do you see. Even when words were treated as predictive of other words (a more ecologically valid assumption but one that requires infants to attend to conditional statistics), the model ended up with quite a few multi-word ‘words’ - 15% of the proposed ‘words’ still contained more than one actual word. A similar pattern is found when distributional regularity (co-occurring together frequently and in a variety of contexts) and phonotactic constraints are used to segment speech (Brent & Cartwright, 1996). At least 10% of the proposed word candidates are classified as ‘concatenation errors’ where two or more words were not properly segmented (since the authors only report this number for one of their simulations, we cannot know how widespread they were in the other models).

In all of the computational simulations reviewed above, the proposed lexicon had a substantial proportion of multi-word ‘words’. Whatever segmentation criterion was applied, it led to the treatment of some multi-word sequences as one unit. Instead of treating these as undesirable errors we can use them to learn about the kinds of multi-word sequences that are more likely to be treated as one chunk. A closer look at these
'errors' reveals several interesting patterns. Many of them fall in the category of socialized routines: expressions like *bye bye, thank you,* or *night night*; and positive exclaimations like *good girl!* or *clever boy!* Such expressions are unlikely to appear in diverse linguistic or social contexts (we don’t often say *thank them*), and hence may be treated as an unanalyzed chunk.

Determiner-noun sequences are also repeatedly identified as one word. Of the 70 most frequent multi-word ‘words’ in Goldwater et al. 2009, almost 20% consisted of determiner-noun combinations (tellingly, the only example in Brent & Cartwright (1996) of a concatenation error is the determiner-noun chunk *thekittie*). These findings lend credibility to the claim, made in chapter 5, that children treat articles and nouns as a chunk early on. Several different models, using a variety of cues, end up with article-noun ‘words’ (cross-linguistically, articles and nouns often form a single prosodic unit, Selkirk, 1996).

Among the remaining examples are frequent frames like *do you want,* wh-questions like *who’s that,* auxiliary-noun combinations like *is he* or *are you,* and labeling frames like *that’s a* and *it’s a.* I want to outline two possible functions of such chunks in learning. These units may aid learners in discovering grammatical patterns by providing a relatively constrained space for comparison. For example, by noting the difference between *is he* and *are you* the child can learn that certain auxiliaries appear with certain pronouns, a pattern that may be harder to discern from larger, more complex sequences of speech. Labeling frames in particular may serve another purpose. Processing them as one unit may free up resources to focus on the new word being offered (Fernald & Hurtado, 2006).

Whatever the function of such units in learning, it is clear that the postulated lexicons include a considerable number of multi-word ‘words’. If children use segmentation strategies that are anything like the ones used in the simulations (not an unlikely scenario since these were inspired by behavioral findings), then they too will end up with quite a few multi-word units. The simulations offer further insight into the kinds of sequences that remain chunked even after segmentation has taken place; they include socialized routines, determiner-noun combinations, and certain frequent frames.
6.2.3 Units of production or perception?

So far, I’ve spoken about the units of learning without distinguishing between production and perception, but the units children rely on in the two modalities may differ (Clark & Hecht, 1983). Children often comprehend more than they can produce – their receptive vocabulary is typically much larger than their productive one. They also seem to have more adult-like units in comprehension compared to production. One classic example is the ‘fis phenomenon’ (Berko & Brown, 1960) where children reject an adult use of their own mispronunciation. If an adult says *fis*, following the child’s own use, the child will reject it, accepting only the correct *fish* pronunciation. Another example is that of a child that comprehends the difference between *mouse* and *mouth* but produces the same form (/maus/) for both (Smith, 1973).

These discrepancies, which are not limited to the realm of word learning, suggest that children have distinct representations for production and comprehension – what they produce may be different from what they comprehend or learn from. Learning, in fact, may advance by aligning these two representations - by reducing the difference between their own speech and that of adult speakers (Clark & Hecht, 1983). The distinction between units of production and units of perception is important because it can reconcile the role of larger units with children’s own more truncated productions.

One possible objection to the Starting Big Hypothesis stems from children’s production: if children start out with larger chunks, why is much (but not all) of their early speech limited to single words? One answer, alluded to above, is that children produce only a portion of what they represent. Just as children’s truncated word forms (like *apu* for *apple*, Clark & Hecht, 1983) don’t mean that they lack a representation for the whole word, single-word utterances do not show they lack larger units. In fact, children can recognize whole words even when the same words are truncated in their own productions (e.g., Dodd, 1975). The perceptual evidence described above makes a similar point: Pre-linguistic infants have memory for multi-word utterances, even though they clearly cannot produce them. Children’s truncated productions, therefore, cannot serve as evidence against the role of larger units in learning.
6.2.4 Identifying multi-word chunks in production

Multi-word chunks are also more prevalent in child speech than often assumed. Children attempt to produce larger units early on (chapter 2), and draw on such units in their production of smaller ones (chapter 3). Their frequency in early speech is probably grossly under-estimated - such utterances often go undetected because researchers are looking for more conventional units like words (Peters, 1977). Even when children produce utterances that appear to be made up of adult-like words, they may still be relying on multi-word chunks.

The computational and perceptual findings reviewed above show that children extract multi-word units from speech, and offer ways to determine what these units correspond to. But these findings are limited to units of perception. We need a parallel measure to identify multi-word units in production: units that could be generated, but may be used as one chunk.

Wong-Fillmore (1976, 1979), in a pioneering study of child second language learning of English, offers several criteria for identifying what she calls ‘formulaic expressions’ in child speech. She classifies sequences as formulaic if they exhibit any combination of the following properties: (1) they have a single invariant form, (2) are used repeatedly, (3) are situationally dependent, (4) are grammatically advanced compared to other constructions the child produces (e.g., show correct subject-verb agreement not found in other utterances), or (5) are community-wide routines (greetings, exclamations, idioms, etc.). Between 53% and 100% (!) of children’s language at the earliest time of study (the percent varies from child to child) was defined as formulaic, suggesting that early utterances are much less productive than they seem. Some utterances are easy to classify. After only 7 weeks of exposure to English, Nora (aged 5;0, one of the learners), produced the unquestionably formulaic utterance *I don’t know, I think I oughta know* coupled with the appropriate gesture. Others are harder to identify. Wong-Fillmore’s method leaves much room for subjective judgment, and is dependent on what the researcher happened to record – missing one creative use could change an utterance’s characterization from productive to formulaic.

Other studies have relied on productivity measures, calculated from child speech, to identify chunked language. An utterance was considered novel only if it wasn’t produced
before, in its entirety, by the child. In one study, only 35% of two-year-olds utterances were classified as novel, and of those, 75% could be derived from previous utterances by using a single combinatorial operation like addition or substitution (Lieven et al. 2003, see also Lieven, Salomo, & Tomasello, 2009). Early language, then, isn’t as productive as it may seem. More importantly, 50% of the first 400 identifiable multi-word utterances were classified as frozen: there was no evidence that children had productive knowledge of their parts (Lieven, Pine & Baldwin, 1997). To be classified as frozen, the utterance could not contain words that occurred on their own, or in the same sequential position in another multi-word utterance (I return to this definition later).

Recently developed computational tools may offer a more systematic way of identifying multi-word units in children’s production. Borensztajn, Zuideme & Bod (2009) use data oriented parsing (Bod, 1998), in combination with statistical parsing techniques, to “identify the most likely primitive units that were used […] to produce the utterances in a given corpus”. Utterances are decomposed into the fragments they were most likely generated from - these fragments are the models’ building blocks. The authors used this method to show that children’s grammars become more abstract with age: early fragments (called constructions in the paper) had more lexical nodes than later ones.

This method can also be used to uncover multi-word chunks at different stages in development – at any given point, the fragments that have no slots (non-terminal nodes), and have more than one word in them, are likely to be stored and used as one unit. The authors don’t say how many of the fragments they extracted fit this definition (they only looked at the ones that did), except to say that there were many. But we can use their method to detect candidate chunks in development, and follow their decomposition. By looking at children’s utterances over time, we can see the way earlier fragments are broken up into smaller more productive units. Unlike the productivity measure used by Lieven and her colleagues, the computational analysis can be applied to a much larger corpus. Having to code all of the child’s previous utterances for lexical content and sequential position to classify a unit as ‘frozen’ (Lieven et al. 1993) naturally limits the size of the corpus that can be investigated. The computational approach also offers a quantitative way to assess the size and level of abstraction in children’s productive units.
6.2.5 Segmenting early units

In the previous sections, I focused on what children’s early units look like, and concluded that they contain quite a few multi-word units (both in production and in perception). In this section, I want to briefly discuss how early units may be segmented. There is a huge literature on the cues that infants use to segment speech, among them transitional probabilities, prosodic cues, phonotactic constraints, stress patterns, and more (see Gervain & Werker, 2008; Jusczyk & Luce, 2002 for reviews).

Some of these cues can be used to segment early units – calculated over the larger chunks; they can differentiate the smaller units within them. Acoustic markers of unit boundaries seem especially useful in this process, but other cues, like stress may also be used. Children’s increasing sensitivity to smaller prosodic boundaries (e.g., Soderstrom et al. 2003) will allow them to segment larger prosodic units into smaller ones, discovering syntactic and semantic regularities along the way (as has been argued for in the prosodic bootstrapping literature, Morgan, 1986; Morgan & Demuth, 1996; Peters, 1983). Children can also discover parts, and the relations between them, by analogy between stored units (Bod, 2009; Peters, 1983; Wong-Fillmore, 1976) For example, comparing the phrases *give it to me* and *give it to Daddy* marks the final noun as a separate unit, and signals a similarity relation between ‘me’ and ‘Daddy’ in the syntactic and semantic roles these terms can fill. A concrete model of how exactly such analogy leads to segmentation and abstraction is beyond the scope of this dissertation, but is a much needed endeavor in developmental research.

I use an example from Lily Wong-Fillmore’s study of children learning English as a second language to show what the process of segmenting larger chunks may look like (Wong-Fillmore, 1976, 1979). This study is unique in the detailed attention to larger chunks and their use over time\(^{12}\). The example is taken from Nora, a 5-year-old child who was highly successful in learning English, and produced many long sequences of

\(^{12}\) I return to this study when I discuss different kinds of learners. The study targets children learning a second language; these learners differ from first language learners in many ways. Nevertheless, the study uncovers many strategies that are applicable to first language learners. In particular, it illuminates processes that remain hidden in L1 because of the communicative, attentional, and articulatory limitations of infants. In a way, looking at child language learners may allow us to observe the utterances that infants would have produced, if they could.
language from the start. I focus on her use of the phrase *how-do-you-do-dese*. Nora used this phrase frequently, and at first produced it in appropriate contexts. Only when she began using it with nouns, as in *How do you do dese little tortillas* did it become clear that she hadn’t analyzed it correctly. Nora quickly proceeded to look for parts. As the constituent following *you* became analyzed she extracted the frame *How do you X* used to produce forms like *How do you make the flower*. This was followed soon after by the incorrect *how do make it*, where the frame *how do* on its own was used as a question frame. Following this process of analysis, which resulted in uses that were less accurate than her earlier ones, Nora began to build-up correct productive knowledge of *how*-questions.

6.2.6 Individual differences

Wong-Fillmore’s study highlights another important factor in language learning, which has not been discussed yet in this dissertation - that of individual differences among children in learning language. Despite claims to the contrary, children do not follow the exact same path in learning. Some talk more, and earlier, than others. Some children wait until they can get things right while others are less bothered by using incorrect forms (Bates et al. 1983; Bates, Bretherton & Snyder, 1991). They differ in the kind of words they start out with – even within the same language some children produce more nouns, others more verbs (Bates et al. 1994). Much research has examined internal factors that underlie these differences, that is, factors pertaining to the child’s psyche like introverted or extroverted, degree of sociability, and goals in using language (e.g. communicating, getting things, doing things, Nelson, 1973, 1981). Other studies have examined how the amount of input a child receives influences later linguistic abilities (Hart & Risley, 1995).

In keeping with the general theme of this dissertation, I want to look at individual differences through the lens of early units. Children may differ both in the units they extract, and in the ways they segment them. These differences may be driven by internal factors (what the child finds interesting) and external ones (properties of the speech directed to him or her). If, as I’ve argued in this dissertation, early units influence subsequent learning, then we may find a link between children’s initial linguistic inventories and their various trajectories. Put differently, what individual children start
out with may influence how they proceed. I want to outline several factors that could affect what early units look like in different children.

Children’s input may lead them to extract different units. Their ability to break into the speech stream may be affected by the prosodic properties and the length of the utterances directed to them. Parents who talk more clearly, and who make utterance-level boundaries more prominent, will make it easier for children to extract manageable units for further analysis. There is in fact some evidence that the prosodic clarity of caretaker speech correlates positively with children’s speech perception skills – infants ability to perceive phonetic contrasts is correlated with how clear their mothers’ speech is (how expanded her vowel space was, Liu, Kuhl & Tsang, 2003). Even more striking is the negative correlation between the proportion of unanalyzed chunks in children’s early production and the prominence of word boundaries in their care-takers’ speech. Parents who produced more words in isolation, and who tended to highlight words by repeating them with different phrases (e.g., *That’s a doggie! Look at the doggie, what a nice doggie*), had children with a lower proportion of frozen multi-word chunks (Pine, Lieven & Rowland, 1997).

More generally, we might expect children to acquire units that are easy to detect in the input addressed to them: Children may pick up on a frame or a word that a parent often uses, and use it as a building block for further learning. Indeed, 75% of children’s early words were used in isolation by their care-takers (Brent & Siskind, 2001). Children are more likely to learn common nouns when such nouns are frequently found in their care-takers’ speech (Furrow & Nelson, 1984). The relation between child and child-directed speech is bi-directional. Parents say things that interest their children – parents of a car-obsessed child will find themselves talking about cars and their attributes a disproportionate number of times. What children first extract from their input (because of their own interests or because it is salient), affects (1) the units they learn from, and (2) what they will hear more about.

Children may also differ in the way they segment early units. They may be more or less sensitive to prosodic information. If they are slower to detect smaller prosodic units, they will end up with larger units, for a longer time. Their speech may, as a result, be characterized by a more prolonged use of chunked multi-word utterances. We may be
able to identify these children on the basis of their sensitivity to acoustic boundaries\(^\text{13}\). Children may also differ in their ability to extract information and use that to segment existing units, again leading to a higher proportion of unanalyzed or mis-analyzed chunks. Adults differ in their ability to extract statistical regularities from an artificial language, and this correlates with their native language processing skills (Misyak & Christiansen, 2007; Misyak, Chrsitiansen, & Tomblin, 2009). There is no equivalent research on children’s differential ability to attend to distributional information and how that may affect their later language skills.

In sum, children may differ in the specific units they extract from speech and in the ways they go about analyzing them. These differences will result from an interaction of the children’s own interests, their ability to attend to prosodic and distributional cues, and the input they are exposed to. By highlighting the role of early units, we can allow for the how differences between care-takers, as well as between children, influence the path of learning.

6.2.7 The advantage of Starting Big

I’ve suggested that using stored utterances to learn grammar is not only a good description of how children learn language (e.g., Tomasello, 2003), but that this whole-to-part process is a unique and beneficial characteristic of first language learning. Chapter 5 provided some indirect evidence for this: Adults showed better and more child-like learning when exposed first to larger chunks of language, that are presumably more like the ones children employ.

There are several other indications of the benefits of ‘Starting Big’. Models of unsupervised grammar induction often implement a ‘Starting Big’ approach implicitly by not limiting the size of the relevant linguistic units that can be assumed. The advantage of larger chunks is more clearly seen in Bod’s experience-based models of grammar (Bod, 1998; 2006; 2009). Here, syntax is learned by starting with arbitrarily large fragments of language and deducing syntactic structure through statistical inference. Accuracy

\(^{13}\) It is hard to learn from existing research whether such differences exist since children are lumped into age groups, with little attention given to individual differences in acoustic perception (e.g., Jusczyk et al. 1992; Soderstrom et al. 2003).
increases with fragment size: The larger the units you allow the model to start with (in terms of word length and dependency depth), the better it is at generating the correct parse for an unseen sentence (Bod, 1998, 2001). In these simulations, syntactic structure is learned better when starting with larger units. While the simulations are not like first language learning in many ways (notably the lack of communicative intent and the large amount of data the model receives all at once), they corroborate the prediction that bigger is better for learning grammar.

A similar pattern can be seen in the domain of visual processing. Here too, infants attend to global features of the stimulus before more local ones (this is tested by presenting letter-like shapes arranged in a larger configuration, where the shape of the larger configuration is the global feature and the shape of the letter-like forms is the local feature, Frick, Colombo, & Allen, 2000). We can speculate that learning from less differentiated, more holistic representations may be beneficial whenever the learner has to break into complex patterns without knowing what the relevant dimensions for analysis are. Those dimensions emerge from comparing and analyzing many holistic representations, hence circumventing the chicken-and-egg problem of how the child knows what to attend to. More research is needed, using both experimental and computational methods, to map out the possible advantages of Starting Big for grammar learning.

6.2.8 Processes in learning

We can think of the first language learning as involving three general processes. First, they need to store in memory any relatively fixed chunks corresponding to utterance-level prosodic units. Given the nature of child-directed speech (Fisher & Takura, 1996), these should include single words, certain multi-word units (e.g., article-noun combinations), and short utterances. Second, they start work early on segmenting their initial inventory using prosodic, distributional, and semantic cues, to extract grammatical and referential regularities. The third process, which continues into adulthood, involves the gradual

---

14 In the 2001 paper, accuracy increased until fragments were longer than 12 words, at which point it decreased slightly. This finding is of little consequence to first language learning since children probably hear very few (if any) sentences longer than 12 words.
build-up of a rich inventory of linguistic units and constraints on how to combine them, and the accumulation of fine-grained distributional information about the relations between them (e.g., verb-bias, the probability of a word given the preceding word, the frequency of a four-word phrase, and so on).

These processes overlap in time. The first starts at birth and continues through the early stages of production (probably between 2;0 and 3;0, but this depends on the individual child), at this point in time, the child produces lots of ‘frozen’ or semi-productive language. The second overlaps with the first (children start segmenting very early on), and is characterized by the introduction of more advanced syntactic and morphological knowledge, and the appearance of over-regularized forms. The process of segmentation lasts well into the third year. Following it, children enter the third stage (lasting into adulthood).

6.3 Adult Language Learning

Unlike child learning, adult learning may progress in a different manner, especially in the early stages. Because of their prior knowledge, and their ability (and tendency) to focus on meaning-carrying units like noun-labels or verbs, adults will start out with smaller and more segmented units than the ones children extract. The nature of adult-to-adult speech, which is most likely what adult learners are exposed to, may exacerbate this tendency: sentences are longer (average length of 7.5 words in the Switchboard corpus of spoken English), there is much less repetition, and the prosody is less pronounced than in child-directed speech, making it harder to extract manageable multi-word units (ones that are not too long or complicated). Effectively, adults will skip over the initial stages of learning (collecting an inventory of different sized chunks and segmenting them), and start out directly from the third stage: learning about grammar by learning how differentiated units co-occur.

Starting from this stage will lead to an impoverished learning outcome characterized by poorer mastery of grammatical relations between adjacent units (e.g., between articles and nouns), and a non-native, overly flexible treatment of larger chunks of language (idioms, formulaic language, collocations). Both of these properties (difficulty with grammatical relations and non-native use of formulaic language) have long been noted as
characteristics of non-native speech (e.g., Fillmore, 1979; Pawley & Syder, 1980; Wray, 2002). What I propose here is an explanation of how they come about.

6.3.1 The units adults learn from

As in the case of child language, it is difficult to detect the units that adults are learning from. It is hard to show that adults depend on more fragmented units. What we can do is look at the language that adult learners produce. There are few examples of under-segmented speech, where two units are treated as one (e.g., give-it the ball) in adult learners’ speech. While such examples are abundant in studies of early child language (see chapter 2), I have yet to find similar reports in descriptions of adult learners.

On the contrary, the language of such learners is often described as overly-flexible, or non-formulaic (e.g., Fillmore, 1979; Pawley & Syder, 1983; Wray, 2002). Rather than erroneously treating two-units as one, adult L2 learners seem to treat chunks of language as more flexible than they actually are. Some of this literature was reviewed in chapter 5 in the context of L2 processing of idioms: non-native speakers process idioms as a less cohesive unit than do native speakers (Gibbs, 1980; Swinney & Cutler, 1979; Tabossi et al., 2009). Here I want to focus on some examples from production.

An illuminating case-study is presented in Wray (2004). The author followed a contestant on the TV show ‘Welsh in a week’. In this show, which encourages viewers to learn Welsh, participants have four days to learn enough Welsh to achieve a real-world task (the specific contestant had to present a cookery demonstration). They do so by memorizing fixed phrases, building up to a whole memorized script. This method eliminates one of the main challenges L2 learners are faced with - identifying what should be treated as a fixed unit. But even so, the learner often changed the forms: “the adult learner’s mind appears intent on interfering with this apparently simple business, by editing forms unnecessarily” (p. 262). A similar pattern was found in a more experimental setting, here too, L2 learners consistently deviated from the conversational formulas they were asked to memorize (Wray & Fitzpatrick, 2008). Even when asked to memorize language, L2 learners just can’t seem to leave it alone.

This over-flexibility is also found in their spontaneous production. L2 learners produce deviant collocations like being taking care of or put more attention to (Yorio,
1989, p. 62). They are overly creative with fixed phrases, for instance producing Lady first when there is only one woman present (Wray, 2008, p. 263). In general, their use of collocations and formulaic expressions is more flexible than that of native-speakers (De Cock et al. 1998; Granger, 1998). The speech of L2 learners highlights the adult tendency to ‘look for parts’, and supports the prediction that they learn from more fragmented units.

6.3.2 The role of multi-word units in learning

Even when adults do acquire chunked units, they seem to use them differently in learning. Adult learners clearly learn some fixed expressions, like greetings or requests for information, early on. But they don’t engage in the kind of scaffolding evident in children’s language. They don’t play around with chunks to extract productive frames; and they don’t extend their uses beyond the specific situation they were learned in (something that children do, often with temporarily inappropriate results…). The few formulaic expressions that adults master are not used to further their grammatical development (Wray, 1999, 2000; Yorio, 1989).

For instance, Turkish immigrants in Germany picked up a stock of phrases that enabled only minimal communication – and limiting themselves to these phrases protected them from having to interact more than they wanted (Rehbein, 1987). Even when adult learners seem to want to expand their repertoire, they don’t follow the same route that children do. Shapira (1978) notes how Ziola, the learner she followed, stuck to several simple phrases used in a limited number of contexts, without further analysis or expansion. Similar patterns are reported in other longitudinal studies of adult L2 learners from a variety of language backgrounds learning in an immersion setting (e.g., Hanania & Gradman, 1977; Huebner, 1983; Schumann, 1978).

Formulaic expressions are even less prominent in the speech of adults learning in a classroom (I use adult here to refer middle-school students and older, see further discussion in the next section). Unlike immigrants, classroom learners don’t need to quickly acquire some useful phrases to be understood. That they don’t is evident from the large pedagogical literature that focuses on how to teach formulaic language better (Wray, 2000 and references therein). The few studies that look for chunked language in
L2 learners’ speech have found that even advanced learners produce fewer formulaic sequences than do native-speakers in both spoken and written production (DeCock et al. 1998; Granger, 1998; Howarth, 1998). The ones that they do produce are often non-native-like in form or function – for instance: saying watch up instead of watch out (Scarcella, 1979).

6.3.3 Different kinds of L2 learners

So far, I’ve focused on the contrast between child and adult learners to suggest that children, more than adults, draw on multi-word units in learning, and that this is one of the factors that contribute to their better proficiency. In making this comparison, I’ve conflated two factors: whether the learner already has a first language, and whether the learner is an adult or a child. Luckily, the diversity of second language learners offers a way to disentangle those factors. There are many kinds of L2 learners: children learning a second language at different ages (and with different levels of first language proficiency); adults learning in different settings (immersion vs. classroom), and for different reasons (immigration vs. enrichment). Of most interest here is the comparison between child L2 learners who are already relatively proficient in their first language (e.g., middle-school age), and adult L2 learners. By looking how these two groups of learners use multi-word chunks, we can start to dissociate the effect of age from that of already speaking a language.

Overall, the reliance on multi-word chunks seems more related to age (child vs. adult) than to knowing a first language, or to the setting a second language is learned in (see Wray, 1999 for slightly different conclusions). Several studies have looked at formulas (larger frozen chunks of language) in child L2 learners, and they all describe many commonalities with how such units are used in first language learning. Like child L1 learners, children learning a second language use multi-word utterances early on, but their utterances have to serve a wider range of social functions (Hakuta, 1974; Huang & Hatch, 1978; Karniol, 1990; Willett, 1995; Wong Fillmore, 1976, 1979): from expressing their needs and desires (much like an infant) to communicating with their peers and establishing their social position (less like an infant). An utterance like I’ll beat’em up after school, produced by a 6-year-old L2 learner (Wong-Fillmore, 1976: 343) is not
likely to be one of the first sentences produced by a first language learner... Still, the similarity between the multi-word chunks that child L1 and L2 learners produce is striking. Many of the frozen chunks produced by the five children studied by Wong-Fillmore (1976, 1979) are also found in child L1 speech (Borensztajn et al. 2009; Lieven et al. 2003) and are frequently repeated in speech directed to children learning a first language (Cameron-Faulkner et al. 2003), among them are strings like what happened, what is this and thank you.

Child L2 learners, like children learning a first language, use larger chunks to support grammatical development. They break them down into parts and use them creatively (deviating from the memorized form), even when this leads to temporarily erroneous productions. Uguisi, a Japanese child (age 5;6) learning English, used the phrase I know how to early on, juxtaposing it with other actions to produce sentences like I know how to read it this (Hakuta, 1974). Extracting the frame I know how + X, led to incorrect, but more elaborate productions like I know how do you write this. Ultimately, as in first language learning, this creative process seems to lead to improved grammatical knowledge (Wong-Fillmore, 1976).

Child L2 learners share another interesting feature with first language learners: both engage in self-directed talk, possibly as a way of improving their linguistic skills. Such talk has been labeled ‘crib-talk’ in first language learning -- pre-sleep monologues made by young children in bed. In it, children often repeat and expand on utterances from the day. A typical instance may look something like this: on the blanket, under the blanket, what a blue blanket (produced by Anthony at age 2;6, Weir, 1962, see also Kuczaj, 1983; Nelson, 1989). While less studied, children learning a second language also seem to engage in self-talk with a “clear practice intention” (Linnakyla, 1980: 385) where they expand on utterances they have produced before. Interestingly, a substantial portion of this speech consists of multi-word utterances rather than single words, suggesting its main purpose is to facilitate syntactic and morphological learning. In sum, children learning a second language draw on larger chunks of language, and they do so in a way that resembles first language learners.

Adult L2 learners, whether learning in an immersion setting or in a classroom, rely less on larger chunks of language compared to child L2 learners. As discussed in the
previous section, this is true both in the early stages, and throughout the learning process. Learners in an immersion environment may initially use more frozen chunks than those studying in a classroom (Yorio, 1989, vs. Wray, 2000), but in both settings, these chunks remain frozen and are not used in further analysis. This stands in stark contrast with the way multi-word chunks are used by children learning a second language. The relevant distinction seems to be between children and adults, and not between learners who already speak one language and ones that do not. Knowing one language influences how you learn a second, but in terms of how multi-word units are used, the learner’s age emerges as the most influential factor.

There is an important caveat to this claim. All the studies of child L2 learners, with one exception (Willett, 1995), have looked at pre-school aged children who are not yet reading (Willett looked at first-graders who probably aren’t reading yet either). All the adult studies have looked at literate adults. Amount of schooling has been flagged as an important predictor of L2 proficiency (Flege, Yeni-Komshian, & Liu, 1999). But there is no research that asks how it affects the initial inventory of units: knowing to read may result in more fragmented early units. We cannot tell from the existing literature how much of the difference between child and adult L2 learners might be attributed to literacy: is reliance on chunks dependent on not having a channel of input (written text) where word boundaries are clearly marked?

We can start to address this question in two ways: (1) we can study language learning in illiterate adult populations, to see if they are more child-like in their use of chunks, and (2) we can compare children who can already read to ones who cannot, while keeping the learning environment maximally similar (for example, by looking at 3rd graders vs. 1st graders). In this way, we may be able to tease apart the effect of literacy and age-group. Even if literacy is an influential factor, there are other characteristics of learning in childhood that encourage the use of larger units in learning.

This should not be understood as a critical-period argument - I am not proposing that there is a biological age limit on the use of multi-word chunks in learning. Instead, age (child vs. adult) stands as a proxy for several factors that make the experience of learning language as a child (1) very different from that of an adult, and (2) more conducive for chunk learning. Children are more likely to learn in an immersion setting; their
motivation to communicate and the resulting need for catch-phrases is higher. Their input provides more opportunities to extract such phrases – pre-school is filled with repetitive routines (and the repeated language that occurs with them). It is no accident that one of the first phrases used by Juan, another of the children studied by Wong-Fillmore, was *es time to clean up*. The language children hear in pre-school is more similar (in its repetitiveness, length and amount of socialized routines) to child-directed speech than to adult speech (Wong-Fillmore, 1976). Children are less likely to be exposed to written input, where word boundaries are clearly marked. Finally, some cognitive factors make it more likely that children will attend to larger chunks, specifically, their difficulty with directing attention to parts (Thomson-Schill, Ramscar, & Chrysikou, 2009). All in all, being a child comes with a set of characteristics that make learning from chunks more likely and more facilitative.

6.3.4 Critical period effects in language learning

The difference between children and adults is one of degree. They differ in the amount of ‘chunked’ language they acquire, and in the extent that they rely on such chunks in discovering grammar. It is a difference fueled by multiple factors: the input they hear, what they already know about language, their cognitive abilities, and their communicative needs. The difference is age-related only in that age is correlated with changes in all those dimensions (experience, cognitive ability, learning environment).

What follows is a more nuanced view of age-related declines in language learning, often termed ‘critical period effects’

15 In its most stringent use, this term predicts a sharp, age-dependent cutoff point after which achieving native-like linguistic mastery is impossible (Lenneberg, 1967). L2 learners exposed to the second language before the cutoff point may show an age-related decline in ability: The sooner you learn the

---

15 I focus here on critical period effects for second language learning, and not on the way the term is used in debates about first language learning. There may well be a critical period in the sense that children not exposed to any language before a certain point in development will never learn to talk. However, these claims are often based on case-studies that conflate linguistic and cognitive development. Many of the children in question (e.g., Genie) were also cognitively impaired, and it is hard to tell whether they didn’t learn language because they were not exposed to it, or because of the devastating emotional and cognitive consequences of their neglect.
language, the better you are. Exposure after that point will lead to a uniformly poor learning outcome (Johnson & Newport, 1989), often attributed to losing access to Universal Grammar (UG).

There is ample evidence against postulating an age-dependent cutoff point for language learning. While there are clear differences in the proficiency levels of first and second language learners, there is little evidence that they are primarily driven by age of exposure (cf. Bialystok & Hakuta, 1999; Hakuta, 2001). For starters, there is little agreement on what the cutoff point is: researchers have set it at age 5 (Krashen, 1973), age 12 (Lenneberg, 1967), or age 15 (Johnson & Newport, 1989). Of more significance are studies showing no evidence of a change in learning outcome around a specific age. Using a very large sample of census data, Hakuta, Bialystok & Wiley (2003) find a linear decline in second language proficiency with age, but no sign of a “critical point”. Finding late L2 learners that are indistinguishable from native speakers (e.g., Birdsong, 1992), and finding that age ceases to be significant once other factors (like years of exposure and education) are controlled for (Flege et al. 1999), similarly undermine the notion of an age-dependent critical period.

At the same time, there is plenty of evidence for experience-based differences in learning: at a very basic level, what you know shapes what you learn. The phonetic contrasts your first language has; the neurological areas you have already committed; all these shape how you learn a new language (Flege, 1999; Kuhl et al. 1992; Kuhl, 2004; Kuhl et al. 2005). Building on this, we can formulate a ‘soft’ experience-based view of age-related differences in language learning in general, and the units of learning in particular. Prior experience shapes default patterns of learning – it affects the sound contrasts learners perceive; the semantic and syntactic contrasts they assume, and the units they attend to. To some degree, these default preferences can be modified. In the case of the units of learning, the kind of exposure you get can influence what you learn from. Under this view, the interesting question is not whether there are age-related differences but rather, what underlies them. We want to identify these preferred patterns of learning, understand their sources (cognitive, communicative, input-related), and find out how much they can be modulated.
In many ways, this view of critical period effects parallels Slobin’s ‘Thinking for speaking’ approach to linguistic relativity (Slobin, 1996). There are well-documented differences between speakers of different languages in their performance on non-linguistic tasks, most notably in the domain of spatial cognition (e.g., Levinson et al. 2002) and color perception (Berlin & Kay, 1969; Winwar et al., 2007). But the question is how to interpret these differences: what do they tell us about the relation between language and thought? Slobin’s answer is that they reflect preferred modes of expression. Speakers prefer to make the distinctions that their language encodes. Language affects thought by making certain dimensions of an event easier to express. In a similar way, your first language shapes how you approach learning a second, but these preferences can, to some degree, be affected by new experience. Slobin’s view on linguistic relativity, like the outlook of L2 learning described above, emphasizes the subtle ways in which previous experience makes itself known.

6.4 The Units of Language

Up to now, I’ve discussed the findings in this dissertation with relation to what they tell us about the linguistic units that are used in learning. But there are also implications for how we think of the units of language more generally.

6.4.1 Multi-word phrases in the adult lexicon

As reported in chapter 4, compositional phrases showed whole-form frequency effects like those displayed by simple and inflected words: adults were sensitive to the frequency of the whole phrase when substring frequency was controlled for. Following the morphological literature (e.g. Alegre & Gordon, 1999; Baayen et al. 1997; Taft, 1979, see chapter 4 for a more elaborate discussion), I took whole-form frequency as an indicator of whole-form representation. Multi-word phrases, like non-inflected words, are part of the adult linguistic inventory.

What exactly this means depends on our definition of ‘the mental lexicon’ and of ‘representation’ more generally. In contrast with previous claims (e.g., Chomsky and Halle, 1968; Pinker, 1999), the mental lexicon does not seem like a repository for atomic elements. Even though phrases can be generated, speakers still have memory traces of the
whole-form. The fact that speakers show similar effects for regularly inflected words (e.g., Baayen et al. 1997) and productive noun-compounds (Bien, Levelt & Baayen, 2005) further suggests that “minimizing storage” is not a defining characteristic of the mental lexicon. Instead, the mental lexicon may contain much redundancy, with speakers having complementary representations of different grain-sizes (in a sense, such redundancy already exists in any model that has lexical entries for words instead of deriving them each time from the phonemic inventory).

This still doesn’t answer the question of what it means to have multi-word phrases in the mental lexicon. I am not proposing that speakers store a list of phrases. Or that these phrases have no internal structure and cannot be generated from their parts. But I do want to suggest that phrases are as psychologically real to speakers as more conventional linguistic units like words, and that this reflects the usage-based (and usage-driven) nature of representation: speakers’ linguistic knowledge is an accumulation of their linguistic experience. That experience includes, among other things, multi-word sequences.

Many studies seek to identify the primary units of language: the building blocks that speakers draw on in speech production and perception. But this investigation makes sense only if we think such a static, well-defined inventory exists. Alternatively, we can think of linguistic representation as a constantly evolving system that is fundamentally rooted in interaction (De Bot, Lowie & Verspoor, 2007; Elman, 2009; Goldinger & Azuma, 2003; Grossberg, 1980; Van Geert, 1991). Language users attend to many kinds of linguistic patterns - linguistic units pick out salient and consistent aspects of this experience. But what serves as a primary unit (for production or perception) may depend on many factors, including the task at hand, and the feedback of the interlocutor. For example, whether the syllable or the phoneme is more prominent in speech perception depends on what speakers are asked to do (Goldinger & Azuma, 2003). The fit between the traditional units of linguistic analysis (phonemes, morphemes, words) and the units that speakers rely on is good, but not perfect: there are additional units, like multi-word ones, that speakers can also draw on.

We should be able to find mental traces for linguistic patterns of all kinds, as seems to be the case. The effect that different units have on processing will depend on the task at
hand (e.g., are speakers being asked to recognize words or phrases?), and on their characteristics: Does the unit (word or phrase) have a special meaning? Does it serve a special communicative function? Do its parts appear together more often than would be expected? Certain multi-word units, like idioms or collocations, may show more of a processing advantage than others (like compositional multi-word phrases) because they exhibit more unit-inducing features.

More broadly, viewing language as a constantly changing system, and language use as a reflection of this vivacity, shifts the focus of investigation from identifying static entities to uncovering the complex dynamic of learning and interaction. It allows for a much richer interpretation of the kinds of processes and units that speakers rely on when they use language. The dynamic perspective is also useful for thinking about learning. Instead of viewing it as a process with a clear beginning and end state, we can characterize language learning as the development (and departure from) attractor states that is (a) influenced by the initial state of the system, (b) affected by individual differences, and (c) in constant interaction with the environment (this explains why children learn the language they are exposed to, and why native-speakers share a common form of linguistic expression).

6.4.2 Multiple-access models of the lexicon

Allowing for complementary (and redundant) representations of linguistic material creates multiple ways to produce the same complex form, using components of varying levels of abstractness. An utterance like *give me the ball* could be produced by inserting single words into the dative construction, combining the stored chunk *give me* with a noun-phrase, or accessing the frozen utterance *give me the ball* (unlikely, but maybe more likely in the case of the more collocated *give it to me*). Any of these would allow the speaker to convey the desired meaning.

A similar observation holds in the morphological literature. Under dual-access models, inflected forms can be accessed as a whole or through their parts (Baayen, 1992; Baayen et al, 1997; Stemberger & MacWhinney; Stemberger & Middleton, 2003). Which path is chosen depends on many factors: the nature of the task, the relative frequency of the whole form and the parts, and more. To extend these ideas to larger chunks of
language we need to allow for multiple forms of access. A complex form can be produced from any combination of chunks and parts. One way to understand the pervasiveness of formulaic language (up to 80% of adult speech in some estimates, Alternberg, 1998), is that sequences that are used often develop a path for whole-form access: the more you use something, the easier it becomes to produce. This preference will become more entrenched over time for utterances that are highly collocated overall, or used frequently in specific situations (allowing for meaning and social function to enter the equation).

There are parallels between this description and processes of grammaticalization where frequently occurring elements may become fused into one unit, changing semantic or syntactic features along the way (Hopper & Traugott, 2003). The synchronic effect of chunk frequency on language users (i.e. the increased cohesion of units that co-occur often that is reflected in their privileged status in production and comprehension) may play into diachronic change. Units that are likely to develop a whole-form access route will also be the ones that will be likely to ‘merge’ over time.

Chunk frequency may also play a role in processes of re-syllabification, where elements that started off as two units, are mis-analyzed and segmented into differently sounding units. For example, in English, an ewt became a newt: the final consonant of the indefinite article became the initial consonant of the noun. We can see the reverse process in the change from a napron to an apron where the initial consonant of the noun was mis-analyzed as the final consonant of the indefinite article. Similar examples can be found in many other languages (French: lonce (lynx) > l’once, Italian: l’astirco (pavement) > lastrico, Spanish: el lagarato (alligator) > alagarto). These processes may be more likely to take place in more chunked units - ones where the elements are found together more often than expected given their individual distribution.

In this chapter, I’ve discussed some implications and further refinements of the Starting Big Hypothesis. In the coda that follows, I present a more general view of language, the main tenants of the SBH, and discuss a few limitations and directions for future work.
CHAPTER 7: STARTING BIG REVISITED

7.1 Tenets and Predictions

The Starting Big Hypothesis can be refined from its early incarnation in the following way:

1. Children learn language by analyzing stored exemplars. Their initial exemplars correspond to prosodic, not lexical boundaries. Given the nature of child-directed speech, these exemplars will include a mix of single words, phrases and short multi-word utterances.

2. Grammatical and lexical knowledge develops by segmenting and analyzing early units using many cues: distributional, prosodic cues, semantic. This process, of learning grammar from unanalyzed chunks, creates stronger links between elements that were once part of the same unit. It is better to learn grammar and segmentation at the same time.

3. Children gradually accumulate a rich inventory of linguistic units and a fine-grained understanding of the relations between them. The adult inventory contains linguistic units of varying sizes and levels of abstraction. Adult-like linguistic knowledge is the sum of past linguistic experience. Previous exposure influences the future processing of all patterns (words, phrases, constructions).

4. Adult L2 learners, because of their prior knowledge, the input they hear, and the communicative setting they learn in, learn from more fragmented representations, where word boundaries are more distinct. This leads to poorer mastery of grammatical relations between adjacent units (e.g., articles and nouns), and a non-native, overly flexible treatment of larger chunks of language (idioms, formulaic language, collocations).

The Starting Big Hypothesis offers a novel framework for understanding and investigating L1/L2 differences in language learning. At the same time, it expands on previous accounts of first language learning by giving a systematic account of children’s
early units and the way they affect learning. Learning from multi-word phrases is advantageous, and is in fact, one of the features that distinguish child language learning from adult language learning. The process of learning is viewed as usage-based, exemplar-driven, and involving a gradual process of abstraction. The product of learning is a hierarchically structured network of linguistic patterns, of varying sizes, all processed by the same cognitive mechanism.

7.2 What is Language? Looking Back for Answers

The first volume of *Forum Linguisticum*, published in 1976, contains a scathing critique of Chomskyan views of language. In clear language, the authors dismantle the most basic assumptions of the ‘Chomskyan Revolution’: the distinction between competence and performance, the scientific nature of the endeavor (Gray, 1976), its divergence (and compatibility) with linguistic traditions of the past (Izzo, 1976). Bolinger’s article in particular (Meaning and memory, 1976) offers a compelling alternative to the dogmas of the time. As in other cases (e.g., the effect of semantics on accent placement, Bolinger, 1972), the ideas Bolinger presented then, are 30 years later becoming commonly held.

I want to highlight several claims in particular, and present my take on them, based on the findings presented and discussed in this dissertation. Together, they illustrate what I believe to be a promising approach to what it means to know, learn and use language (all the examples are taken from Bolinger’s article).

a) “Speakers do at least as much remembering as they do putting together” (Bolinger, 1976: 2). Language use is more memorized than creative. This is obviously true for idioms, and archaic forms (like *nothing loth*) where speakers cannot produce the forms from scratch. But many other aspects of language lack the productive freedom so valued by linguists. We have idiosyncratic restrictions on the ways words can be combined. For instance, we can say *somewhere else* but not *sometime else, a long time ago* but not *an extended time ago*. We combine words differently with different lexical items. We produce blueprints but not buildings, and invent stories but not novels. We prefer certain combinations over others: we say *salt and pepper* but not
pepper and salt. In all these cases, there is something about the output that is unpredictable given the rules. We could still describe language as (mostly) generative if these unpredictable cases were easily distinguished from generated language. But similar unruly restrictions show up all over the place: why does *hurt her terribly* suggest emotional injury while *hurt her badly* implies physical pain? We can detect a degree of “unfreedom” (Ibid., p.9) in much, if not all, of the language we produce.

b) “It is not because the generative mechanism is lacking” (Ibid., p. 4). Native-speakers are aware of these restrictions, but they can play around with them (often for pragmatic reasons), and are capable of producing novel utterances. The rigidity and flexibility of language come from the same source – the way language is used. Our sense that language has rules comes from the consistency in usage-patterns. Our ability to generate new sentences comes from the analysis and abstraction of what we hear. This ability, like language itself, is constantly changing and evolving, and is never entirely separate from concrete examples (it makes little sense to talk about an NP > Det + N rule without taking into account the concrete tokens that led to its formulation).

c) “Learning goes on constantly – but especially in young children - in segments of collocation size as much as it does in segments of word size. Much if not most of our later manipulative grasp of words is by way of analysis of collocations” (Ibid., p. 8). Children discover the patterns of language by analyzing stored exemplars. They listen to their care-takers, extract what they can, and use this early inventory as the basis for learning. Multi-word units are especially important in this process because they reveal the intricate ways that words can co-occur: they allow children to learn about the grammatical and distributional relations between words (to achieve native-like fluency, one has to learn both). We no longer need two learning mechanisms: one for rules, and one for exceptions. Instead, both grow out of the same process of usage-driven generalization.

d) “We must retain the collocations even after the individual words have become entities in their own right” (Ibid., p. 9). In the process of learning, speakers
build-up an inventory of linguistic units of varying sizes and levels of abstraction. Larger units, used as building blocks, don’t disappear, but instead co-exist along with their parts (words, morphemes). The structure of this inventory is continuously affected by use: some units grow closer, some move apart. We are sensitive to the usage profiles of all these units: how often they appear, with which other units, and in what contexts. How we produce and comprehend language is affected by this information.

We can find historical precursors for many of these views. Bolinger himself echoes voices from the past. In another article in the same volume Izzo (1976) criticizes the treatment of historical accounts of language in Chomskyan linguistics. In particular, he laments the neglect of Renaissance scholars, whose views on language resemble current usage-based accounts much more than they resemble Universal Grammar. I leave the reader with some of those quotations (taken from Izzo’s article), which outline, in my view, some productive assumptions about what language is.

I. Speroni 1542: I firmly believe that the languages of every nation, Arabic or Indian, just as much as Latin or Greek, are of equal value and were shaped by mortals for the same purpose.

II. Varchi 1570: I say that languages must be learned from the people, that is from the usage of those who speak them. Language consists of the usage of those who speak them.

III. Buonmattei 1623: We have said then, that pronunciation is customary usage; for all the rules of languages are nothing but customary usage. In the eyes of those who are versed in these matters, those people appear ridiculous who think – and worse, teach – that the rules of language should be deduced from reason.

7.3 Future Research

The research presented in this dissertation leaves many open questions about the role of larger units in learning and processing. There are four domains in particular, spanning child learning, adult learning and adult processing, that call for further investigation: (1)
the status of unanalyzed multi-word units in learning, (2) the way these units are segmented, (3) the nature of early units in adult learners, and (4) the dimensions that influence the processing of multi-word phrases in adults. In the next section, I discuss these issues and present several ideas for exploring them further.

7.3.1 The status of unanalyzed chunks in child learning.

There is quite a lot of evidence suggesting that children draw on multi-word units in both production and perception: infants can remember them; young children attempt to produce them; there is computational and corpus-based support for chunked units in development. What remains less explored is their role in learning. How can I show that children utilize larger units in learning about smaller ones? Chapter 3 documented a relation between phrases and words, but did not show how that relation developed. The fact that 4;6 year-olds produced teeth better in brush your teeth does not mean that they started off with brush your teeth as a single unit. In fact, given their age (and the amount of linguistic knowledge they already have), it is more likely that they produced the form better because they were sensitive to the frequent co-occurrence of the frame and the word, just like adults are.

We may be able to use longitudinal data to address this question. If children use phrases to learn about words, I would expect the relation between the two to be stronger in younger children. More than that, we may see a U-shaped pattern in the effect of phrases on word production. Early on, words will be tightly linked to the early phrases they appear in – hearing a phrase will enhance word production. In a later stage, the emphasis on segmentation and analysis will weaken these links, and phrases that occur frequently in child-directed speech will no longer be as facilitative. Finally, as children are exposed to more language, the ties between those phrases and words that keep appearing together will again become strong.

I can make the following predictions. Using the same design (of irregular plural elicitation), I can expect the following longitudinal pattern: Two-year-olds will produce irregular forms better following frequent phases; three-year-olds, on the other hand, who are at the intense segmentation phase, will not show this effect (or show it less). Together
with the data I already have from the 4;6 year-olds, this would document a shift from chunked knowledge (where phrases and words are closely linked), through segmentation (where the links weaken), to the accumulation of distributional information that we see in adults.

### 7.3.2 Chunk Age-of-Acquisition.

Another way to investigate the primacy of unanalyzed chunks in early learning is by looking for Age-of-Acquisition (AoA) effects for multi-word units. The term comes from the lexical literature, and is used to refer to the privileged status of words that are acquired early on. Some words (like *cat*) are almost certainly acquired before others (like *vase*). AoA can be seen as a proxy for the order of acquisition of words: which words were acquired first, and which serve as the early building blocks for the lexicon. Speakers show processing advantages for early acquired words: controlling for frequency (and many other factors), early-acquired words are named faster, read faster and recognized faster (see Juhasz, 2005 for a review).

One explanation for lexical AoA effects is provided by network plasticity accounts. In this framework, AoA effects are a natural consequence of a learning mechanism that loses plasticity as learning progress. Using connectionist simulations (Ellis & Lambon Ralph, 2000; Monaghan & Ellis 2002), researchers showed that patterns seen early in training have more influence on the final weights of the network.

If multi-word units are processed like words, and if they serve as building blocks for learning grammar, then they too should exhibit AoA effects. I should be able to find traces of early acquired chunks in adult processing. To do this, we need to identify early acquired chunks, and then see whether they show any processing advantages. As in the word AoA literature, we can use frequency as a proxy for age-of-acquisition – let’s assume that chunks that are frequent in child-directed speech are acquired earlier. I can now compare pairs of multi-word phrases that are equally frequent in adult speech but where only one was acquired early (based on frequency counts). For example, *to get a little* is more frequent than *to get a lot* in child-directed speech, but both phrases are equally frequent in adult-to-adult speech. The phrase that is more frequent in child-
directed speech is classified as early acquired (it is hard, but not impossible to construct such pairs). If there are chunk AoA effects, then the early-acquired phrase should be faster to process despite equal frequency in the adult language.

7.3.3 The process of learning: from chunk to segment to cohesive unit.

The three-staged model of learning that I described in the General Discussion makes several predictions that have not yet been tested. Multi-word utterances will start out chunked, then enter a period of analysis, and finally become fused again through repeated use. This could be tested by looking at the duration of early multi-word utterances. Duration can be used to measure how chunked a unit is: words are reduced in frequently occurring sequences (e.g., Bybee & Schiebman, 1999). The process of segmentation could be examined by looking at how duration changes over time.

To do this, I need to look at multi-word utterances that are produced early on, and that seem unanalyzed (there is no evidence for productive knowledge of their parts). Candidates include expressions like what-is-this or how-are-you. Their duration in early productions could be shorter than their duration in later periods, when they are somewhat analyzed. In fact, there may be three distinct stages: (a) short durations early on which (b) become longer as segmentation occurs, and finally (c) reduce again as a result of repeated usage.

7.3.4 Are U-shaped patterns related to increased segmentation?

U-shaped patterns of learning have fascinated many researchers because of the display of what seems like non-linear or even regressed learning. They are cases where learning seems to deteriorate before improving. In language, the best studied examples are of over-regularization errors where children initially produce the correct irregular form, but then go through a phase of also producing an incorrect form before settling on the correct one. The following exchange in Hebrew is a great example of such a pattern. The conversation, cited by Berman (1985), takes place between three siblings, aged 3, 4, and 5 who are engaged in a bedtime ritual. The important thing to know is that the word for
night in Hebrew has feminine phonology but is masculine (layla). The correct form of the phrase good night is layla tov.

A: Layla tov (3-year-old)
   Night-Masc good-Masc (correct)

B: lo, layla tova (4-year-old sibling)
   No, night-Masc good-Fem (over-regularized)

C: LO, omrim layla tov (correct) (5-year-old sibling)
   No, say night-Masc good-Masc

The three-year-old produces the correct form, the four-year-old over-regularizes, and the five-year-old is back to producing the correct form. One fruitful way of interpreting these patterns (which are also found in motor development) is found in dynamic-system theory (Gershkoff-Stowe & Thelen, 2004). In this account, U-shaped patterns reflect a re-organization of the system, a shift from one cue to another, an increase in the complexity of the system. That is, they demonstrate progression, not regression. Adopting this perspective, I want to explore the idea that U-shaped patterns in language reflect a productive stage of increased segmentation. I want to see if there is a way to evaluate how ‘segmented’ a child’s language is, and to tie that to their production of over-regularized forms. This would require devising a method for assessing degree of ‘segmentation’.

7.3.5 The units adults learn from.

The speech of adult L2 learners can be described as more fragmented than that of native speakers. They use fewer collocations and idioms, and even when they use them, they tend to be more flexible than they should. This pattern fits well with the idea that L2 learners have a tendency to look for parts, and that this can become a disadvantage in language learning. But it is hard to (1) identify the units that L2 learners use in learning,
and (2) quantify the perceived difference between children and adults in the use of multi-word chunks.

We can start to address both issues by applying the same computational models (Borensztajn et al. 2009) and productivity calculations (Lieven et al. 2003) used to identify chunks in child speech, to the speech of L2 learners. If adults use more fragmented units, then there should be a smaller proportion of multi-word units in their productions. Ideally, we would use a sample of speech similar to the one used for children: naturalistic, spoken, and interactive. But such a corpus is not currently available for L2 learners (compiling such a corpus would be of great benefit to the study of second language learning in general).

What is available is the International Corpus of Learner English: a 2.5 million word collection of 3,640 essays written by learners from eleven European languages (Granger, Dagneaux, & Meunier 2002), and its non-learner equivalent, the Louvain Corpus of Native English Essays (LOCNESS), a smaller corpus (323,300 words) of essays written by native English-speaking university students (Granger, Sanders & Connor, available upon request). These two corpora can be used to investigate ‘chunked’ language in L2 writing, and compare it to native-speaker text elicited under similar conditions.

7.3.6 Which dimensions influence the processing of multi-word units?

In my studies, I used string frequency as an estimate for sequence strength: irregular nouns were produced better in frequent phrases; adults were faster to recognize higher frequency phrases. This is an external measure of how often a specific sequence appears. But there are other, ‘internal’ measures that are probably equally important: the predictability of a word given a previous or later sequence, the degree of mutual information between the components, how often they appear in other constructions (e.g. teeth probably appears in fewer sequences than hand). If the processing of multi-word units is similar to that of words, there will be many factors in addition to token frequency (in this case, phrase or frame frequency) that affect processing. I plan to explore those factors using the same phrasal-decision task used to look at phrase frequency.
In sum, in this dissertation I’ve focused on the units of learning as a way to address questions in first language acquisition, adult processing, and second language learning. I’ve argued for a usage-based view of learning that emphasizes the importance of multi-word phrases in the construction of grammar. By doing so, I’ve been able to offer a novel perspective on the difficulty that adults experience in learning a second language. Adults’ increased reliance on more segmented units (because of their prior knowledge and learning situation) negatively impacts their ability to achieve native-like fluency, in particular in learning how to use idiomatic and formulaic language, and in learning grammatical relations that are more semantically opaque. The study of compositional multi-word phrases allowed me to contrast single-system and dual-system approaches to language and present evidence in support of an emergentist view of language where all linguistic experience (be it atomic or complex) is processed by the same cognitive mechanism. Drawing on additional findings, I’ve formulated the Starting Big Hypothesis as a framework for studying the processes involved in language learning and use.
### A.1 Experiment 3 Stimuli

#### A.1.1 High-frequency bin (with frequency per million words in the Fisher corpus)

<table>
<thead>
<tr>
<th></th>
<th>Phrase</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A lot of places</td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td>A lot of days</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>A lot of work</td>
<td>14.70</td>
</tr>
<tr>
<td></td>
<td>A lot of years</td>
<td>1.90</td>
</tr>
<tr>
<td>3</td>
<td>All over the place</td>
<td>21.45</td>
</tr>
<tr>
<td></td>
<td>All over the city</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>Don't have to worry</td>
<td>15.30</td>
</tr>
<tr>
<td></td>
<td>Don't have to wait</td>
<td>1.40</td>
</tr>
<tr>
<td>5</td>
<td>Don't know how much</td>
<td>12.80</td>
</tr>
<tr>
<td></td>
<td>Don't know how many</td>
<td>7.80</td>
</tr>
<tr>
<td>6</td>
<td>Go to the doctor</td>
<td>16.70</td>
</tr>
<tr>
<td></td>
<td>Go to the beach</td>
<td>5.65</td>
</tr>
<tr>
<td>7</td>
<td>How do you feel</td>
<td>29.60</td>
</tr>
<tr>
<td></td>
<td>How do you do</td>
<td>4.95</td>
</tr>
<tr>
<td>8</td>
<td>I don't know why</td>
<td>35.15</td>
</tr>
<tr>
<td></td>
<td>I don't know who</td>
<td>7.00</td>
</tr>
<tr>
<td>9</td>
<td>I have a lot</td>
<td>26.45</td>
</tr>
<tr>
<td></td>
<td>I have a little</td>
<td>8.95</td>
</tr>
<tr>
<td>10</td>
<td>I have to say</td>
<td>15.40</td>
</tr>
<tr>
<td></td>
<td>I have to see</td>
<td>0.95</td>
</tr>
<tr>
<td>11</td>
<td>I want to go</td>
<td>9.10</td>
</tr>
<tr>
<td></td>
<td>I want to know</td>
<td>2.95</td>
</tr>
<tr>
<td>12</td>
<td>It's kind of hard</td>
<td>13.30</td>
</tr>
<tr>
<td></td>
<td>It's kind of funny</td>
<td>7.20</td>
</tr>
</tbody>
</table>
13. On the other hand 27.15
On the other end 3.95
14. Out of the house 9.75
Out of the game 0.70
15. We have to talk 9.70
We have to say 0.65
16. Where do you live 44.80
Where do you work 2.60

A.1.2 Low frequency bin (with frequency per million words in the Fisher corpus)

1. A lot of rain 4.65
   A lot of blood 0.20
2. Don't have any money 2.35
   Don't have any place 0.25
3. Going to come back 1.35
   Going to come down 0.40
4. Have to be careful 5.90
   Have to be quiet 0.15
5. I have a sister 4.90
   I have a game 0.10
6. I have to pay 1.80
   I have to play 0.10
7. I want to say 3.60
   I want to sit 0.20
8. It was really funny 2.65
   It was really big 0.15
9. Out of the car 2.00
   Out of the box 0.20
10. We have to wait 1.65
    We have to leave 0.25
11. We have to talk 9.70
   We have to sit 0.20
12. You like to read 1.55
   You like to try 0.10

A.2 Experiment 4 Stimuli

A.2.1 Mid frequency bin (with frequency per million words in the Fisher corpus)

1. A lot of problems 9.60
   A lot of power 0.85
2. All over the country 9.55
   All over the house 0.75
3. Be able to go 8.40
   Be able to see 3.95
4. Do you know how 6.40
   Do you know when 1.85
5. Go back to school 6.75
   Go back to work 4.00
6. How do you get 6.95
   How do you go 1.05
7. I don't really care 6.20
   I don't really need 0.85
8. I don't see how 9.20
   I don't see them 2.45
9. It takes a lot 7.25
   It takes a little 1.45
10. Know what that is 6.25
    Know what that was 1.05
11. Not going to get 7.95
    Not going to see 0.75
12. Out of the house 9.75
   Out of the car 2.00
13. Take care of them 6.90
   Take care of things 0.80
14. To have a lot 6.55
   To have a little 2.70
15. We used to go 7.05
   We used to be 2.25
16. You know how many 5.40
   You know how long 3.40
17. You're going to be 9.70
   You’re going to do 4.05

A.3 Experiment 5 Stimuli

A.3.1 Artificial language vocabulary and English meaning (by alphabetical order)

1. Etkot Key
2. Fertsot Sock
3. Geesoo Hat
4. Gorok Bike
5. Hekloo Bath
6. Hertin Iron
7. Jarkad Cup
8. Jatree Pan
9. Panjol Television
10. Perdip House
11. Toonbot Clock
12. Slindot Piano
13. Sodap Spoon
14. Viltord Plane
REFERENCES


Velleman, S. L. & Vihman, M. M (2002). Whole-word phonology and templates trap, bootstrap, or some of each? Language, Speech and Hearing Services in Schools, 33, 9-23.


Granger (Eds.), *Phraseology in language learning and teaching* (pp. 123-148). Amsterdam: John Benjamins.
